Chinese Journal of Chemical Engineering 88 (2025) 222–238

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

journal homepage: www.elsevier.com/locate/CJChE

Full Length Article

Source term estimation of hazardous gas leakages under turbulent atmospheric transport dispersion scenarios

Chuantao Ni [1], Ziqiang Lang [1, 2, *], Bing Wang [1], Ang Li [1], Chenxi Cao [1], Wenli Du [1], Feng Qian [1]

[1] *Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China*
[2] *Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom*

ARTICLE INFO

Article history:
Received 17 March 2025
Received in revised form
17 June 2025
Accepted 23 June 2025
Available online 13 August 2025

Keywords:
Computation fluid dynamics
Hazardous gas leakage
Optimization
Source term estimation
Turbulent flow

ABSTRACT

Source term estimation (STE) of hazardous gas leakages in chemical industrial parks (CIPs) is important for addressing environmental pollution and improving safety and reliability in engineering practice. To achieve real-time STE, least squares-based STE methods have recently been developed. However, these methods require the number and locations of potential hazardous gas leakage sources are known as a priori, which is difficult in many practical scenarios. To address this limitation, we propose a new data-driven STE approach, which enables the STE to be implemented in real time and applicable to complicated turbulent dispersion scenarios. The linear independent analysis in data science is applied to historically collected concentration data of a hazardous gas of concern from a network of sensors to extract the sensor data which represent independent hazardous gas leakage scenarios (IHGLSs). An appropriate Gaussian model approximation to a high-fidelity computational fluid dynamics (CFD) model that must be used to represent the hazardous gas leakage scenarios of concern is built, and the off-line STE of IHGLSs using the approximating Gaussian model is then performed to build the data-driven STE model. The performance of the proposed approach is evaluated by using data that are generated by simulating ethane leakage scenarios in a CIP using a CFD model. Results indicate that the leakage localization accuracy is 100% and the mean relative estimation error for the leakage strength is 6.76%. Moreover, the proposed approach is validated with real data in Prairie Grass field dispersion experiments, demonstrating the practical applicability of the proposed approach.

1. Introduction

Recently, chemical industrial parks (CIPs) have been rapidly emerging and developing in many countries around the world, and accidental leakages of hazardous gases in CIPs are becoming an issue of great concern [1]. Hazardous gas leakages in CIPs can lead to fire and explosion, casualties, environmental pollution, economic losses, and public panic [2]. Therefore, it is necessary to quickly and accurately estimate the location and strength of hazardous gas leakages. Once this information of hazardous gas leakages has been determined, decision-makers in the CIP can promptly take emergency response actions such as air purification, evacuation, risk assessment, and hazard warnings to address the risks of hazardous gases to the environment and public safety [3–5].

Source term estimation (STE) is concerned with estimating the unknown parameters of the sources of hazardous gas leakages using different inversion methods from available information of meteorological conditions, ground geometry, as well as hazardous gas concentration measurements [6]. The concentration measurements can be from a network of static sensors or mobile monitoring stations. Based on static-sensor networks, many STE methods have been developed and can be classified into four main categories: nonlinear optimization (NO) approaches, machine learning (ML)-based techniques, data assimilation (DA) approaches, and least squares (LS) methods.

* Corresponding author. Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom.
E-mail address: z.lang@sheffield.ac.uk (Z. Lang).

https://doi.org/10.1016/j.cjche.2025.06.025
1004-9541/

Generally, the NO-based STE methods have three critical factors: sensor measurements, an atmospheric transport and dispersion (ATD) model representing the source-sensor relationship, and an estimating algorithm that uses the sensor measurements and the ATD model to obtain the STE results [7]. The ATD model is used to predict the concentrations of leaking hazardous gases at the locations of the sensors. The predicted concentrations are then compared with the sensor measurements, and the discrepancies are used by either an optimization algorithm or a Bayesian inference algorithm to search and derive the unknown source parameters that minimize a cost or likelihood function [8–14]. The efficiency and accuracy of the forward ATD model greatly affects the performance of the NO-based STE methods. Therefore, some ML-based ATD models have been developed and used as the forward ATD models to perform the NO-based STE. The particle swarm optimization algorithm was applied to identify the leakage source parameters with integrated Gaussian-machine learning network models being the ATD models [15]. A ML algorithm model was proposed in Ref. [16] and coupled with Markov chain Monte Carlo method to estimate the source terms of pollutants. A forward ATD model based on radial basis function neural network and Gaussian model was presented in Ref. [17] and then applied to determine the leakage source parameters. These can greatly improve the efficiency and accuracy of ATD model calculations during the optimization process. Nevertheless, it is still difficult to achieve real-time STE due to the complexity associated with the required on-line optimization processes. Furthermore, in the case of simultaneous leakage from multiple sources, the number of leakage sources is usually required to be known a priori by existing methods [18], which may not be fulfilled in many practical scenarios.

The ML-based techniques aim to use machine learning models, such as artificial neural networks (ANNs) [19–21], convolutional neural networks [22–24], attention-based long short-term memory (LSTM) [25], gated recurrent units [26], LSTM auto-encoder [27], and LSTM recurrent neural networks [28], to extract useful knowledge from massively simulated or experimentally collected data and establish an inverse machine learning model that can directly use real-time sensor measurements to determine unknown source parameters. The ML-based STE techniques have the potential to address the speed and efficiency issues and break the limits with using physical principle-based ATD models for the STE. However, the STE results may be poor once the real-time sensor measurements are beyond the range of training data [24,29].

To achieve real-time STE, certain studies are delving into the DA-based STE methods. By assimilating the sensor measurements into an appropriate ATD model continuously, the DA-based STE seeks to estimate the strength of the leakage source and predict the spatio-temporal concentration distribution of hazardous gases. The DA-based STE can achieve optimization by analyzing the interactions in the physical field, including spatial and variable correlations, thereby avoiding complex matrix derivations and computational processes [30,31]. Therefore, the revised spatio-temporal distribution of hazardous gas concentrations and reasonable source parameters can be obtained in a relatively real-time manner. As a kind of sequential DA method, the ensemble Kalman filter algorithm is easy to be combined with different ATD models and able to account for uncertainties in STE results [32]. Several studies have verified its superiority in real-time parameter estimation and forecast correction [31–35]. However, so far, existing DA-based STE approaches can only deal with the scenarios of single source leakage with a known location and can, therefore, still not be applied in practice.

Concurrently, the LS-based STE methods have been proposed to facilitate the real-time STE. The LS-based STE methods assume that there is a linear relationship between the sensor measurements and the strengths of hazardous gas leakages [36]. The basic idea is to use this linear relationship to determine a response matrix that represents how the sensor measurements are affected by the strengths of hazardous gas leakages, and to use the response matrix to estimate the leakage strengths in real time. To apply the LS-based STE methods, the response matrix should be pre-determined. Most methods used the ATD model to directly calculate the response matrix [36–40], while the impulse response method, where the components in the response matrix are determined using the concentration response for the unit impulse release of the leakage source at a given location, was suggested in Refs. [41,42] to address this issue. However, these LS-based STE methods often cannot be applied in practice as the methods require the number and locations of possible hazardous gas leakage sources are known a priori, but the condition may not be satisfied in many practical applications. In addition, even if the condition is satisfied, the response matrix that is determined from dedicated ATD model simulations or field experiments may not be applicable in practice. This is because, without properly exploiting the information of multi-sensor data routinely measured in a CIP, the response matrix that has been determined may not be representative and cannot well describe the relationship between the strengths of gas leakages and the multi-sensor data in practical scenarios of concern.

Very recently, a multi-sensor data-driven STE (MSDD-STE) approach was proposed in Ref. [29], which can potentially address the problems with the existing LS-based STE methods. The idea is to offline establish a STE model from historically collected sensor data measured during a period that cover the situations known as independent hazardous gas leakage scenarios (IHGLSs) in the CIP and then online apply the established STE model to process field-measured sensor data to estimate the strengths of hazardous gas leakages in real time. Further, a new MSDD-STE approach that can work over variable wind directions is developed in Ref. [43], where a radial basis function-based ANN was determined and applied to represent the inverse of the response function matrix valid over a range of wind directions. Instead of requiring knowing the number and locations of possible hazardous gas leakage sources, both approaches determine this information and then build the STE models directly from historically collected sensor data, providing a solution that can be directly relevant to the concerned practical scenarios and can, therefore, have a better potential of practical applications. However, the problem with the methods proposed in Refs. [29,43] is that they are only applicable in the hazardous gas leakage scenarios when the Gaussian model can be used as an ATD model so that the required offline STE can readily be solved.

In order to resolve the limitation with the existing MSDD-STE approaches, a new method is proposed in the present study. The idea is to integrate the DA technique into the MSDD-STE framework. This is to find an appropriate Gaussian model approximation to the high-fidelity ATD model such as a computational fluid dynamics (CFD) model that has to be used to represent the hazardous gas leakage scenarios of concern and then solve the problem using the MSDD-STE approach. The new idea can extend the application of the MSDD-STE idea to more complicated hazardous gas leakage scenarios such as scenarios where turbulent ATD models have to be used to represent hazardous gas leakage scenarios. Comprehensive simulation studies are performed to verify the effectiveness and robustness and demonstrate potential practical applications of the new method.

The remainder of this paper is arranged as follows. In Section 2, the hazardous gas leakage STE problem under turbulent ATD scenarios in a CIP is described. In Section 3, the new STE method and corresponding implementation algorithm are proposed.

Section 4 establishes, based on the CFD model-based simulation of the turbulent ATD scenarios, a concentration database of leaking hazardous gas for the CIP under study. Section 5 applies the new STE method to the CFD model simulation data generated in Section 4 and discusses the STE results that have been achieved. Finally, in Section 6, conclusions are reached and possible future studies are summarized.

2. Problem Description

2.1. Turbulent ATD scenarios

As a viscous fluid, the atmosphere has two flow states: laminar flow and turbulent flow. Normally, the turbulent flow velocity of the atmosphere is very high. In this case, the atmosphere moves randomly and irregularly and is in the state of turbulent flow state with a continuous increase in flow velocity [44,45]. An atmospheric turbulent field is mainly characterized by the irregular and ever-changing motion of wind flow in a three-dimensional space. This motion is extremely chaotic and has multi-scale characteristics, including vortices from tiny scale to larger scale vortex structures [46].

2.2. STE problem description

Fig. 1 illustrates the STE problem considered under a turbulent ATD scenario. The terrain of the target area in the present study is open and flat with few obstacles. In practice, the fringes of the CIP, *i.e.*, the part of the park bordering the outside area, and the area reserved for the development of the park are mostly flat and with few obstacles and can therefore be represented by the leakage scenarios considered in this study. The problem assumes there are N potential hazardous gas leakage sources with known locations in the target area of a CIP. M sensors are deployed at M different locations in the target area for the purpose of STE. The meteorological condition in terms of wind direction and wind speed is assumed to be known and fixed. A basic assumption of the proposed approach is that the historically collected sensor data can cover IHGLSs of concern. From the practical application perspective, this assumption implies.

(1) There exist regular hazardous gas leakages in a CIP.
(2) Sufficient historical sensor data can be collected from the CIP such that the data can cover the IHGLSs of concern.

According to Refs. [29,38], (1) is the case in many practical scenarios and this is why regular monitoring is often conducted by local governments to make sure overall hazardous gas leakages do not exceed the allowed limit otherwise fines will be imposed [36]. (2) can, in principle, be satisfied if a sufficiently long time is allowed to collect sensor data. This is because the sensor data thus collected will cover all regular hazardous gas leakage scenarios hence the IHGLSs of interest. The proposed approach does not require data from numerous accident cases, including both pollutant concentrations and corresponding leakage strengths, to construct the response matrix under various working conditions. For example, if there are 10 sources releasing hazardous gas, only 10 sets of sensor data representing 10 IHGLSs need to be extracted from the historical data to determine the response matrix.

$H > M$ sets of sensor measurements have been collected over a period, and these data can cover the scenarios where there exist hazardous gas leakages from n actual leakage sources with $n \le N$. Assume that the hazardous gas dispersion from the n actual sources can be detected by m sensors with $4n \le m \le M$ to ensure that there exist sufficient sensor measurements that can be used to determine $4n$ source parameters including $3n$ parameters for source locations and n parameters for leakage strengths. In addition, among the H sets of sensor data, n sets of data cover n IHGLSs as defined in Ref. [29], which can be satisfied by taking a sufficient time for the data collection.

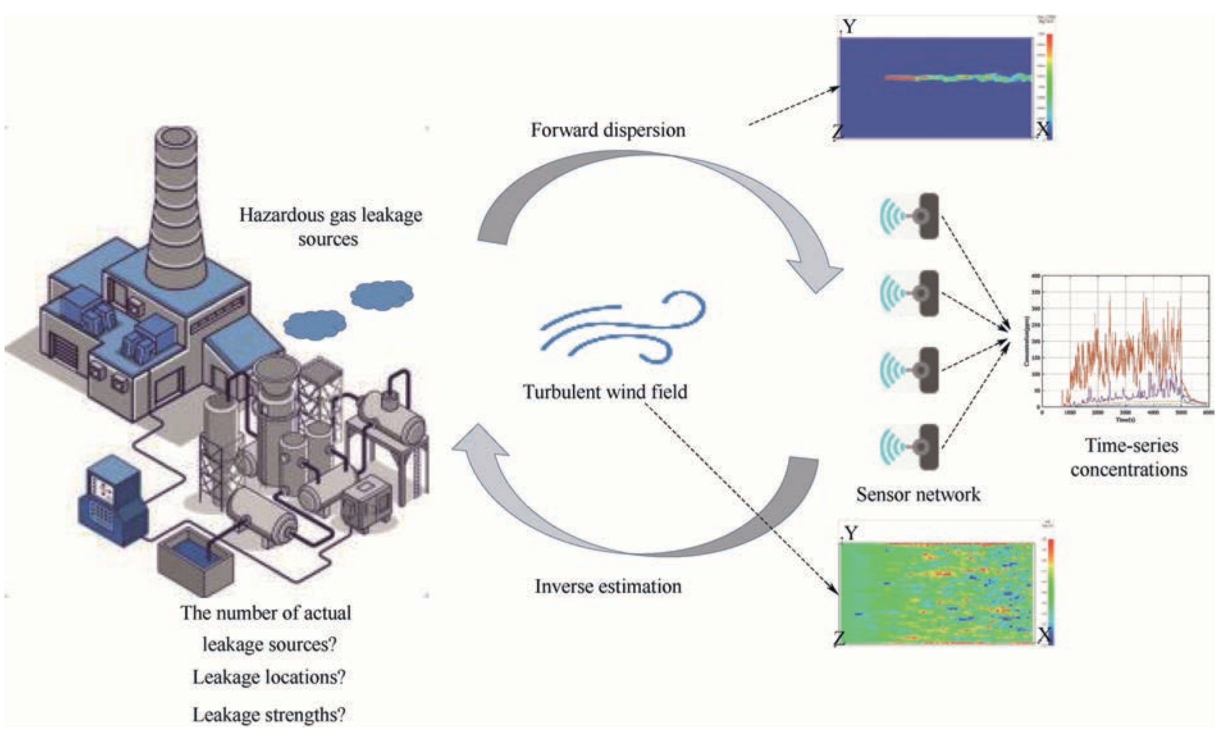

Fig. 1. An illustration of the STE problem under a turbulent ATD scenario.

For non-reactive hazardous gases, the total concentration at a location of CIP can be described by the summation of the concentrations produced by individual sources of hazardous gas [36], expressed by Eq. (1).

$$\boldsymbol{c} = \boldsymbol{A}\boldsymbol{q} \tag{1}$$

where $\boldsymbol{c} = [c_1, c_2, \cdots, c_m]^{\mathrm{T}}$ represents the measurements of the m sensors, $\boldsymbol{q} = [q_1, q_2, \cdots, q_n]^{\mathrm{T}}$ represents the strengths of the n hazardous gas leakage sources, $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n]$ is what is referred to as the response matrix [37], where $\boldsymbol{a}_j = [a_{1,j}, a_{2,j}, \cdots, a_{m,j}]^{\mathrm{T}}, j = 1, 2, \cdots, n$ with $a_{i,j}$ ($i = 1, 2, \cdots, m, j = 1, 2, \cdots, n$) representing how the ith sensor measurement is affected by the strength of the jth leakage source.

The data-driven STE under a turbulent ATD scenario is concerned with off-line STE model building and on-line application of the STE model for real-time STE. The off-line STE model building involves (I) determining the number n of IHGLSs from the H sets of historically collected sensor data, (II) finding the sensor data collected from each of the n IHGLSs, and (III) determining the STE results for each IHGLS using the sensor data obtained in (II), including the locations and strengths of n actual hazardous gas leakage sources. The on-line application is to estimate $\boldsymbol{q}$ in real time using the off-line determined STE model and online-measured $\boldsymbol{c}$. If the estimated $\boldsymbol{q}$ exceeds the allowed limits, the leaking hazardous gases will cause serious harm to the surrounding environment and local residents and the decision-makers in the CIP should take corresponding emergency response actions to minimize the losses.

3. STE of Hazardous Gas Leakages under Turbulent ATD Scenarios

3.1. Data pre-processing

Generally, the time-series concentration data of leaking hazardous gases recorded by sensors in a CIP show strong fluctuations due to turbulent flow. This will increase the difficulties in off-line STE model building and affect the final STE result. Therefore, the raw concentration data are first smoothed by the wavelet threshold de-noising algorithm [47]. Then, the average of the smoothed concentration data over a period when the concentration values have achieved the steady state is taken, and the average value below a certain threshold ($1.00 \times 10^{-6}\ \mathrm{kg \cdot m^{-3}}$) is set to 0 to simulate a more realistic situation.

3.2. Gaussian plume model approximation

3.2.1. CFD modelling

The CFD modelling has now been developed into a mature technology and applied in various fields such as chemical spills, fire, and explosions. In this study, an open-source code, namely FDS, is used to perform CFD modelling and simulate the hazardous gas dispersion under turbulent scenarios. The FDS is created by National Institute of Standards and Technology [48,49].

The FDS allows a large eddy simulation (LES) and even direct numerical simulation (DNS), and is not based on a Reynolds average Navier-Stokes formulation [49]. In this study, the LES is used for turbulence over grids and the Deardorff turbulent viscosity model [50] is used for in-grid vortex. The LES is based on averaging, but instead of averaging all turbulence, only turbulent motions smaller than a given size (*e.g.*, the cell size of the computational grid) are averaged, while any turbulence that can be resolved on the grid is simulated directly [46].

In the present study, the concentrations of leaking hazardous gases at different locations of concern are needed. The procedure for generating the concentration data using FDS is summarized as follows.

Step (i) Build a 3D geometric model of the target area in the CIP of concern using FDS.
Step (ii) Mesh the computational domain after grid independence analysis.
Step (iii) Initialize main simulation parameters (*e.g.*, source location, leakage strength, wind speed and direction, simulation step and time, type of hazardous gas, *etc.*) and boundary conditions.
Step (iv) Start simulating the considered hazardous gas leakage scenario using FDS and export the concentration data of hazardous gases across mesh grids of concern at each time point of the simulation. Fig. 7 shows, for example, the simulated evolution of the concentration of ethane with time at a grid point for a scenario in a CIP with relevant parameters given in Section 4.

3.2.2. Gaussian plume model

Gaussian plume models are widely used for gas leakage and dispersion because it is simple and efficient and can be used for iterative calculations [51]. Under the assumption of spatially homogenous flow over the flat terrain, the general expression for the Gaussian plume model is given as [52]

$$c(x,y,z) = \frac{q}{2\pi v \phi_y \phi_z} \exp\left[-\frac{(y-y_0)^2}{2\phi_y^2}\right]\left\{\exp\left[-\frac{(z-z_0)^2}{2\phi_z^2}\right] + \exp\left[-\frac{(z+z_0)^2}{2\phi_z^2}\right]\right\} \tag{2}$$

where $c(x, y, z)$, in $\mathrm{kg \cdot m^{-3}}$, is the concentration of leaking hazardous gases at (x, y, z), in m; q, in $\mathrm{kg \cdot s^{-1}}$, is the strength of a hazardous gas leakage source; v, in $\mathrm{m \cdot s^{-1}}$, is the uniform wind speed at the height of z, in m; y_0 and z_0, in m, are the coordinates of the leakage source in the crosswind and vertical directions, respectively, and ϕ_y and ϕ_z, in m, represent the standard deviations of a statistically normal plume in the crosswind and vertical directions, respectively, which are functions of the downstream distance between the leakage source and the sensor and come in various forms. In this study, the power-law functions for ϕ_y and ϕ_z suggested in Ref. [53] are adopted as

$$\phi_y = \alpha_1 (x - x_0)^{\alpha_2} \tag{3}$$

$$\phi_z = \alpha_3 (x - x_0)^{\alpha_4} \tag{4}$$

where x_0, in m, is the coordinate of the leakage source in the downwind direction, and α_1, α_2, α_3, and α_4 are dispersion coefficients, which are related to atmospheric stability classes (ASCs) and terrain types.

3.2.3. Dispersion coefficients optimization for approximating CFD simulated concentration data using a Gaussian plume model

As the terrain of the target area in the present study is open and flat with few obstacles, the Gaussian plume model (2) can be used to approximately represent the source-receptor relationship of FDS [51]. This can be achieved by optimizing the dispersion coefficients as defined in Eqs. (3) and (4) using FDS simulated concentration data. In an ASC of concern, the FDS simulations cover different strengths for every one of the N potential leakage sources.

An optimization code can then be used to obtain optimum numerical values of the dispersion parameters that can be used in Gaussian plume models. This allows for quick estimations of hazardous gas dispersion without excessively sacrificing the accuracy of the results.

The dispersion coefficients are optimized by minimizing the sum of squared deviations of the concentration data between the Gaussian plume model and the FDS model, expressed as

$$\widehat{\alpha}_j = \arg\ \min \sum_{i=1}^{b} \left(c_{\mathrm{FDS},i,j} - c_{\mathrm{Gau},i,j}(\boldsymbol{\alpha}) \right)^2, j = 1, 2, \cdots, J \tag{5}$$

where b is the number of concentration sampling locations, satisfying $b \gg m$ to ensure high modelling accuracy, J is the number of simulated scenarios, $c_{\mathrm{FDS},i,j}$ and $c_{\mathrm{Gau},i,j}$ are the concentrations simulated by the FDS and predicted by the Gaussian plume model, respectively, at the ith concentration sampling location in the jth simulated scenario, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]^{\mathrm{T}}$, and $\widehat{\alpha}_j = [\widehat{\alpha}_{1,j}, \widehat{\alpha}_{2,j}, \widehat{\alpha}_{3,j}, \widehat{\alpha}_{4,j}]^{\mathrm{T}}$ represent the optimization results of dispersion coefficients in the jth scenario.

Because of the effect of turbulence, the solutions to the four dispersion coefficients in the J scenarios are all different. Therefore, the average of the optimization results in the J cases is taken as the result, where $\overline{\boldsymbol{\alpha}} = [\overline{\alpha}_1, \overline{\alpha}_2, \overline{\alpha}_3, \overline{\alpha}_4]^{\mathrm{T}}$.

$$\overline{\boldsymbol{\alpha}} = \frac{1}{J} \sum_{j=1}^{J} \widehat{\boldsymbol{\alpha}}_j \tag{6}$$

3.3. Approximating Gaussian plume model-based data-driven STE

The concept of IHGLSs was introduced in Ref. [29], representing linearly independent hazardous gas leakage scenarios. When there are n actual hazardous gas leakage sources, there exist only n leakage scenarios that are linearly independent, *i.e.*, there are n IHGLSs.

Denote $\tilde{\boldsymbol{Q}} = [\tilde{\boldsymbol{q}}_1, \tilde{\boldsymbol{q}}_2, \cdots, \tilde{\boldsymbol{q}}_n]$, where $\tilde{\boldsymbol{q}}_j = [\tilde{q}_{1,j}, \tilde{q}_{2,j}, \cdots, \tilde{q}_{n,j}]^{\mathrm{T}}$ $(j = 1, 2, \cdots, n)$ represent the strengths of hazardous gas leakages in the jth of the n IHGLSs, and denote $\tilde{\boldsymbol{C}} = [\tilde{\boldsymbol{c}}_1, \tilde{\boldsymbol{c}}_2, \cdots, \tilde{\boldsymbol{c}}_n]$, where $\tilde{\boldsymbol{c}}_j = [\tilde{c}_{1,j}, \tilde{c}_{2,j}, \cdots, \tilde{c}_{m,j}]^{\mathrm{T}} (j = 1, 2, \cdots, n)$ represent the corresponding sensor measurements in the jth of the n IHGLSs. It is known from Eq. (1) that

$$\tilde{\boldsymbol{C}} = \boldsymbol{A}\tilde{\boldsymbol{Q}} \tag{7}$$

Since the n column vectors in square matrix $\tilde{\boldsymbol{Q}}$, that is, $[\tilde{q}_{1,j}, \tilde{q}_{2,j}, \cdots, \tilde{q}_{n,j}]^{\mathrm{T}}$, $j = 1, 2, \cdots, n$, are linearly independent, the inverse of matrix $\tilde{\boldsymbol{Q}}$, *i.e.*, $\tilde{\boldsymbol{Q}}^{-1}$, exists. Therefore, the response matrix $\boldsymbol{A}$ can be obtained as

$$\boldsymbol{A} = \tilde{\boldsymbol{C}}\tilde{\boldsymbol{Q}}^{-1} \tag{8}$$

Mathematically, Eq. (1) is uniquely solvable if and only if $\boldsymbol{A}$ has a column rank [54]

$$\mathrm{rank}(\boldsymbol{A}) = n \tag{9}$$

When the number of sensors is not less than that of potential hazardous gas leakage sources, condition (9) is satisfied, and the STE problem (1) is traceable. The leakage strengths of all n sources can be estimated from multi-sensor measurements by the least squares method, which minimizes the summation of squares of estimation error as

$$\widehat{\boldsymbol{q}} = \arg\ \min_{\boldsymbol{q}} \|\boldsymbol{c} - \boldsymbol{A}\boldsymbol{q}\| \tag{10}$$

Singular value decomposition (SVD) is used to solve Eq. (10). The SVD of $\boldsymbol{A}$ is the factorization

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \tag{11}$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are unitary, and $\boldsymbol{D} = \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \in \mathbb{R}^{m\times n}$, where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n)$ with σ_i, $i = 1, 2, \cdots, n$ being non-zero singular values and satisfying $\sigma_1 \geq \sigma_1 \geq \cdots \geq \sigma_n > 0$. Then, the LS estimation of leakage strengths can be obtained as

$$\widehat{\boldsymbol{q}} = \boldsymbol{V}\Sigma^{-1}\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{c} \tag{12}$$

where $\boldsymbol{U}_1$ is the first n columns of $\boldsymbol{U}$, that is, $\boldsymbol{U}_1 = \boldsymbol{U}(:, [1:n])$.

However, the response matrix $\boldsymbol{A}$ is often ill-posed in practice so that the solution (12) is sensitive to even tiny errors in $\boldsymbol{c}$ and becomes unstable. To quantitatively analyze whether the response matrix $\boldsymbol{A}$ is ill-posed or not, the spectral condition number is introduced as

$$\kappa_2(\boldsymbol{A}) = \frac{\sigma_{\max}}{\sigma_{\min}} \tag{13}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the maximum and minimum singular values, respectively. It is known from Eq. (13) that a very small $\sigma_{\min}$ can lead to a very large condition number, which makes the response matrix $\boldsymbol{A}$ be ill-posed. Therefore, the singular value shrinkage operator (SVSO) simply applying a soft-threshold rule [55] is introduced to the singular values σ_i, $i = 1, 2, \cdots, n$ to shrink these toward zero. For each $\tau \geq 0$, the SVSO Θ_τ is defined as

$$\Theta_\tau(\Sigma) = \mathrm{diag}\left(\{\sigma_i - \tau\}_+\right), i = 1, 2, \cdots, n \tag{14}$$

where $\{\sigma_i - \tau\}_+ = \max\{0, \sigma_i - \tau\}$, $i = 1, 2, \cdots, n$. If some of the singular values of $\boldsymbol{A}$ are below the threshold τ, they will be excluded, thereby leading to a lower rank and condition number of $\boldsymbol{A}$ and improving the stability of the LS solution (12).

Consequently, Eq. (12) is implemented as

$$\widehat{\boldsymbol{q}} = \tilde{\boldsymbol{V}}\tilde{\Sigma}^{-1}\tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{c} \tag{15}$$

where $\tilde{\boldsymbol{V}} = \boldsymbol{V}(:, [1:n^*])$ with n^* being the number of retrained singular values after introducing the SVSO, $\tilde{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_{n^*})$, and $\tilde{\boldsymbol{U}} = \boldsymbol{U}(:, [1:n^*])$. Eq. (15) is what is known as the data-driven STE model.

Consequently, performing the data-driven STE requires the following steps.

(a) Find the number of actual leakage sources n.
(b) Find the n IHGLSs and corresponding sensor measurements $\tilde{\boldsymbol{c}}_j, j = 1, 2, \cdots, n$
(c) Determine the strengths of the n leakage sources in the n IHGLSs $\tilde{\boldsymbol{q}}_j, j = 1, 2, \cdots, n$.
(d) Determine the response matrix $\boldsymbol{A}$ and then determine $\tilde{\boldsymbol{V}}$, $\tilde{\Sigma}$, and $\tilde{\boldsymbol{U}}$.
(e) Estimate $\boldsymbol{q}$ in real time for any scenario of concern from online-measured sensor data $\boldsymbol{c}$ using Eq. (15).

The implementation of Step (c) needs a forward ATD model. However, it is extremely difficult to implement Step (c) by directly using the FDS-based CFD model as the forward ATD model. This is because of the need to store the transient wind

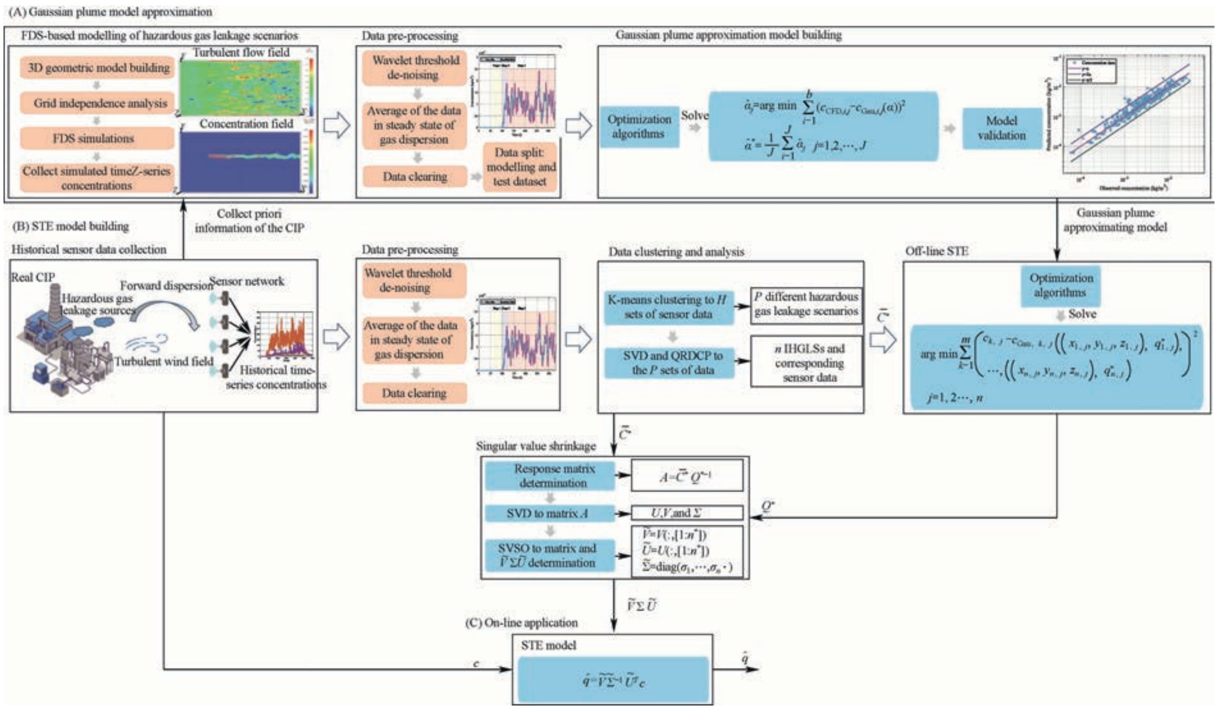

Fig. 2. The framework of the proposed data-driven STE under turbulent ATD scenarios.

fields for all time steps in order to run backward time-stepping simulations using the FDS model, which requires impractically large size of storage space given the small time steps and high grid resolution in the LES method. Therefore, the approximating Gaussian plume model with the dispersion parameters optimized for the same ASC as covered in the IHGLSs is proposed in the present study. To solve the problem, the approximating Gaussian plume model is used as the forward ATD model to predict concentrations at the locations of sensors. The predicted concentrations are then compared with the sensor measurements $\tilde{\boldsymbol{c}}_j (j = 1, 2, \cdots, n)$ in each IHGLS, and the discrepancies are used by a well-established optimization algorithm to search and derive the locations and strengths of hazardous gas leakages that minimize a cost function:

$$f = \arg\ \min \sum_{k=1}^{m} \left(\tilde{c}_{k,j} - c_{\mathrm{Gau},k,j} \left(\left(x_{1,j}, y_{1,j}, z_{1,j} \right), \tilde{q}_{1,j}, \cdots, \left(x_{n,j}, y_{n,j}, z_{n,j} \right), \tilde{q}_{n,j} \right) \right)^2, j = 1, 2, \cdots, n \tag{16}$$

where the subscripts j and k represent the serial numbers of the IHGLS and the sensor, respectively, $(x_{i,j}, y_{i,j}, z_{i,j})$ $(i = 1, 2, \cdots, n, j = 1, 2, \cdots, n)$ representing the location of the ith source in the jth IHGLS, and c_{Gau} is the concentration predicted by the approximating Gaussian plume model. Since there are a total of $4n$ undetermined source parameters including $3n$ parameters for source locations and n parameters for leakage strengths in each IHGLS as expressed by Eq. (16), at least $4n$ sensor measurements are required to make the STE problem (16) traceable [54]. If the number of sensor measurements m is less than $4n$, the STE problem (16) will be partially traceable [36].

3.4. Implementation algorithm

The overall framework of the proposed method is shown in Fig. 2. As shown in Fig. 2, three main steps are used for data-driven STE: (A) Gaussian plume model approximation, (B) data-driven STE model building, and (C) real-time application of the STE model.

(A) Gaussian plume model approximation: This step includes (A.1) FDS-based modelling of hazardous gas leakage scenarios, (A.2) data pre-processing, and (A.3) Gaussian plume approximating model building. Through Steps (A.1) and (A.2), the concentration data in each of the J cases are obtained. In terms Step (A.3), a well-established optimization algorithm, *e.g.*, covariance matrix adaptive evolution strategy (CMA-ES) [56], is applied to solve the optimization problem (5) from the obtained concentration data in each of the J cases. This yields an optimal solution $\hat{\alpha}_j^* = [\hat{\alpha}_{1,j}^*, \hat{\alpha}_{2,j}^*, \hat{\alpha}_{3,j}^*, \hat{\alpha}_{4,j}^*]^{\mathrm{T}}$. Then, taking the average of $\hat{\alpha}_j^*$, $j = 1, 2, \cdots, J$ yields $\overline{\alpha}^* = [\overline{\alpha}_1^*, \overline{\alpha}_2^*, \overline{\alpha}_3^*, \overline{\alpha}_4^*]^{\mathrm{T}}$.

(B) Data-driven STE model building: This step includes (B.1) historical sensor data collection over a sufficiently long time, (B.2) data pre-processing, (B.3) sensor data clustering and analysis, (B.4) off-line STE, and (B.5) singular value shrinkage. Through Steps (B.1) and (B.2), H sets of sensor data $[c_1^j, c_2^j, \cdots, c_m^j], j = 1, 2, \cdots, H$ are obtained. In terms Step (B.3), the K-means clustering algorithm is firstly applied to the H sets of sensor data to divide the data into P clusters covering P different hazardous gas leakage scenarios with

$$\left[c_1^j(p), c_2^j(p), \cdots, c_m^j(p) \right], j = 1, 2, \cdots, g(p), p = 1, 2, \cdots, P \tag{17}$$

$$\sum_{p=1}^{P} g(p) = H \tag{18}$$

Then, generate P sets of representative sensor measurements with and each of the P sets of sensor measurements representing one hazardous gas leakage scenario of concern.

$$[\bar{c}_1(p), \bar{c}_2(p), \cdots, \bar{c}_m(p)],\ p = 1, 2, \cdots, P \tag{19}$$

$$\bar{c}_i(p) = \frac{1}{g(p)} \sum_{j=1}^{g(p)} c_i^j(p),\ i = 1, 2, \cdots, m \tag{20}$$

Subsequently, singular value decomposition (SVD) and QR decomposition with column pivoting (QRDCP) are applied to matrix $\Xi = [\bar{\boldsymbol{c}}_1, \bar{\boldsymbol{c}}_2, \cdots, \bar{\boldsymbol{c}}_m]$, where $\bar{\boldsymbol{c}}_j = [\bar{c}_j(1), \bar{c}_j(2), \cdots, \bar{c}_j(P)]^{\mathrm{T}}$, to determine the number of actual hazardous gas leakage sources, *i.e.*, n, and n sets of sensor measurements that cover n IHGLSs [29,43].

$$\left[\bar{c}_{1,j}^*, \bar{c}_{2,j}^*, \cdots, \bar{c}_{m,j}^*\right]^{\mathrm{T}},\ j = 1, 2, \cdots, n \tag{21}$$

In terms Step (B.4), the approximating Gaussian plume model with the dispersion parameters optimized for the same ASC as in the n IHGLSs is used as the forward ATD model since the approximating model is different in different ASCs, and the CMA-ES strategy is then applied to solve the optimization problem (16) from each of the n sets of sensor data obtained in Step (B.3) producing

$$\left(\hat{x}_{i,j}, \hat{y}_{i,j}, \hat{z}_{i,j}\right),\ i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n \tag{22}$$

and

$$\left[q_{1,j}^*, q_{2,j}^*, \cdots, q_{n,j}^*\right]^{\mathrm{T}},\ j = 1, 2, \cdots, n \tag{23}$$

with $(\hat{x}_{i,j}, \hat{y}_{i,j}, \hat{z}_{i,j})$ representing the estimated location of the ith source in the jth IHGLS and $[q_{1,j}^*, q_{2,j}^*, \cdots, q_{n,j}^*]^{\mathrm{T}}$ representing the estimated strengths of hazardous gas leakages in the jth IHGLS. Then, the average values of estimated source locations for the n IHGLSs are taken as the results.

$$\frac{1}{n}\left(\sum_{j=1}^{n} \hat{x}_{i,j}, \sum_{j=1}^{n} \hat{y}_{i,j}, \sum_{j=1}^{n} \hat{z}_{i,j}\right),\ i = 1, 2, \cdots, n \tag{24}$$

In terms Step (B.5), with the obtained results

$$\overline{\boldsymbol{C}}^* = [\bar{\boldsymbol{c}}_1^*, \bar{\boldsymbol{c}}_2^*, \cdots, \bar{\boldsymbol{c}}_n^*] \tag{25}$$

where $\bar{\boldsymbol{c}}_j^* = [\bar{c}_{1,j}^*, \bar{c}_{2,j}^*, \cdots, \bar{c}_{m,j}^*]^{\mathrm{T}}, j = 1, 2, \cdots, n$, and

$$\boldsymbol{Q}^* = [\boldsymbol{q}_1^*, \boldsymbol{q}_2^*, \cdots, \boldsymbol{q}_n^*] \tag{26}$$

where $\boldsymbol{q}_j^* = [q_{1,j}^*, q_{2,j}^*, \cdots, q_{n,j}^*]^{\mathrm{T}}, j = 1, 2, \cdots, n$, the response matrix can be determined as

$$\boldsymbol{A} = \overline{\boldsymbol{C}}^* \boldsymbol{Q}^{*-1} \tag{27}$$

Applying the SVD to matrix $\boldsymbol{A}$ following Eq. (11) yields $\boldsymbol{U}$, $\boldsymbol{V}$, and Σ. Applying the SVSO to matrix Σ following Eq. (14) yields n^*. Then, $\tilde{\boldsymbol{V}}$, $\tilde{\Sigma}$, and $\tilde{\boldsymbol{U}}$ can be determined as $\tilde{\boldsymbol{V}} = \boldsymbol{V}(:, [1 : n^*])$, $\tilde{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_{n^*})$, and $\tilde{\boldsymbol{U}} = \boldsymbol{U}(:, [1 : n^*])$, respectively.

(C) Real-time application of the STE model: The data-driven STE model (15) is built using $\tilde{\boldsymbol{V}}$, $\tilde{\Sigma}$, and $\tilde{\boldsymbol{U}}$ that are obtained in Step (B.5). $\hat{\boldsymbol{q}}$ is finally determined using the STE model (15) and the online-measured $\boldsymbol{c}$. In principle, for each wind direction, one STE model needs to be built using the proposed approach. The variation in wind speed can be accommodated using the same model so can readily be addressed. To deal with variable wind directions, numerous STE models can be built with each model covering the area of a sector. Thus, the proposed approach can be extended to deal with practical scenarios with different wind speeds and directions. This has been demonstrated in our previous work [43]. To implement Step (C), it is the data-driven STE model corresponding to the on-line meteorological condition that is needed. If the meteorological conditions such as wind directions change over a considerable range, the new data-driven STE approach that enables the incorporation of variable wind directions into the data-driven STE framework can be used to solve the problem [43], provided that along the new wind direction, the fundamental assumption that the hazardous gas dispersion from the n hazardous gas leakage sources can be detected by m sensors with $4n \le m$ satisfied.

In practical applications, to ensure the performance of the resulting response matrix, a STE model validation procedure can be used, which is summarized below.

(1) When determining the average concentration data for each cluster using Eq. (20), the sensor data in each cluster are segmented into STE model training data and STE model validation data. The averages of the STE model training data for all the clusters are used to first produce matrix $\Xi = [\bar{\boldsymbol{c}}_1, \bar{\boldsymbol{c}}_2, \cdots, \bar{\boldsymbol{c}}_m]$, then $[\bar{c}_{1,j}^*, \bar{c}_{2,j}^*, \cdots, \bar{c}_{m,j}^*]^{\mathrm{T}}, j = 1, 2, \cdots, n$, and then off-line STE results $\boldsymbol{q}_j^* = [q_{1,j}^*, q_{2,j}^*, \cdots, q_{n,j}^*]^{\mathrm{T}} (j = 1, 2, \cdots, n)$, and finally the estimated response matrix $\boldsymbol{A}$ using Eq. (27).

(2) The averages of the STE model validation data for the clusters which cover the IHGLS, denoted by $\bar{\boldsymbol{c}}_{\mathrm{v}j}, j = 1, \cdots, n$, are used to validate the performance of the estimated response matrix as follows:

(i) Determine $\hat{\bar{\boldsymbol{c}}}_{\mathrm{v}j} = \boldsymbol{A}\boldsymbol{q}_j^* (j = 1, 2, \cdots, n)$

(ii) Evaluate the difference between $\bar{\boldsymbol{c}}_{\mathrm{v}j}, j = 1, \cdots, n$ and $\hat{\bar{\boldsymbol{c}}}_{\mathrm{v}j}, j = 1, \cdots, n$ as follows

$$Re = \frac{\sum_{j=1}^{n} \left\| \bar{\boldsymbol{c}}_{\mathrm{v}j} - \hat{\bar{\boldsymbol{c}}}_{\mathrm{v}j} \right\|}{n} \tag{28}$$

(iii) Use the result of Eq.(28) to evaluate the performance of the resulting response matrix.

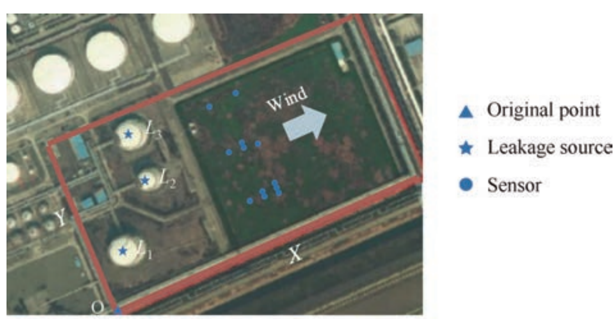

Fig. 3. Layout of the target area in the CIP.

Table 1
Coordinates of 3 potential hazardous gas leakage sources and 18 sensors.

Source & sensor	Coordinate/m	Sensor	Coordinate/m
Source L_1	(39.00, 75.00, 10.00)	Sensor 9	(230.00, 245.50, 10.00)
Source L_2	(115.00, 157.00, 10.00)	Sensor 10	(280.00, 70.00, 10.00)
Source L_3	(115.00, 243.00, 10.00)	Sensor 11	(280.00, 75.00, 10.00)
Sensor 1	(180.00, 75.00, 10.00)	Sensor 12	(280.00, 80.00, 10.00)
Sensor 2	(180.00, 157.00, 10.00)	Sensor 13	(280.00, 152.00, 10.00)
Sensor 3	(180.00, 243.00, 10.00)	Sensor 14	(280.00, 157.00, 10.00)
Sensor 4	(230.00, 72.50, 10.00)	Sensor 15	(280.00, 162.00, 10.00)
Sensor 5	(230.00, 77.50, 10.00)	Sensor 16	(280.00, 238.00, 10.00)
Sensor 6	(230.00, 154.50, 10.00)	Sensor 17	(280.00, 243.00, 10.00)
Sensor 7	(230.00, 159.50, 10.00)	Sensor 18	(280.00, 248.00, 10.00)
Sensor 8	(230.00, 240.50, 10.00)		

In the present study, the performance of the determined response matrix has been validated by the online STE results in Table 8.

The proposed method as implemented using the above Steps (A)–(C) has the following advantages.

- No requirement to know the number and locations of actual hazardous gas leakage sources.
- Extend the basic data-driven STE approach to more practical scenarios where there exists turbulent flow.

4. CFD Simulation

4.1. Layout of the CIP under study

Consider a real CIP in Shanghai, China, with the size of a target area prone to hazardous gas leakages being 500 m (Length) × 260 m (Width) × 20 m (Height). A map of the target area is illustrated in Fig. 3, where the original point of the coordinate system is marked as a blue triangle. As shown in Fig. 3, there are three potential hazardous gas leakage sources L_1, L_2, and L_3 in the target area, which are represented by blue pentagrams. In the present study, only storage tanks are selected as the potential leaking sources. Hazardous gas leakages from components such as valves or pipelines are not considered, which will be the focus of future studies. Assume that the wind direction is fixed and along the direction of positive X axis. 18 sensors represented by blue circles are deployed in the area to monitor the leaking hazardous gases. The coordinates of the 3 potential hazardous gas leakage sources and the 18 sensors are listed in Table 1.

4.2. FDS configuration

A box-shaped computational domain with the size being 500 m (X) × 260 m (Y) × 20 m (Z) is established in the FDS. The leakage source is represented by a square nozzle with a constant jet. The gas leakage strength is chosen empirically from 0 to 1.5 kg·s^{-1}, which basically covers representative small and medium hole leakage scenarios in the CIP. Ethane is selected as the leaking hazardous gas. Each ethane leakage starts at the 60th second when the wind flow covers the computational domain completely, and lasts for 270 s until the end of simulation. The simulation time step is 1 s. The wind speed at the height of 10 m is 3.5 m·s^{-1}. The vertical wind profile is set through the built-in "RAMP" function. The ASC in the target area is 'F'. In FDS, the aerodynamic roughness parameter characterizes the terrain type. In the present study, the aerodynamic roughness is set to 0.03, representing an open

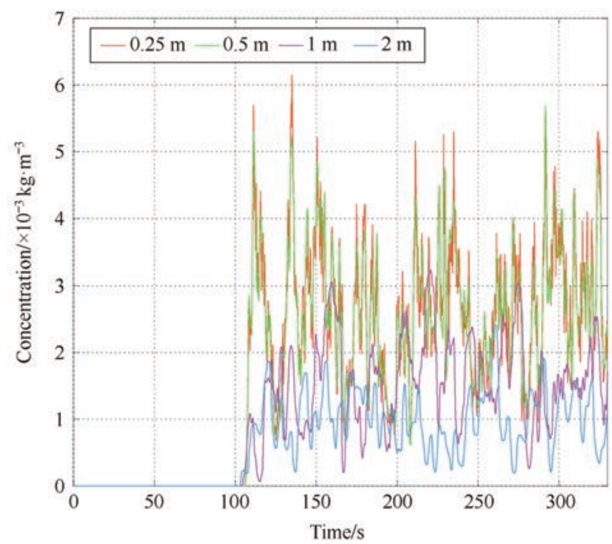

Fig. 5. Grid independence analysis.

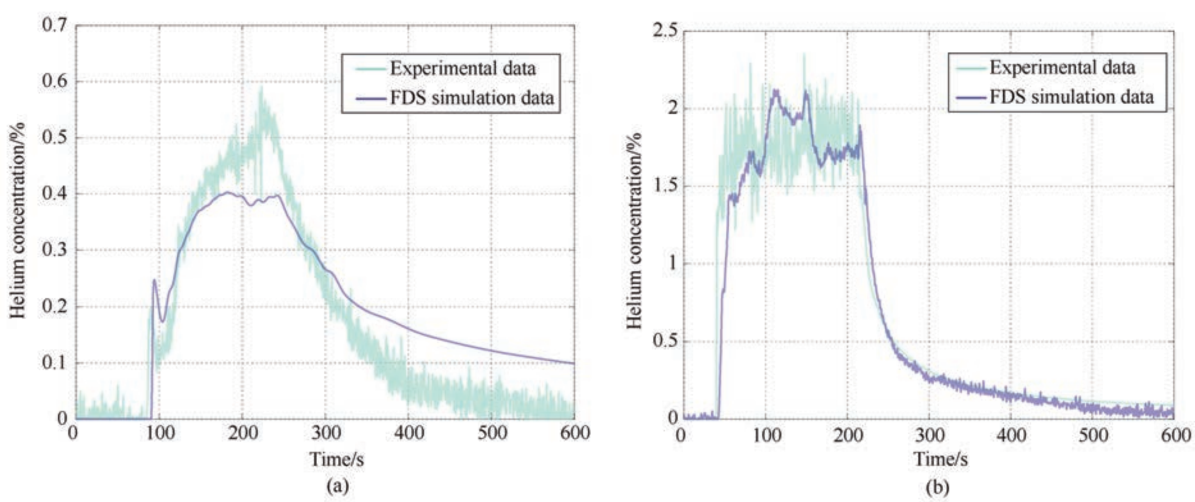

Fig. 4. Comparison between experimental data and FDS simulation data. (a) Sensors IX. (b) Sensor XII.

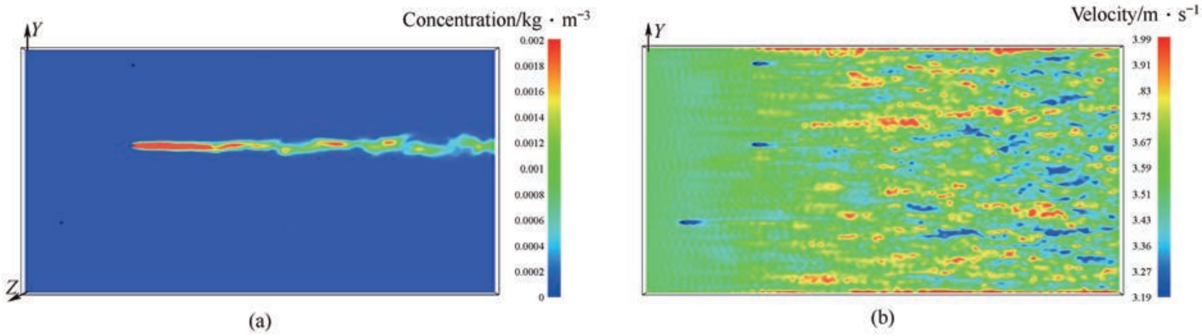

Fig. 6. A top view of the simulated transient: (a) concentration and (b) turbulent flow fields at the 275th second over the spatial plane of $Z = 10$ m.

terrain. The pressure, temperature, humidity, and other meteorological parameters are set to default parameters. For the boundary conditions, the bottom face of the domain is set to "Ground" and all other faces are set to "OPEN".

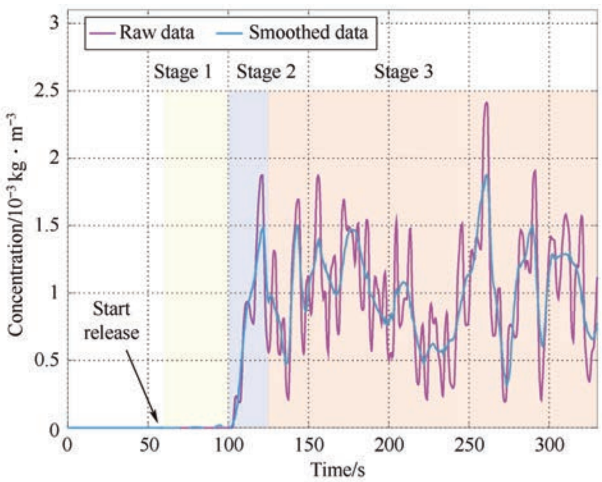

Fig. 7. Comparisons between the raw and smoothed ethane time-series concentration data.

Table 2
The locations and strengths of ethane leakages for the scenarios that are used for Gaussian plume approximation model building and validation.

Modelling scenario No.	Leakage source	Leakage strength /kg·s^{-1}	Test scenario No.	Leakage source	Leakage strength /kg·s^{-1}
1	L_2	0.10	1	L_1	0.30
2	L_2	0.20	2	L_1	0.60
3	L_2	0.40	3	L_1	0.90
4	L_2	0.50	4	L_1	1.20
5	L_2	0.70	5	L_1	1.50
6	L_2	0.80	6	L_2	0.30
7	L_2	1.00	7	L_2	0.60
8	L_2	1.10	8	L_2	0.90
9	L_2	1.30	9	L_2	1.20
10	L_2	1.40	10	L_2	1.50
			11	L_3	0.30
			12	L_3	0.60
			13	L_3	0.90
			14	L_3	1.20
			15	L_3	1.50

4.3. Model validation and grid independence analysis

To ensure the reliability of the FDS model in numerical simulations of hazardous gas dispersion, the FDS model is validated using experimental data on hydrogen leakage in a scale fuel cell vehicle garage [57]. In the experiment, helium leaking from a standard 40 L helium bottle was used as a hydrogen surrogate for safety reasons. The gas flow condition was completely dependent on the natural air currents. 12 Sensors were deployed on the ceiling of the garage model to measure helium concentrations at 12 different locations. The comparison between the numerical simulation results and the experimental results at the locations of sensors IX and XII is presented in Fig. 4, where the ordinate is the volume fraction of helium. The comparison results demonstrate the effectiveness of the FDS model in accurately predicting the hydrogen dispersion in the scaled hydrogen fuel cell vehicle garage. Additionally, since the validity of FDS for hazardous gas leakage and dispersion has been verified in Refs. [23,24,27,51], it is expected that the FDS model for simulating ethane leakage scenarios in the target area would also be representative.

FDS only supports structured cuboid grids [58]. To investigate the effect of grid size on the ethane dispersion, the grid size is set to 0.25 m, 0.5 m, 1 m, and 2 m, respectively. Sensor 14 is selected to record simulated ethane concentrations when the strengths of the leakage source L_2 is set to 0.2 kg·s^{-1}. Corresponding results are shown in Fig. 5. It shows that the concentration time-series tend to stabilize with increasing simulation time for all grid sizes. In addition, the results from the 0.5 m grid deviate significantly from those of the 1 m and 2 m grids but deviate less significantly from those of the 0.25 m grid. To balance the calculation accuracy and time, the 0.5 m grid is chosen for further simulations.

Table 3
Coordinates of 180 spatial locations for recording ethane concentration data simulated by FDS.

	X coordinate/m	Y coordinate/m	Z coordinate/m
60 locations for	165	[71.5: 0.5: 78.5]	10
recording ethane	215	[68: 1: 82]	10
concentrations	265	[64.5: 1.5: 85.5]	10
from source L_1	315	[61: 2: 89]	10
60 locations for	165	[153.5: 0.5: 160.5]	10
recording ethane	215	[150: 1: 164]	10
concentrations	265	[146.5: 1.5: 167.5]	10
from source L_2	315	[143: 2: 171]	10
60 locations for	165	[239.5: 0.5: 246.5]	10
recording ethane	215	[236: 1: 250]	10
concentrations	265	[232.5: 1.5: 253.5]	10
from source L_3	315	[229: 2: 257]	10

Table 4
The locations and strengths of ethane leakages in each of the 40 scenarios that are used to generate historical sensor data for data-driven STE modelling.

Scenario No.	Leakage source	Leakage strength /kg·s^{-1}	Scenario No.	Leakage source	Leakage strength /kg·s^{-1}
1	L_1	0.36	21	L_3	1.46
2	L_2	0.20	22	L_1	1.37
3	L_3	1.48	23	L_2	1.12
4	L_1	0.34	24	L_3	0.34
5	L_2	0.36	25	L_1	0.44
6	L_3	0.07	26	L_2	0.07
7	L_1	0.75	27	L_3	0.33
8	L_2	0.82	28	L_1	0.61
9	L_3	0.09	29	L_2	0.51
10	L_1	0.88	30	L_3	1.27
11	L_2	1.42	31	L_1	1.46
12	L_3	0.70	32	L_2	0.7
13	L_1	0.99	33	L_3	0.47
14	L_2	1.11	34	L_1	0.70
15	L_3	0.75	35	L_2	1.34
16	L_1	0.50	36	L_3	1.24
17	L_2	1.26	37	L_1	0.36
18	L_3	0.49	38	L_2	0.05
19	L_1	1.27	39	L_3	0.33
20	L_2	0.05	40	L_1	0.70

4.4. Simulated ethane leakage scenarios under turbulent flow

Different ethane leakage scenarios are simulated by varying the source location (L_1–L_3) and the leakage strength (0–1.50 kg·s^{-1}). The ethane leakage takes place only from one of the three potential sources in each scenario. For the present study needs, a total of 145 scenarios are simulated, including 10 scenarios for Gaussian plume approximation modelling, 15 scenarios for Gaussian plume approximation model validation, 40 scenarios for data-driven STE modelling, 40 scenarios for STE model test, and 40 scenarios for STE model robustness investigation. In terms of Gaussian plume approximation model building and validation, the locations and strengths of ethane leakages in each of the 25 scenarios are shown in Table 2. Since a large number of simulated ethane concentration data are needed to approximate the Gaussian plume model, 180 spatial locations with coordinates listed in Table 3 are selected as concentration sampling points to record ethane concentration data. In terms of data-driven STE modelling, the locations and strengths of ethane leakages in each of the 40 scenarios are shown in Table 4.

Fig. 6 shows a top view of the simulated transient concentration and turbulent flow fields, respectively, over the spatial plane of $Z = 10$ m, when the strength of the leakage source L_2 is 0.2 kg·s^{-1}. Correspondingly, the time-series concentrations monitored by sensor 14 are shown in Fig. 7. Fig. 7 also shows a comparison between the raw and smoothed concentration data. As shown in Fig. 7, the data oscillations due to turbulent flow are smoothed while preserving the key treads of the curve. Moreover, the ethane dispersion can be divided into three stages. The ethane concentration is zero in Stage 1 as no ethane is detected by sensor due to the long distance between the sensor and leakage source. The ethane concentration increases promptly in Stage 2 and reaches a certain steady level in Stage 3. Consequently, the average of the concentration data from the 150th to the 330th second in Stage 3 is taken as the sensor measurement.

Table 5
Optimal dispersion coefficients of the Gaussian plume model for ethane dispersion in the ASC 'F'.

Dispersion coefficients	α_1	α_2	α_3	α_4
Optimal result	0.4523	0.4839	0.1566	0.4940

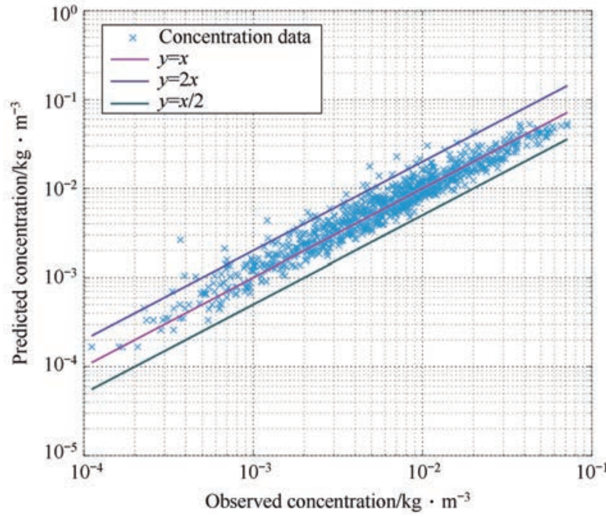

Fig. 8. Comparisons between the concentrations predicted by the approximating Gaussian plume approximation model and the concentrations simulated by FDS.

4.5. Hardware and platform

The FDS package (version 6) is laughed on a server with Intel (R) Gold 6242R CPU @ 3.10 GHz to generate a simulated ethane concentration dataset. It takes about 188 h to simulate all needed scenarios. The algorithm of the proposed approach is implemented *via* MATLAB 2020a on a workstation with Intel (R) Xeon (R) E–2236 CPU @ 3.40 GHz.

5. Results and Discussion

The estimation accuracy of the strengths of ethane leakages is measured by the mean relative error (MRE) and the mean absolute error (MAE), which are derived from Eqs. (29) and (30), respectively.

$$\mathrm{MRE} = \frac{1}{s}\sum_{i=1}^{s}\frac{\left\|\boldsymbol{q}^{(i)} - \widehat{\boldsymbol{q}}^{(i)}\right\|_2}{\left\|\boldsymbol{q}^{(i)}\right\|_2} \times 100\% \tag{29}$$

$$\mathrm{MAE} = \frac{1}{s\cdot n}\sum_{i=1}^{s}\sum_{j=1}^{n}\left|q_j^{(i)} - \widehat{q}_j^{(i)}\right| \tag{30}$$

where s is the number of test scenarios, $\boldsymbol{q}^{(i)}$ and $\widehat{\boldsymbol{q}}^{(i)}$ represent the vectors of real and estimated ethane leakage strengths, respectively, in the ith test scenario, and $q_j^{(i)}$ and $\widehat{q}_j^{(i)}$ represent the real and estimated leakage strengths of the jth source, respectively, in the ith test scenario.

5.1. Results of data-driven STE under turbulent ATD scenarios

The details and results obtained in each of the three steps of the proposed data-driven STE approach for ethane leakages under turbulent ATD scenarios are as follows.

Table 6
Evaluation indicators of the approximating Gaussian plume model.

Evaluation indicator	R^2	FAC2	FB	NMSE
Value	0.9326	0.9889	−0.0007	0.0625

Step (A): CMA-ES is used to search for an optimal solution to the dispersion coefficients of Gaussian plume model. Main parameters of the CMA-ES are set as follows. The initial step size is 0.3. The population size and parent number are 300 and 150, respectively. The initial values of four dispersion parameters are uniformly distributed between 0 and 1. The number of stopping iterations is 1000. For ASC 'F', the optimal dispersion parameters of the Gaussian plume model for ethane dispersion are obtained as shown in Table 5.

The performance of the approximating Gaussian plume model is evaluated by the gap between the predicted and observed values on the test concentration dataset, with a small gap indicating a good performance of the model. Fig. 8 shows a comparison between the concentrations predicted by the approximating Gaussian plume model and the concentrations simulated by FDS. It shows that the predicted concentrations are in good agreement with the observed concentrations. Several statistics are used to quantify the performance of the approximating Gaussian plume model [59]. These statistics are the R square (R^2) given by

$$R^2 = 1 - \frac{\sum_{i=1}^{r}\left(c_{o,i} - c_{p,i}\right)^2}{\sum_{i=1}^{r}\left(c_{o,i} - \bar{\mathbf{c}}_p\right)^2} \tag{31}$$

the fraction of predictions within a factor of two of observations (FAC2) given by

$$\mathrm{FAC2} = \frac{\sum_{i=1}^{r}\xi_i}{r}, \text{s.t.}, \xi_i = \begin{cases} 1, 0.5 \le \dfrac{c_{p,i}}{c_{o,i}} \le 2 \\ 0, \text{otherwise} \end{cases} \tag{32}$$

the fractional bias (FB) given by

$$\mathrm{FB} = \frac{2(\bar{\mathbf{c}}_o - \bar{\mathbf{c}}_p)}{\bar{\mathbf{c}}_o + \bar{\mathbf{c}}_p} \tag{33}$$

and the normalized mean square error (NMSE) given by

$$\mathrm{NMSE} = \frac{\overline{(\mathbf{c}_o - \mathbf{c}_p)^2}}{\overline{\mathbf{c}_o}\,\overline{\mathbf{c}_p}} \tag{34}$$

where r is the number of observations or model predictions, superscript i denotes the ith concentration data, c_o is the concentration simulated by FDS, c_p is the concentration predicted by the approximating Gaussian plume model, $\mathbf{c}_o = [c_{o,1}, c_{o,2}, \cdots, c_{o,r}]^{\mathrm{T}}$, $\mathbf{c}_p = [c_{p,1}, c_{p,2}, \cdots, c_{p,r}]^{\mathrm{T}}$, and overbars denote averages over the sample. The evaluation indicators are summarized in Table 6. As shown in Table 6, the evaluation indicators are satisfactory, as described below. The R^2 is 0.9326, showing a great goodness of fit. The FAC2 is 0.9889, indicating that most predicted concentrations are between the fitting lines $y = 2x$ and $y = x/2$. The FB and NMSE are both near zero, showing the high prediction accuracy of the approximating model.

Step (B):

Step (B.1): Apply the K-means clustering algorithm to $H = 40$ sets of historical sensor data yielding the clustering results shown in Fig. 9. Fig. 9 (a) shows the 40 sets of data are divided into $P = 36$ subgroups, which is consistent with the real situation as these data really cover 36 scenarios with scenarios 1, 20, 27, and 34 being the same as scenarios 37, 38, 39, and 40 as shown in Table 4. Fig. 9 (b) shows the cluster number of each of the 40 sets of data. The average of the concentration data in each subgroup is obtained to produce matrix $\Xi \in \mathbb{R}^{36\times 18}$.

Apply SVD and QRDCP to matrix Ξ yielding $n = 3$ and 3 sets of sensor data

Table 7
Off-line STE results for three IHGLSs.

IHGLS No.	Type	q_{L_1}/kg·s^{-1}	q_{L_2}/kg·s^{-1}	q_{L_3}/kg·s^{-1}
1	Real value	0	0	1.48
	Estimated value	0	0	1.5652
2	Real value	1.37	0	0
	Estimated value	1.2305	0	0
3	Real value	0	1.42	0
	Estimated value	0	1.5107	0

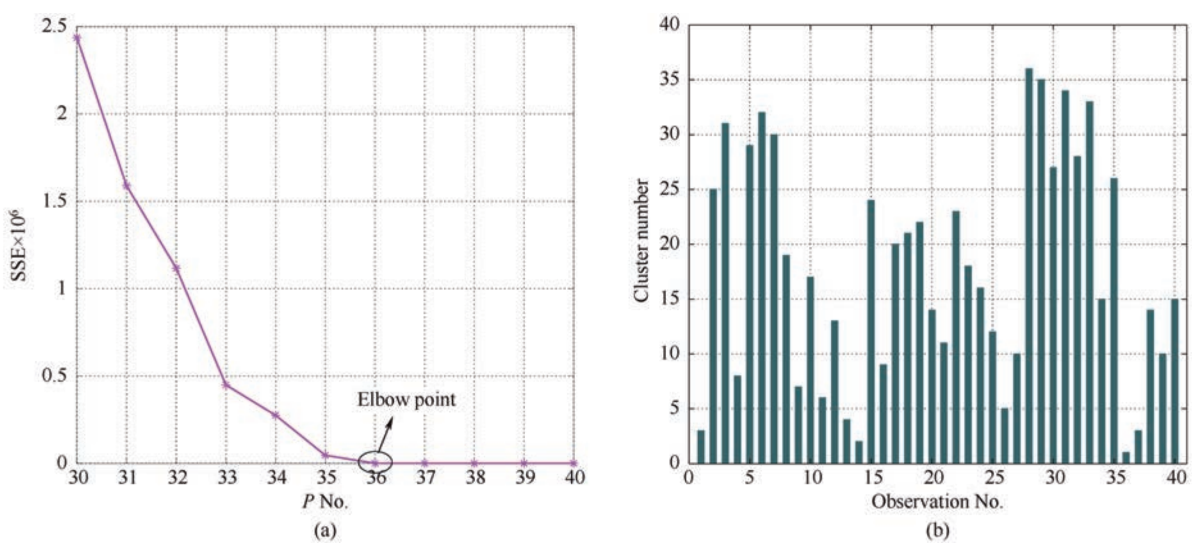

Fig. 9. Clustering results of 40 sets of historical sensor data. (a) Total sum of distance. (b) Cluster number of each dataset.

Table 8
On-line STE results obtained in Step (C).

Test scenario No.	Real strengths of ethane leakages/kg·s^{-1}	Estimated strengths of ethane leakages/kg·s^{-1}	Test scenario No.	Real strengths of ethane leakages/kg·s^{-1}	Estimated strengths of ethane leakages/kg·s^{-1}
1	(0.49, 0, 0)	(0.5192, 0, 0)	21	(0, 0, 0.67)	(0, 0, 0.6197)
2	(0, 0.66, 0)	(0, 0.7091, 0)	22	(0.51, 0, 0)	(0.4636, 0, 0)
3	(0, 0, 0.77)	(0, 0, 0.7388)	23	(0, 0.09, 0)	(0, 0.0938, 0)
4	(0.65, 0, 0)	(0.5406, 0, 0)	24	(0, 0, 0.88)	(0, 0, 0.7947)
5	(0, 0.80, 0)	(0, 0.8161, 0)	25	(1.12, 0, 0)	(0.8269, 0, 0)
6	(0, 0, 1.07)	(0, 0, 1.0738)	26	(0, 0.97, 0)	(0, 1.0148, 0)
7	(0.10, 0, 0)	(0.1028, 0, 0)	27	(0, 0, 1.12)	(0, 0, 1.0009)
8	(0, 1.18, 0)	(0, 1.2063, 0)	28	(0.73, 0, 0)	(0.5979, 0, 0)
9	(0, 0, 0.25)	(0, 0, 0.2449)	29	(0, 0.46, 0)	(0, 0.4366, 0)
10	(0.76, 0, 0)	(0.6582, 0, 0)	30	(0, 0, 0.58)	(0, 0, 0.5952)
11	(0, 0.21, 0)	(0, 0.2063, 0)	31	(0.12, 0, 0)	(0.1261, 0, 0)
12	(0, 0, 0.35)	(0, 0, 0.3661)	32	(0, 0.12, 0)	(0, 0.1306, 0)
13	(0.91, 0, 0)	(0.7774, 0, 0)	33	(0, 0, 0.12)	(0, 0, 0.1264)
14	(0, 0.44, 0)	(0, 0.4305, 0)	34	(0.16, 0, 0)	(0.1546, 0, 0)
15	(0, 0, 0.23)	(0, 0, 0.2344)	35	(0, 0.22, 0)	(0, 0.2340, 0)
16	(0.41, 0, 0)	(0.4302, 0, 0)	36	(0, 0, 0.29)	(0, 0, 0.2930)
17	(0, 0.14, 0)	(0, 0.1419, 0)	37	(0.47, 0, 0)	(0.4549, 0, 0)
18	(0, 0, 0.55)	(0, 0, 0.4857)	38	(0, 0.61, 0)	(0, 0.6115, 0)
19	(0.77, 0, 0)	(0.6490, 0, 0)	39	(0, 0, 0.66)	(0, 0, 0.6833)
20	(0, 0.79, 0)	(0, 0.8208, 0)	40	(0.66, 0, 0)	(0.5557, 0, 0)

$$\Xi(31,:) = [0, 0, 0.0489, 0, 0, 0, 0, 0.0183, 0.0240, 0, 0, 0, 0, 0, 0, 0.0086, 0.0149, 0.0078]$$

$$\Xi(23,:) = [0.0189, 0, 0, 0.0093, 0.0119, 0, 0, 0, 0, 0.0049, 0.0091, 0.0050, 0, 0, 0, 0, 0, 0]$$

$$\Xi(6,:) = [0, 0.0463, 0, 0, 0, 0.0217, 0.0179, 0, 0, 0, 0, 0, 0.0112, 0.0147, 0.0042, 0, 0, 0]$$

that cover 3 IHGLSs to produce matrix $\overline{\boldsymbol{C}}^* = \begin{bmatrix} \Xi(31,:) \\ \Xi(23,:) \\ \Xi(6,:) \end{bmatrix}^{\mathrm{T}}$.

Step (B.2): As the locations of $N = 3$ potential leakage sources are known and $n = 3$, the source locations in each IHGLS do not need to be re-estimated. The Gaussian plume model (2) is applied offline with $v = 3.5\ \mathrm{m \cdot s^{-1}}$, $\alpha_1 = 0.4523$, $\alpha_2 = 0.4839$, $\alpha_3 = 0.1566$, and $\alpha_4 = 0.4940$ as the forward ATD model to determine the strengths of the three actual leakage sources in each of the three IHGLSs determined in Step (B.1). The CMA-ES is used to search for an optimal solution to the strengths of ethane leakages in each IHGLS. The results are shown in Table 7, which are used to produce matrix $\boldsymbol{Q}^* = \begin{bmatrix} 0 & 1.2305 & 0 \\ 0 & 0 & 1.5107 \\ 1.5652 & 0 & 0 \end{bmatrix}$.

Step (B.3): With the obtained results $\overline{\boldsymbol{C}}^*$ and $\boldsymbol{Q}^*$, the response matrix $\boldsymbol{A}$ can be determined following Eq. (27). Applying the SVD to matrix $\boldsymbol{A}$ following Eq. (11) yields $\boldsymbol{U}$, $\boldsymbol{V}$, and Σ. With $\tau = 0.02$, applying the SVSO to matrix Σ following Eq. (14) yields $n* = 3$. Then, $\tilde{\boldsymbol{V}}$, $\tilde{\Sigma}$, and $\tilde{\boldsymbol{U}}$ can be determined as $\tilde{\boldsymbol{V}} = \boldsymbol{V}(:, [1:3])$, $\tilde{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3)$, and $\tilde{\boldsymbol{U}} = \boldsymbol{U}(:, [1:3])$, respectively.

Step (C): The data-driven STE model is established with the obtained results $\tilde{\boldsymbol{V}}$, $\tilde{\Sigma}$, and $\tilde{\boldsymbol{U}}$. Then, the model is applied to the sensor data from 40 test scenarios, respectively, to determine the strengths of ethane leakages in each case. The real and estimated strengths of ethane leakages in each test scenario are listed in Table 8. The maximum and minimum relative errors are 26.17% and 0.25% in scenarios 25 and 38, respectively. The maximum and minimum absolute errors are 0.2931 and 0.0015 in scenarios 25 and 38, respectively. The MRE and MAE are 6.76% and 0.0149, respectively, indicating a satisfactory STE accuracy. In addition, the accuracy of the leakage strength estimation for sources L_2 and L_3 is better than that for source L_1. This is mainly due to the large error in the estimation of the leakage strength of source L_1 in IHGLS 2 as shown in Table 7. The leakage location where the estimated leakage strength is not zero are regarded as the estimated source location. For example, it is known from the estimated leakage strengths (0.5192, 0, 0), in kg·s^{-1}, in the test scenario 1 that the estimated leakage location is L_1. As shown in Table 7, source locations are correctly estimated for all 40 test scenarios.

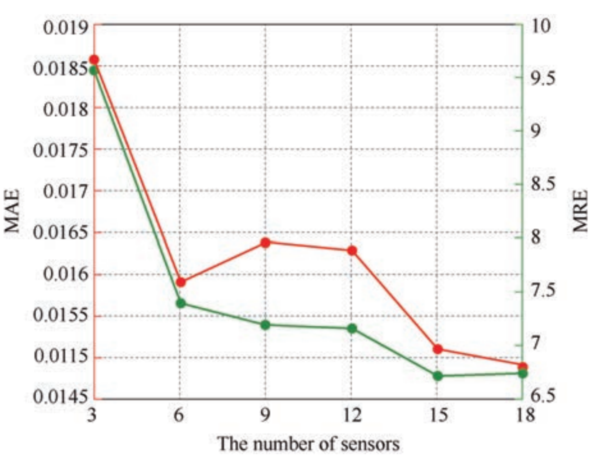

Fig. 10. Variations of the MRE and MAE with the number of sensors.

5.2. Effect of the number of sensors

To investigate the effect of the number of sensors on the performance of STE, the number of sensors is set to 3, 6, 9, 12, and 15, respectively, to build the data-driven STE model (15), and the built model is then used to online process the measurements from the corresponding sensors and perform STE. The variations of the MRE and MAE with the number of sensors are shown in Fig. 10. It shows that the performance of STE is essentially better with more sensors. Moreover, the MRE and MAE do not change greatly after the

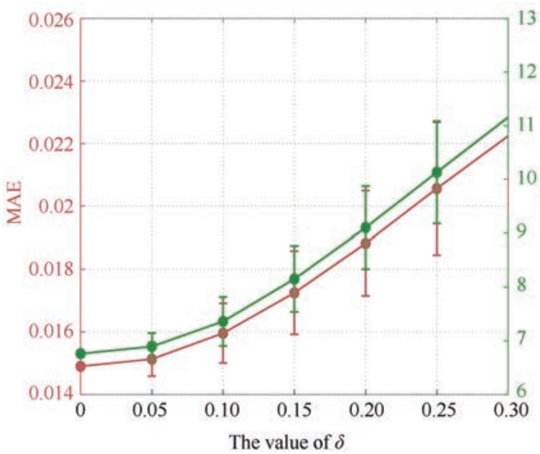

Fig. 11. Variations of the mean and standard deviation of evaluation indicators in the 2000 MCSs with δ.

number of sensors has been increased to 6, indicating 6 sensors are sufficient for this case.

5.3. Effect of measurement errors of online-measured sensor data

To investigate the effect of errors of online-measured sensor data on the performance of STE, the sensor measurements in each of the 40 test scenarios are corrupted with random noise such that

$$c_j^{(i)} \leftarrow c_j^{(i)}(1+\xi),\ i=1,2,\cdots,40,\ j=1,2,\cdots,18 \tag{35}$$

where $c_j^{(i)}$ is the concentration at the jth sensor location in the ith test scenario, and ξ is the noise intensity that is assumed to follow the uniform distribution $\xi \sim U(-\delta,\delta)$ with δ representing the maximum noise intensity. Then, the STE model obtained in Step (C) of Subsection 5.1 is used to process the noise-corrupted sensor measurements and perform STE. Here, δ is set to 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30, respectively. For each δ, 2000 Monte Carlo simulations (MCSs) with random noise added to the sensor measurements are conducted. The variations of the mean and standard deviation of the evaluation indicators in the 2000 MCSs of δ are shown in Fig. 11. It can be observed that the two line charts show a steady upward trend, indicating as δ increases, the performance of STE is worse. Additionally, as δ increases, the standard deviation of the STE results gradually increases, implying that the STE becomes unstable. Nevertheless, the slopes of the two line charts are not significant and the performance of STE is satisfactory when δ is not greater than 0.10, showing the proposed method has a degree of immunity to interference.

5.4. STE results for simultaneous leakages of ethane from multiple sources

In practice, simultaneous leakages of ethane from multiple sources in the CIP is inevitable. Therefore, it is necessary to evaluate the performance of STE in the case of simultaneous leakages of ethane from multiple sources. The STE model obtained in Step (C) of Subsection 5.1 is applied to the sensor data in 40 different ethane leakage scenarios, respectively, to perform STE in each case. The real and estimated strengths of ethane leakages in each of the 40 scenarios are shown in Table 9. Fig. 12 shows the STE results of leakage strengths on the 40 sets of test data. As shown in Fig. 12, apart from a few cases, most estimated leakage strengths are close to true values. Even in these few cases, the differences between the estimated and true leakage strengths are not significant. The MRE and MAE are 8.0336% and 0.0321, respectively. In addition, the accuracy of the leakage strength estimation for sources L_2 and L_3 is better than that for source L_1. The leakage locations where the estimated leakage strengths are not zero are regarded as the estimated source locations. For example, it is known from the estimated leakage strengths (0.2382, 0.4152, 0), in $\mathrm{kg \cdot s^{-1}}$, in the test scenario 1 that the estimated leakage locations are L_1 and L_2. Fig. 13 shows the STE results of leakage locations from the 40 sets of test data. As shown in Fig. 13, source locations are correctly estimated for all 40 test scenarios. These demonstrate that the established STE model can effectively process the sensor data to carry out STE when there exist simultaneous leakages from multiple leakage sources.

Table 9
STE results for simultaneous leakages of ethane from multiple sources.

Scenario No.	Real strengths of ethane leakages/$\mathrm{kg \cdot s^{-1}}$	Estimated strengths of ethane leakages/$\mathrm{kg \cdot s^{-1}}$	Scenario No.	Real strengths of ethane leakages/$\mathrm{kg \cdot s^{-1}}$	Estimated strengths of ethane leakages/$\mathrm{kg \cdot s^{-1}}$
1	(0.22, 0.39, 0)	(0.2382, 0.4152, 0)	21	(1.05, 1.46, 0.39)	(0.7798, 1.5043, 0.3960)
2	(0, 0.18, 0.08)	(0, 0.1773, 0.0768)	22	(0.38, 0.32, 0)	(0.3920, 0.3235, 0)
3	(0.38, 0, 0.65)	(0.3804, 0, 0.6211)	23	(0.70, 0.89, 0)	(0.5344, 0.8087, 0)
4	(0.49, 0, 0.71)	(0.4471, 0, 0.6359)	24	(1.07, 1.27, 0)	(0.8739, 1.3588, 0)
5	(0, 0.29, 0.33)	(0, 0.3269, 0.3290)	25	(0.60, 0.98, 0)	(0.5168, 1.0078, 0)
6	(0.20, 0.37, 0.56)	(0.2452, 0.3523, 0.5538)	26	(0, 0.22, 1.25)	(0, 0.2238, 1.2577)
7	(0.82, 0.05, 0)	(0.7798, 0.0491, 0)	27	(0, 1.10, 0.15)	(0, 1.0527, 0.1479)
8	(0.59, 0.61, 0)	(0.6157, 0.6433, 0)	28	(0.32, 0.47, 0.63)	(0.2945, 0.4832, 0.6437)
9	(0.20, 0.10, 0)	(0.1988, 0.0999, 0)	29	(0.88, 0.20, 0.06)	(0.7246, 0.2077, 0.0613)
10	(0, 0.33, 0.33)	(0, 0.3128, 0.3334)	30	(0.25, 0.25, 0.25)	(0.2582, 0.2417, 0.2302)
11	(0.36, 0.64, 0.11)	(0.3183, 0.5998, 0.1121)	31	(0.62, 0.35, 0.19)	(0.5568, 0.3771, 0.1916)
12	(0.50, 0.50, 0.07)	(0.4736, 0.4951, 0.0691)	32	(0.82, 0.56, 0.19)	(0.7680, 0.5747, 0.1930)
13	(0.71, 1.20, 0)	(0.6515, 1.1982, 0)	33	(0.73, 0, 0.47)	(0.6057, 0, 0.4765)
14	(1.11, 0.20, 0)	(0.8178, 0.1967, 0)	34	(0.64, 0.81, 0)	(0.6005, 0.7780, 0)
15	(0.82, 0, 0.47)	(0.7377, 0, 0.4603)	35	(0, 0.36, 0.10)	(0, 0.3643, 0.0969)
16	(0.49, 0.91, 0)	(0.4447, 0.9319, 0)	36	(0.80, 0.40, 0.20)	(0.7117, 0.4107, 0.2047)
17	(1.11, 0, 0.16)	(0.8323, 0, 0.1803)	37	(0.57, 0.91, 0)	(0.5319, 0.8193, 0)
18	(0.28, 0, 0.26)	(0.2771, 0, 0.2724)	38	(1.00, 1.02, 0)	(0.7330, 0.9964, 0)
19	(0.33, 0.49, 0)	(0.3379, 0.4974, 0)	39	(0.11, 0.51, 0)	(0.0987, 0.4693, 0)
20	(0, 1.20, 0.55)	(0, 1.2728, 0.5187)	40	(0.66, 0.27, 0.66)	(0.5524, 0.2650, 0.6541)

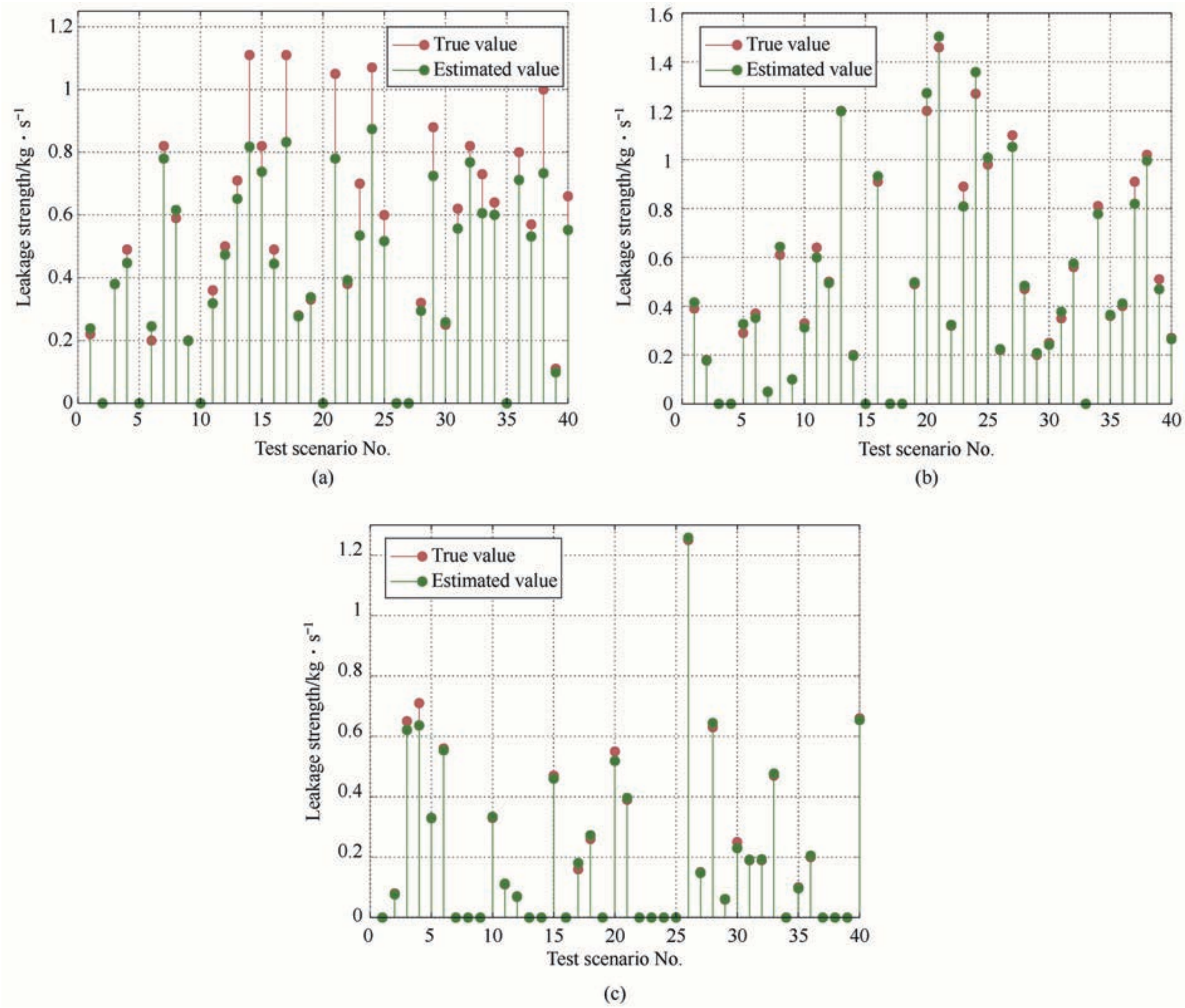

Fig. 12. STE results of leakage strengths from 40 sets of test data. (a) Source 1. (b) Source 2. (c) Source 3.

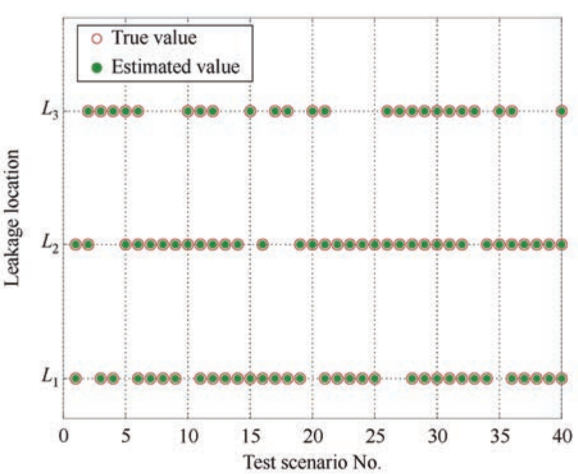

Fig. 13. STE results of leakage locations from 40 sets of test data.

5.5. Application to the real data from Prairie Grass field dispersion experiment

Prairie Grass field dispersion experiments were conducted in O'Neill, Nebraska [60]. The field experiment dataset consists of 68 releases of about 10 min each of trace gas SO_2 from a ground level point source. About 600 mean concentration monitoring sensors were deployed at a height of 1.5 m along five arcs with radii of

Table 10
Optimal dispersion parameters of the Gaussian plume model in the ASC 'D' [62].

ASC	α_1	α_2	α_3	α_4
D	0.32	0.78	0.22	0.78

Table 11
Off-line STE results for the IHGLS.

Type	x_0/m	y_0/m	z_0/m	q/g·s^{-1}
Real value	0	0	0.46	95.9
Estimated value	−3.4716	0.4399	2.7652	105.4467

Table 12
On-line STE results of two test cases of the Prairie Grass filed dispersion experiments.

Test No.	Real leakage strength/g·s^{-1}	Estimated leakage strength/g·s^{-1}
1	89.50	76.5185
2	50.90	51.5870

50 m, 100 m, 200 m, 400 m, and 800 m, measuring 10 min average concentrations. The sensors were deployed at intervals of 2° along four arcs of 50–400 m, and the interval of 1° at 800 m. A large number of meteorological measurements were collected during the experiments. The atmospheric stability class (ASC) associated with each data set is dependent on the time of the release, and the experiments were conducted under different ASCs. Characteristics of the 68 Prairie Grass field experiments have been reported in Ref. [61].

Data preparation: Four sets of experimental data from releases 6, 11, 21, and 33 are used in the present study, which have similar but different meteorological conditions in terms of ASC, wind direction, and wind speed. The ASC in the four experiments is 'D'. The data from releases 11 and 33 are used as historically collected sensor data to build the data-driven STE model, and the remaining data are used as online-measured sensor data to evaluate the performance of the proposed method.

Gaussian plume model approximation: For the Prairie Grass field dispersion experiments, the optimal dispersion parameters as defined in Eqs. (3) and (4) for the Gaussian plume model had been determined in Ref. [62] for different ASCs. The optimal dispersion parameters in the ASC 'D' are summarized in Table 10.

Data-driven STE model building: Apply the K-means clustering to the $H = 2$ sets of historical sensor data producing matrix $\Xi \in \mathbb{R}^{2\times 58}$. Apply SVD and QRDCP to matrix Ξ yielding $n = 1$ and one set of representative sensor data $\Xi(1,:)$ that cover one IHGLS to produce vector $\overline{\boldsymbol{C}}^* = [\Xi(1,:)\ ]^{\mathrm{T}}$. The Gaussian plume model (2) is applied offline with $v = 6.77\ \mathrm{m\cdot s^{-1}}$, $\alpha_1 = 0.32$, $\alpha_2 = 0.78$, $\alpha_3 = 0.22$, and $\alpha_4 = 0.78$ as the forward ATD model to determine the location and strength of the point leakage source in the IHGLS. The CMA-ES is used to search for an optimal solution to the location and strength of the point leakage source. The results are given in Table 11 showing that overall the results are good estimations of the true values. Consequently, we can use $\boldsymbol{Q}^* = 105.4467$.

On-line application: The data-driven STE model is established with the obtained results $\overline{\boldsymbol{C}}^*$ and $\boldsymbol{Q}^*$. Then, the model is applied to process the sensor data from the two test cases, obtaining an estimated leakage strength in each case. The real and estimated leakage strengths at the point leakage source in each case are listed in Table 12. The MRE and MAE are 7.9271% and 6.8343, respectively, indicating a satisfactory STE accuracy even if the meteorological conditions in terms of wind direction and wind speed in the modelling scenarios are different from those in the test scenarios. This demonstrates the practical applicability of the proposed approach.

6. Conclusions

In the present study, a new MSDD-STE approach and associated implementation algorithm that can work under turbulent ATD scenarios are developed. An approximating Gaussian plume model is produced from the CFD model simulated data and used to represent hazardous gas dispersions under turbulent ATD scenarios. This overcomes the difficulties with the off-line STE when applying existing MSDD-STE approaches in turbulent ATD scenarios. Comprehensive numerical simulation studies are conducted to verify the effectiveness and demonstrate the practical significance of the proposed approach. Analysis of the effects of various factors including sensor numbers, measurement errors, and leakage conditions on the STE performance is also conducted, demonstrating the robustness of the proposed approach. Although the proposed method assumes that the terrain of the target area is open and flat, the basic principle of the proposed method can readily be extended to much more general scenarios where there exist complex obstacles. The basic idea is to determine the source-receptor relationship in the more complex scenarios by constructing the unsteady adjoint equation modelled by the LES technique and apply this adjoint equation to perform the off-line STE. Then, the STE problems in the more complex scenario of concern can be solved using the basic MSDD-STE approach. In addition, only one wind direction is considered in the present study. Future studies will be focused on more complicated scenarios where meteorological conditions, such as, *e.g.*, wind directions may change.

CRediT Authorship Contribution Statement

Chuantao Ni: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Ziqiang Lang: Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. Bing Wang: Project administration, Investigation. Ang Li: Investigation. Chenxi Cao: Project administration, Investigation. Wenli Du: Resources, Project administration, Funding acquisition. Feng Qian: Resources, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (Basic Science Center Program 61988101, 62303186, 62203173).

References

[1] S.H. Yang, J.M. Chen, Air pollution prevention and pollution source identification of chemical industrial parks, *Process Saf. Environ. Prot.* 159 (2022) 992–995.
[2] Y.M. Shou, J.Y. Chen, X.X. Guo, J.P. Zhu, L. Ding, J. Ji, Y.F. Cheng, A dynamic individual risk management method considering spatial and temporal synergistic effect of toxic substance leakage and fire accidents, Process, *Saf. Environ. Prot.* 169 (2023) 238–251.
[3] Z.Q. Chen, L. Ye, B. Dai, F. Wang, Feasibility analysis of a single-sensor-based approach for source identification of hazardous chemical releases, *Chin. J. Chem. Eng.* 27 (7) (2019) 1642–1650.
[4] Y.S. Ling, C.F. Liu, Q. Shan, D.Q. Hei, X.J. Zhang, C. Shi, W.B. Jia, J. Wang, Inversion method for multiple nuclide source terms in nuclear accidents based on deep learning fusion model, *Atmosphere* 14 (1) (2023) 148.
[5] C. Rhodes, C.J. Liu, P. Westoby, W.H. Chen, Autonomous search of an airborne release in urban environments using informed tree planning, *Auton. Robots* 47 (1) (2023) 1–18.
[6] H.Y. Jia, H. Kikumoto, Sensor configuration optimization based on the entropy of adjoint concentration distribution for stochastic source

term estimation in urban environment, *Sustain. Cities Soc.* 79 (2022) 103726.
[7] F.Y. Wang, X.Y. Zhou, H. Kikumoto, Improvement of optimization methods in indoor time-variant source parameters estimation combining unsteady adjoint equations and flow field information, *Build. Environ.* 226 (2022) 109710.
[8] J.T. Cai, J.S. Wu, S.Q. Yuan, G. Reniers, Y.P. Bai, Risk-based optimization of emergency response systems for accidental gas leakage in utility tunnels, *Reliab. Eng. Syst. Saf.* 244 (2024) 109947.
[9] N. Gunawardena, K.K. Leang, E. Pardyjak, Particle swarm optimization for source localization in realistic complex urban environments, *Atmos. Environ.* 262 (2021) 118636.
[10] Q.L. He, L.J. Zhou, F. Zhang, D.J. Guan, X. Zhang, Efficient and accurate leakage points detection in gas pipeline using reinforcement learning-based optimization, *IEEE Sens. J.* 24 (17) (2024) 27640–27652.
[11] S. Jang, J. Park, H.H. Lee, C.S. Jin, E.S. Kim, Comparative study on gradient-free optimization methods for inverse source-term estimation of radioactive dispersion from nuclear accidents, *J. Hazard. Mater.* 461 (2024) 132519.
[12] F.Y. Wang, X.Y. Zhou, J. Huang, H.D. Wang, H. Kikumoto, C.Y. Deng, Natural gas leakage estimation in underground utility tunnels using Bayesian inference based on flow fields with gas jet disturbance, *Process Saf. Environ. Prot.* 165 (2022) 532–544.
[13] H.L. Zhang, B. Li, J. Shang, W.W. Wang, F.Y. Zhao, Source term estimation for continuous plume dispersion in Fusion Field Trial-07: Bayesian inference probability adjoint inverse method, *Sci. Total Environ.* 915 (2024) 169802.
[14] J.P. Zhao, J.L. Li, Y.L. Bai, W.J. Zhou, Y.H. Zhang, J.J. Wei, Research on leakage detection technology of natural gas pipeline based on modified Gaussian plume model and Markov chain Monte Carlo method, Process, *Saf. Environ. Prot.* 182 (2024) 314–326.
[15] D.L. Ma, Z.X. Zhang, Contaminant dispersion prediction and source estimation with integrated Gaussian-machine learning network model for point source emission in atmosphere, *J. Hazard. Mater.* 311 (2016) 237–245.
[16] D.L. Ma, J.M. Gao, Z.X. Zhang, H. Zhao, Identifying atmospheric pollutant sources using a machine learning dispersion model and Markov chain Monte Carlo methods, *Stoch. Environ. Res. Risk Assess.* 35 (2) (2021) 271–286.
[17] D.L. Ma, J.Q. Deng, Z.X. Zhang, Comparison and improvements of optimization methods for gas emission source identification, *Atmos. Environ.* 81 (2013) 188–198.
[18] J.L. Wang, B. Wang, J.J. Liu, W. Cheng, J.P. Zhang, An inverse method to estimate the source term of atmospheric pollutant releases, *Atmos. Environ.* 260 (2021) 118554.
[19] W.J. Cui, B. Cao, Q.X. Fan, J. Fan, Y.X. Chen, Source term inversion of nuclear accident based on deep feedforward neural network, *Ann. Nucl. Energy* 175 (2022) 109257.
[20] S. Feintuch, J. Tabrikian, I. Bilik, H. Permuter, Neural-network-based DOA estimation in the presence of non-Gaussian interference, *IEEE Trans. Aero. Electron. Syst.* 60 (1) (2024) 119–132.
[21] Y. Su, J.F. Li, B. Yu, Y.L. Zhao, J. Yao, Fast and accurate prediction of failure pressure of oil and gas defective pipelines using the deep learning model, *Reliab. Eng. Syst. Saf.* 216 (2021) 108016.
[22] S.K. Chen, W.L. Du, X. Peng, C.X. Cao, X.J. Wang, B. Wang, Peripheric sensors-based leaking source tracking in a chemical industrial park with complex obstacles, *J. Loss Prev. Process. Ind.* 78 (2022) 104828.
[23] J.J. Xu, W.L. Du, Q.Y. Xu, J.K. Dong, B. Wang, Federated learning based atmospheric source term estimation in urban environments, *Comput. Chem. Eng.* 155 (2021) 107505.
[24] Q.Y. Xu, W.L. Du, J.J. Xu, J.K. Dong, Neural network-based source tracking of chemical leaks with obstacles, *Chin. J. Chem. Eng.* 33 (2021) 211–220.
[25] X.Q. Zhang, J.H. Shi, M. Yang, X.Y. Huang, A.S. Usmani, G.M. Chen, J.M. Fu, J.W. Huang, J.J. Li, Real-time pipeline leak detection and localization using an attention-based LSTM approach, *Process Saf. Environ. Prot.* 174 (2023) 460–472.
[26] F.M.M. Sousa, A.Z. Selvaggio, F.V. Silva, S.S.V. Vianna, Leakage source localisation employing 3D-CFD simulations and gated recurrent units, *Process Saf. Environ. Prot.* 178 (2023) 540–546.
[27] A. Li, Z.Q. Lang, C.T. Ni, H. Tian, B. Wang, C.X. Cao, W.L. Du, F. Qian, Deep learning-based source term estimation of hydrogen leakages from a hydrogen fueled gas turbine, *Int. J. Hydrogen Energy* 86 (2024) 875–889.
[28] A.Z. Selvaggio, F.M.M. Sousa, F.V. da Silva, S.S.V. Vianna, Application of long short-term memory recurrent neural networks for localisation of leak source using 3D computational fluid dynamics, *Process Saf. Environ. Prot.* 159 (2022) 757–767.
[29] Z.Q. Lang, B. Wang, Y.T. Wang, C.X. Cao, X. Peng, W.L. Du, F. Qian, A novel multi-sensor data-driven approach to source term estimation of hazardous gas leakages in the chemical industry, *Processes* 10 (8) (2022) 1633.
[30] W.X. Qian, H. Gao, Y.Y. Lu, S. Lyu, L. Zhuang, S.T. Hu, L.X. Wang, J. Liu, Optimizing measurement schemes to improve indoor airflow and temperature CFD–EnKF joint simulation, *Build. Environ.* 248 (2024) 111070.
[31] J.S. Wu, J.T. Cai, S.Q. Yuan, X.L. Zhang, G. Reniers, CFD and EnKF coupling estimation of LNG leakage and dispersion, *Saf. Sci.* 139 (2021) 105263.
[32] J.B. Wang, J.S. Zhao, X.H. Lei, H. Wang, An effective method for point pollution source identification in rivers with performance-improved ensemble Kalman filter, *J. Hydrol.* 577 (2019) 123991.
[33] J.T. Cai, J.S. Wu, S.Q. Yuan, D.S. Kong, X.L. Zhang, Prediction of gas leakage and dispersion in utility tunnels based on CFD-EnKF coupling model: a 3D full-scale application, *Sustain. Cities Soc.* 80 (2022) 103789.
[34] J.S. Wu, S.Q. Yuan, C. Zhang, X.L. Zhang, Numerical estimation of gas release and dispersion in coal mine using Ensemble Kalman Filter, *J. Loss Prev. Process. Ind.* 56 (2018) 57–67.
[35] S.Q. Yuan, J.S. Wu, X.L. Zhang, W.Y. Liu, EnKF-based estimation of natural gas release and dispersion in an underground tunnel, *J. Loss Prev. Process. Ind.* 62 (2019) 103931.
[36] X.Y. Zhao, K. Cheng, W. Zhou, Y. Cao, S.H. Yang, J.M. Chen, Source term estimation with deficient sensors: a temporal augment approach, *Process Saf. Environ. Prot.* 157 (2022) 131–139.
[37] X.R. Liu, F. Li, H. Cai, K. Zhang, J.X. Liu, J.H. Xu, X.T. Li, Dynamical source term estimation in a multi-compartment building under time-varying airflow, *Build. Environ.* 160 (2019) 106162.
[38] W. Zhou, X.Y. Zhao, K. Cheng, Y. Cao, S.H. Yang, J.M. Chen, Source term estimation with deficient sensors: error analysis and mobile station route design, *Process Saf. Environ. Prot.* 154 (2021) 97–103.
[39] Y.H. Xu, X.W. Dong, H.Y. Luo, S. Fang, Robust source reconstruction of atmospheric radionuclides from observations of different sparsity with spatial preselection and non-smooth constraints, *J. Hazard. Mater.* 486 (2025) 136919.
[40] S.B. Tang, F. Li, Y.H. Han, Z.B. Feng, A lightweight adjoint method for gaseous pollution source term estimation in urban environments, *Build. Simulat.* 18 (4) (2025) 937–955.
[41] T.F. Zhang, S. Yin, S.G. Wang, An inverse method based on CFD to quantify the temporal release rate of a continuously released pollutant source, *Atmos. Environ.* 77 (2013) 62–77.
[42] J.Y. Zhuang, F. Li, X.R. Liu, H. Cai, L.H. Feng, X.T. Li, An experiment-based impulse response method to characterize airborne pollutant sources in a scaled multi-zone building, *Atmos. Environ.* 251 (2021) 118272.
[43] C.T. Ni, Z.Q. Lang, B. Wang, A. Li, C.X. Cao, W.L. Du, F. Qian, Data-driven source term estimation of hazardous gas leakages under variable meteorological conditions, *J. Loss Prev. Process. Ind.* 94 (2025) 105506.
[44] A.Y. Shikhovtsev, E.A. Kopylov, Structure of atmospheric turbulence, *Atmosphere* 13 (7) (2022) 1107.
[45] F.Z. Wang, W.H. Du, Q. Yuan, D.S. Liu, S. Feng, A survey of structure of atmospheric turbulence in atmosphere and related turbulent effects, *Atmosphere* 12 (12) (2021) 1608.
[46] A. Visscher, Air Dispersion Modeling: Foundations and Applications, John Wiley & Sons, Inc., Hoboken, 2014.
[47] W. Zhang, J.P. Zhao, P. Quan, J.W. Wang, X.Y. Meng, Q. Li, Prediction of influent wastewater quality based on wavelet transform and residual LSTM, *Appl. Soft Comput.* 148 (2023) 110858.
[48] K. McGrattan, S. Hostikka, J. Floyd, et al., Fire Dynamics Simulator: User's Guide, National Institute of Standards and Technology, U.S. Department of Commerce, 2013.
[49] Y. Mouilleau, A. Champassith, CFD simulations of atmospheric gas dispersion using the Fire Dynamics Simulator (FDS), *J. Loss Prev. Process. Ind.* 22 (3) (2009) 316–323.
[50] Y. Kitamura, Improving a turbulence scheme for the terra incognita in a dry convective boundary layer, *J. Meteor. Soc. Jpn. Ser II* 94 (6) (2016) 491–506.
[51] J.K. Dong, W.L. Du, B. Wang, C.X. Cao, S.K. Chen, Q.Y. Xu, Impact analysis of multi-sensor layout on the source term estimation of hazardous gas leakage, *J. Loss Prev. Process. Ind.* 73 (2021) 104579.
[52] J.K. Dong, B. Wang, X.J. Wang, C.X. Cao, S.K. Chen, W.L. Du, Optimization of sensor deployment sequences for hazardous gas leakage monitoring and source term estimation, *Chin. J. Chem. Eng.* 56 (2023) 169–179.
[53] S.R. Hanna, G.A. Briggs, R.P.J. Hosker, Handbook on Atmospheric Diffusion, U. S. Department of Energy, 1982.
[54] K. Cheng, X.Y. Zhao, W. Zhou, Y. Cao, S.H. Yang, J.M. Chen, Source term estimation with deficient sensors: traceability and an equivalent source approach, *Process Saf. Environ. Prot.* 152 (2021) 131–139.
[55] J.F. Cai, E.J. Candès, Z.W. Shen, A singular value thresholding algorithm for matrix completion, *SIAM J. Optim* 20 (4) (2010) 1956–1982.
[56] N. Hansen, A. Ostermeier, Completely derandomized self-adaptation in evolution strategies, *Evol. Comput.* 9 (2) (2001) 159–195.
[57] M.B. Zhao, T. Huang, C.H. Liu, M.J. Chen, S. Ji, D.M. Christopher, X.F. Li, Leak localization using distributed sensors and machine learning for hydrogen releases from a fuel cell vehicle in a parking garage, *Int. J. Hydrogen Energy* 46 (1) (2021) 1420–1433.
[58] B. Wang, F. Qian, W.M. Zhong, Wind field reconstruction for the dispersion modeling of accidental chemical spills on complex geometry, *Chin. J. Chem. Eng.* 27 (11) (2019) 2712–2724.

[59] J.C. Chang, S.R. Hanna, Air quality model performance evaluation, *Meteor. Atmos. Phys.* 87 (1) (2004) 167–196.
[60] M.L. Barad, Project Prairie Grass, a Field Program in Diffusion, Tach. Rep. Geophysical Research Paper, No. 59, Report AFCRC-TR-58-235, Air Force Cambridge Research Center, pp 218.
[61] G. Cervone, P. Franzese, Non-Darwinian evolution for the source detection of atmospheric releases, *Atmos. Environ.* 45 (26) (2011) 4497–4506.
[62] S.S. Mao, J.L. Lang, T. Chen, S.Y. Cheng, J.X. Cui, Z.Y. Shen, F. Hu, Comparison of the impacts of empirical power-law dispersion schemes on simulations of pollutant dispersion during different atmospheric conditions, *Atmos. Environ.* 224 (2020) 117317.