博士論文

Study on Statistical Inference for Unknown Sources of Atmospheric Pollutants in Urban Environment

(都市環境における未知空気汚染発生源の確率的推定に関する研究)

賈　鴻源

Abstract of dissertation

This dissertation studies the statistical inference method for an unknown source of atmospheric pollutants in the complicated urban environment based on Bayesian inference. To enable the inference method to handle the characteristics of the urban area, the main works of the dissertation include: extending the inference's ability to estimate the geometry of the unknown source; increasing the estimation accuracy by introducing a sophisticated dispersion model into the inference; proposing a sensor configuration optimization method to improve the quality of measurements and guide the sensor network design.

Nowadays, a high level of urbanization results in a considerable amount of people gathering in the urban area, where a safe and healthy atmospheric environment has never been more important. However, unexpected atmospheric pollution emitted from unknown sources sometimes occurred because of terrorism, nuclear accidents, illegal industrial emission, and other emergencies, which is a serious threat to humankind and the earth's environment. Therefore, it is important to identify the unknown source as soon as possible after these dispersion emergencies happened. Until now, intensive research has proposed various methods to realize source term estimation (STE). The basic framework of STE is using the estimation algorithm to find the true source based on the measurements obtained during the emergency and the source-receptor relationship (S-RR), which is the concentration prediction modeled in advance. However, the existing methods are still not capable enough of handling complex scenarios in the urban environment because of its unique features as follows.

First of all, possible sources of pollutants are diverse in the urban environment. Although the previous research has considered mobile sources and multiple sources, their estimation algorithm always assumed that the unknown source is an ideal point without geometry. Actually, some sources have non-negligible shapes or volumes, which may fail the estimation based on the point assumption. It is important to extend the STE method for geometry estimation.

Besides, the dispersion mechanism of atmospheric pollutants is extremely complicated in the urban environment. The transportation of pollutants involves complex turbulence caused by buildings, equipment, moving objects, and heat discontinuity. However, the existing simulation techniques for S-RR are still limited in steady models and cannot accurately predict turbulent dispersion. Since STE relies on S-RR as the prior prediction of pollution dispersion, it is necessary to introduce a precise modeling technique for the S-RR in order to promise the accuracy of STE.

What's more, ideal concentration measurements are difficult to acquire in cities because of the dense building distributions and land property limitations. Sensor configurations used in real life are nearly random or empirical, but their effectiveness to all possible sources has no promise. Because the performance of STE is highly dependent on the quality of measurements, it is meaningful to develop a sensor configuration optimization method to guide the sensor deployment and provide high-quality measurements.

Based on this research background, this dissertation aims to develop a statistical inference method for STE in the urban environment by dealing with 3 unsolved problems mentioned above. Because the complexity of real atmospheric pollution is out of the range of a single dissertation, the research subject is limited to one single source with constant emission strength and fixed location in a statistically steady turbulent flow field in a neighborhood-scaled urban area. The main research contents are summarized below.

As a beginning, a new method was proposed to estimate the geometry of unknown sources based on the super-Gaussian function. The coefficients of this function can control its distribution to approximate common shapes: line, rectangular, and ellipse. These coefficients are added into Bayesian inference to realize the geometry estimation. The applicability of the proposed method was first confirmed using a numerical experiment of an ideal boundary layer. The method successfully inferred that the source is line-like without any prior knowledge. Based on this case, the effects of different sensor configurations on the line source estimation were discussed. Because the line source contained more geometric information than point sources, the conventional sensor configuration for the point source may fail in the line source estimation. It was found that the requirements on the sensor configuration become higher. Both the sensors near the source and null-measurement sensors are indispensable.

To examine the robustness of the proposed method against measurement and modeling errors, the second case of a simplified urban square with wind tunnel experiment measurements was conducted. The line source was successfully identified by the proposed method again. By comparing to the conventional STE method with ideal point assumption, it is confirmed that the proposed method can not only provide the geometry estimation but also reduce the inference errors caused by the point source assumption. Hence, it is important to include the geometry estimation when the geometry of the source has unignorable effects on the dispersion.

Then, to improve the accuracy of STE in complex urban applications, large-eddy simulation (LES) was introduced to model the S-RR by unsteady adjoint equations. The LES of adjoint equations has rarely been conducted in the literature because the adjoint equation describes an inverse dispersion process. The time-series flow field data of the entire domain must be produced by forward simulation and stored in advance, thus the volume of data simulated with LES is too large for practical applications. This research proposed to use the wavelet-based compression method to mitigate the storage pressure. The LES flow field can be compressed into a portable database to make the simulation of unsteady adjoint equations easier.

As the first step, to evaluate the accuracy of compression and usefulness of the compressed database, a turbulent flow field in a block-arrayed building group model was simulated by LES and compressed into a database by the wavelet-based compression method. The influence of compression on the quality of the data was checked from the perspectives of a single snapshot and time series. It was confirmed that about 100 times compression can still satisfy the requirement of flow field visualization and afterward simulation. Large-scaled turbulent structures were well preserved after compression, and the dispersion simulation can be reliably reproduced with compression data. Therefore, it is reasonable to expect that the unsteady simulation of adjoint equations can be realized based on the compression database.

Afterward, the compression database above was used in the LES of adjoint equations to model S-RR, which was combined with Bayesian inference as a new STE method. The concentration measurements obtained from a wind tunnel experiment were applied to testify the performance of the proposed method. As a comparison, another STE was also conducted with a conventional method, where steady adjoint equations were simulated with the Reynolds-averaged Navier-Stokes (RANS) model. The results showed that the modeling of S-RR and the

accuracy of STE were significantly improved by the LES of the adjoint equation. The complicated turbulent flows caused by buildings destroyed the reliability of the conventional RANS model. Although the proposed method needs more computational resources, to effectively perform STE in the complicated urban environment, it is valuable to apply LES for adjoint equation simulation.

At last, a sensor configuration optimization method for STE was proposed by the design of an objective function and application of the simulated annealing method. The objective function was set as the information entropy of the spatial distribution of the adjoint concentration field. Its ability to represent the measurement ability of sensor configurations was proved from the views of mathematics and physical meanings. Simulated annealing was applied to find the optimal configuration which owns the largest value of the objective function. The proposed method was utilized to design an optimal sensor configuration for the block-arrayed building group model. The performance of the optimal configuration in STE was compared to uniform and random configurations through estimations for 25 unknown sources. The results revealed that the accuracy of STE is related to the entropy contained in the adjoint concentration of the configuration such that the design of the objective function is reliable. The optimal configuration outperforms the other two in STEs. It is valuable to use the proposed method to guide the configuration design in real applications.

Table of contents

Abbreviations

CDF9/7 : Cohen–Daubechies–Feauveau 9/7

CFD : computational fluid dynamics

DNS : direct numerical simulation

GDH : gradient diffusion hypothesis

LES : large-eddy simulation

MCMC : Monte-Carlo Markov chain

MHMC : Metropolis-Hastings within Gibbs algorithm

PDF : probability density function

PSD : power spectrum density

RANS : Reynolds-averaged Navier-Stokes

RHS : right hand side

SA : simulated annealing

SCO : sensor configuration optimization

SGF : super-Gaussian function

SGS : sub-grid scale

S-RR : source-receptor relationship

STE : source term estimation

UAV : unmanned aerial vehicle

WCM : wavelet-based compression method

WD : wavelet decomposition

WTE : wind tunnel experiment

Chapter 1
Introduction

Abstract

This dissertation studies about statistical estimation method for an unknown source of atmospheric pollutants in the complex urban environment based on Bayesian inference. This chapter is an introduction of the dissertation including the research background, objective, and its structure.

1.1 Research Background

1.1.1 Atmospheric pollution in the urban environment

With the quick development of urbanization, the population is shifting from rural to urban areas. Until 2020, more than half of people in the world, over 4 billion, are living in urban areas (**Fig. 1.1**). Under this circumstance, it is important to keep a healthy atmospheric environment for such a massive population in urban areas. Meanwhile, the situation that massive population gathers in limited urban lands with extremely complicated infrastructures just causes frequent air pollution.

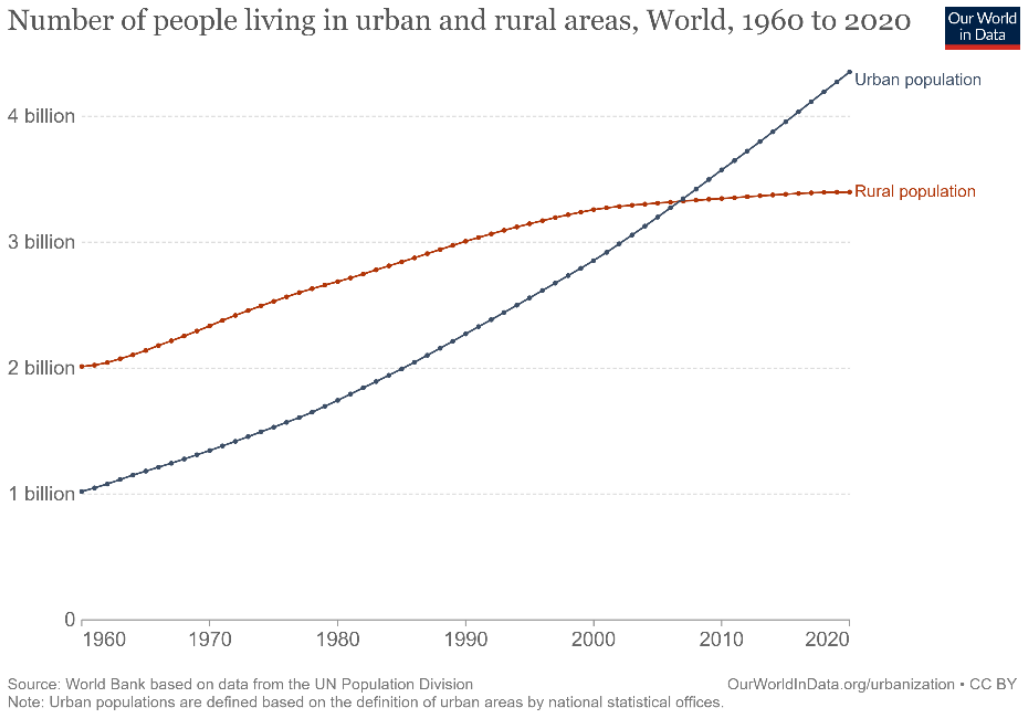


Figure 1.1 The trend of the number of people living in urban and rural areas from 1960 to 2020. (Data source: https://ourworldindata.org/urbanization)

Some air pollutants are well-known by people due to media promotion. For example, carbon dioxide is a famous greenhouse pollutant for its leading effects on global warming. Sulfur oxides produced by various industrial processes often cause consideration environmental damages like acid rain. With the popularity of auto-mobile, nitrogen oxides get the public's attention and become one of the main factors to the atmospheric pollution. The sources of these pollutants are known but are numerous or even infinite. They exist all around people in daily life. In recent years, with international consensus, people are dealing with these pollutants by

technological innovation, legal restriction, and pollution treatment. It has been reported that most pollutions are improving due to these actions.

In contrast, another kind of pollution has different characteristics: (1) the source is single or finite but unknown; (2) the source does not commonly exist in the atmosphere but suddenly appears due to accidents and emergencies. This pollution is caused by the unexpected dispersion of atmospheric pollutants. Although the occurrence is relatively low, people still need a comprehensive countermeasure system to handle it owing to its high level of danger.

1.1.2 Unexpected dispersion of atmospheric pollutants

Unexpected dispersion of atmospheric pollutants jeopardizes the safety of humankind and the earth's environment. These kinds of emergencies can be caused by terrorism. On 20 March 1995, the most severe terrorism attack in Japan, the Tokyo subway sarin incident happened in the central districts of the city. The toxic gas killed 14 people and injured over 6000 people to varying degrees. Nuclear accidents also caused harmful unexpected dispersion. From 11 March 2011 to 15 March 2011, the explosion of 4 reaction units in the Fukushima Daiichi Nuclear Power Plant released a large amount of radioactive contamination into the atmosphere and ocean. The harm caused by this dispersion accident is immeasurable and is still ongoing. Besides, illegal industrial emission also needs special attention. In the spring of 2019, Seoul suffered from its worst-ever atmospheric pollution that the concentration of $PM_{2.5}$ reached almost 4 times the South Korean standard. The reason was found to be the illegal industrial emission of LG Chem and Hanwha Chemical, which has continued for over 4 years until the investigation. Apart from these remarkable emergencies, relatively small-scaled accidents like gas leaks and fire in the common life may also result in severe damage and require attention.

One of the most dangerous properties of unexpected dispersion is that the source of pollutants is unknown at the beginning. People cannot evaluate the emergency or take effective countermeasures until they gain enough information about the source, which often costs considerable time and allows the exacerbation of the emergency. Therefore, it is important to identify the unknown source as soon as possible after the dispersion emergencies happened.

For the realization of this target, the available knowledge generally includes discrete concentration measurements provided by limited sensors, meteorological conditions like wind speed and direction, and geographic information of the incident site. The problem that

identifying the information of the unknown source by utilizing this information is called source term estimation (STE), which is an ill-posed inverse problem with high nonlinearity, strong dependence on input data, and non-single solutions.

1.1.3 Source term estimation

In recent years, intensive research efforts have been devoted to proposing solutions for STE (Hutchinson et al., 2017). Existed methods can be divided into 2 directions by the type of sensors they rely on: mobile sensors and stationary sensors. The development of robotics and unmanned aerial vehicles (UAVs) attributes to the appearance of STE with mobile sensors. People can deploy several sensors after the emergency, and sensors can automatically track down the boundaries or sources of the contaminants. This concept was inspired by biological behaviors like swarms of beetles. The mobile sensors can communicate and cooperate to realize the route planning during STE (Zarzhitsky and Spears, 2005). However, the main problem of this direction is the immature techniques of robots and UAVs, which still have a long way to go before real applications. Besides, its applicability is also doubtful considering the strict regulation of UAVs in urban areas.

The other direction, stationary sensor, is more mature and practical. Many cities already deployed measurement stations at different positions as civil infrastructures. They can provide the basic measurement: the time-averaged concentration of a period. Some stations can also provide advanced information like the time series of concentration and eddy covariance (concentration flux). As a result, more STE methods are based on stationary sensors than mobile ones, and this dissertation also focuses on the STE with stationary sensors as below.

1.1.4 Elements in source term estimation

There are three basic elements in the STE: measurement, source-receptor relationship (S-RR), and estimation algorithm. Measurement is the foundation of STE and is provided by sensors. It dominantly affects the accuracy of STE and is dependent on the sensing strategies, equipment, and techniques, which means sensing in the real application is a complicated task and inevitably involved uncertainties. In the previous research, uniform and random sensors configuration are commonly used to evaluate the performance of STE methods.

S-RR describes the relationship between a source and a sensor. It predicts the concentration measurement of one sensor when a source with certain strength appears

somewhere before the emergency by computational methods. Popular methods include the Gaussian plume model, adjoint equation model, and Markov chain model. Because knowledge of sources before the emergency is unavailable, as a prior database, S-RR should be able to count in a large number of possible sources, which is a high requirement for the calculation cost of methods. Meanwhile, since accurate S-RR is critical to a successful STE, the prediction accuracy of computational methods should be promised. It should be able to capture the influence of meteorological conditions and terrain characteristics on the dispersion.

Gaussian plume model and Markov chain model satisfy the first requirement. By assuming an ideal dispersion in a homogeneous flow field, the Gaussian plume model predicts the concentration of a sensor with an algebraic equation, which is easy and fast to calculate. However, its reliability is vulnerable to real dispersion fields like buildings immersed into a flow. Markov chain model assumes that movements of pollutants among different zones follow a probability matrix, which can be calculated by computational fluid dynamics (CFD) simulation. Although CFD can bring some inhomogeneous information of flow field, to keep the calculation speed, the probability matrix is set as unchanged, which means the turbulent dispersion cannot be properly predicted. Therefore, neither the Gaussian plume model nor the Markov chain model can maintain the prediction accuracy when it comes to complex urban applications.

Considering the main research interest of the thesis is STE in the urban environment, the adjoint equation model is used here to calculate S-RR. This model indeed needs more calculation resources than the other two models, but its prediction accuracy is much higher because the dispersion process is described as partial differential equations, which can theoretically deal with turbulent dispersion. With the help of CFD techniques, adjoint equations can be simulated to predict S-RR for complicated urban flow fields.

The estimation algorithm finds the most credible source as the estimation result based on the measurements and S-RR. Because the estimation algorithm is executed right after the emergency, both the speed and accuracy are important. Existing methods can be classified into two frameworks: optimization and statistics.

The optimization framework applies the objective function to evaluate each candidate, find a candidate that minimizes the value of the objective function and regards it as a unique solution. Different designs of the objective function and plenty of algorithms for minimization

have been developed. Optimization methods are often less computationally expensive than the statistical framework, and their convergence speed is faster. Despite that, sometimes their performance is highly dependent on the initial guess, which may cause strong unsteadiness. More importantly, the estimation is limited to a single solution. In real applications, the inevitable measurements errors in the sensing and numerical errors of S-RR modeling degrade the reliability of the single solution. The optimization may be misled by errors, yield a wrong estimation and the truth is ignored.

In contrast, the statistical framework calculates the probability of each candidate to produce a probability distribution, and consequently, the noise of measurements and simulations can be reflected. In this case, more information like confidence intervals will be provided, and the truth is more likely to be noticed. Most of the statistical methods are based on Bayesian inference, where the posterior probability can be estimated based on the prior information and measurements. Because Bayesian inference can reflect the influence of noise on the estimation results, it is probable to apply it to the STE in the complex urban environment. Therefore, the Bayesian inference is selected as the basic estimation algorithm of STE in the thesis.

1.1.5 Difficulty of source estimation in the urban environment

Although plenty of STE methods, which combine different choices for the above three elements, have been proposed, it is still difficult to claim that the existed methods can handle complex reality problems in the urban environment. In fact, there is still large room for further development of STE because of the features of the urban environment as follows.

First, possible sources of pollutants are diverse in the urban environment. They can be a point or shaped, fixed or moving, single or multiple. In the beginning, the estimation algorithm can only deal with a point source at a fixed position. Afterward, advanced algorithms were proposed to manage moving sources and multiple sources. However, little research is concerned about the estimation of geometry information of the source, which is also valuable for risk management and evaluation of seriousness.

Furthermore, the complicated turbulence field and dispersion phenomenon in the urban environment brought a challenge to the accuracy of S-RR simulation. The turbulent complexity of the atmospheric boundary layer is increased further by buildings, equipment, transportation

systems, and heat discontinuity after it reached the urban area. Therefore, modeling the turbulence flow field and pollutants dispersion in the urban environment is troublesome. In recent years, with the development of computers and numerical simulation techniques, sophisticated CFD models like the large eddy simulation are proposed to reproduce unsteady turbulence fields. However, models for S-RR are still limited that the most accurate solution till now is the steady Reynolds-averaged Navier-Stokes, which may fail to accurately predict the dynamic properties of dispersion behavior. Thus, it is necessary to upgrade the simulation method of S-RR and improve the STE performance.

Last but not least, ideal concentration measurements are difficult to acquire in the urban environment. Even though the existed STE methods perform well with uniform sensor configurations in the previous research, it is impractical to deploy a uniform sensor network in the urban area due to the irregular building distributions and land property. Actually, measurement stations in reality are closer to a random configuration, which is a consequence caused by check and balance of deployment difficulty and sensing efficiency. However, considering that unknown sources may appear anywhere in the city, a random configuration may cause considerable errors in the measurements of certain sources. It is meaningful to develop a sensor configuration optimization method to guide the sensor deployment. Thus, no matter where the unknown source of emergencies appears in the target area, the optimum sensor configuration can provide high-quality concentration measurements that help STE effectively identify the source, and consequently is an important part of the atmospheric pollution detection system.

1.2 Research objective

According to the introduction above, in order to protect the atmospheric quality and strengthen the defense ability of urban areas against dispersion emergencies caused by unknown sources, it is necessary to improve the performance of the current STE for the urban environment. It should be noted that this target cannot be accomplished only by one research given the extreme complexity of the urban environment. As one step of this process, this research mainly tried to realize three terms as follows:

a) Extend the estimation algorithm’s ability to estimate the geometry information of common

sources in the urban area: point, line, rectangular, and ellipse.

b) Apply large eddy simulation in the S-RR simulation and evaluate its effectiveness in STE.

c) Develop a sensor configuration optimization method independent of the prior knowledge of unknown sources.

Concentrating on the statistical STE problem, the above three improvements are made based on three elements as shown in **Fig. 1.2**.

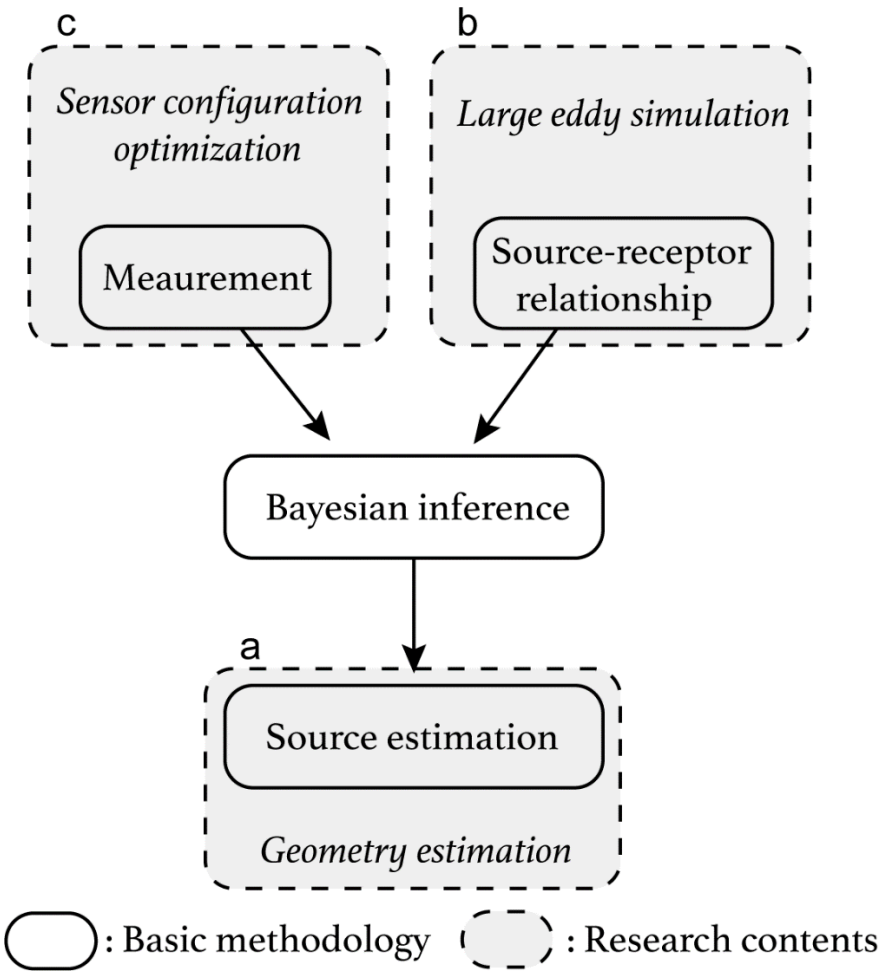


Figure 1.2. The relationship between basic statistical STE methodology and the research contents of the dissertation.

1.3 Research problem

The dispersion phenomena of pollutants are affected by many factors, including but not

limited to unsteady meteorological conditions, coupling of multiple sources, sources' movement, fluctuating emission strength, and chemical reactions. Thereby, a comprehensive STE for the urban environment needs the efforts of a wide range of research. Considering that it is impossible to solve all these issues by single research, to focus on the current target and keep the conciseness, the author needs to limit the range of this thesis by some basic assumptions. The research problem is assumed to be the identification of one single source with constant emission strength and fixed location in a statistically turbulent flow field in a neighborhood-scaled urban area. The emitted pollutants are regarded as passive scalars that will not react with other substances. These assumptions also have been widely used in the previous STE research.

1.4 Structure of the thesis

The thesis consists of 7 chapters as shown in **Fig. 1.3**. The content of each chapter is as follows:

- Chapter 1 introduces the research background, objective, and structure of the thesis.
- Chapter 2 gives a short review of previous STE research and explains the basic framework of the Bayesian inference STE method.
- Chapter 3-6 are the main content of this research. Chapter 3 proposes a new method to estimate the geometry information of the unknown source in STE.
- Chapters 4 & 5 apply large eddy simulation to model S-RR and compare its performance with conventional methods.
- Chapter 6 develops an entropy-based sensor configuration optimization method and evaluates its effectiveness in an ideal urban model.
- Chapter 7 is the concluding remark for this research. The future research plan based on the current progress is also discussed.

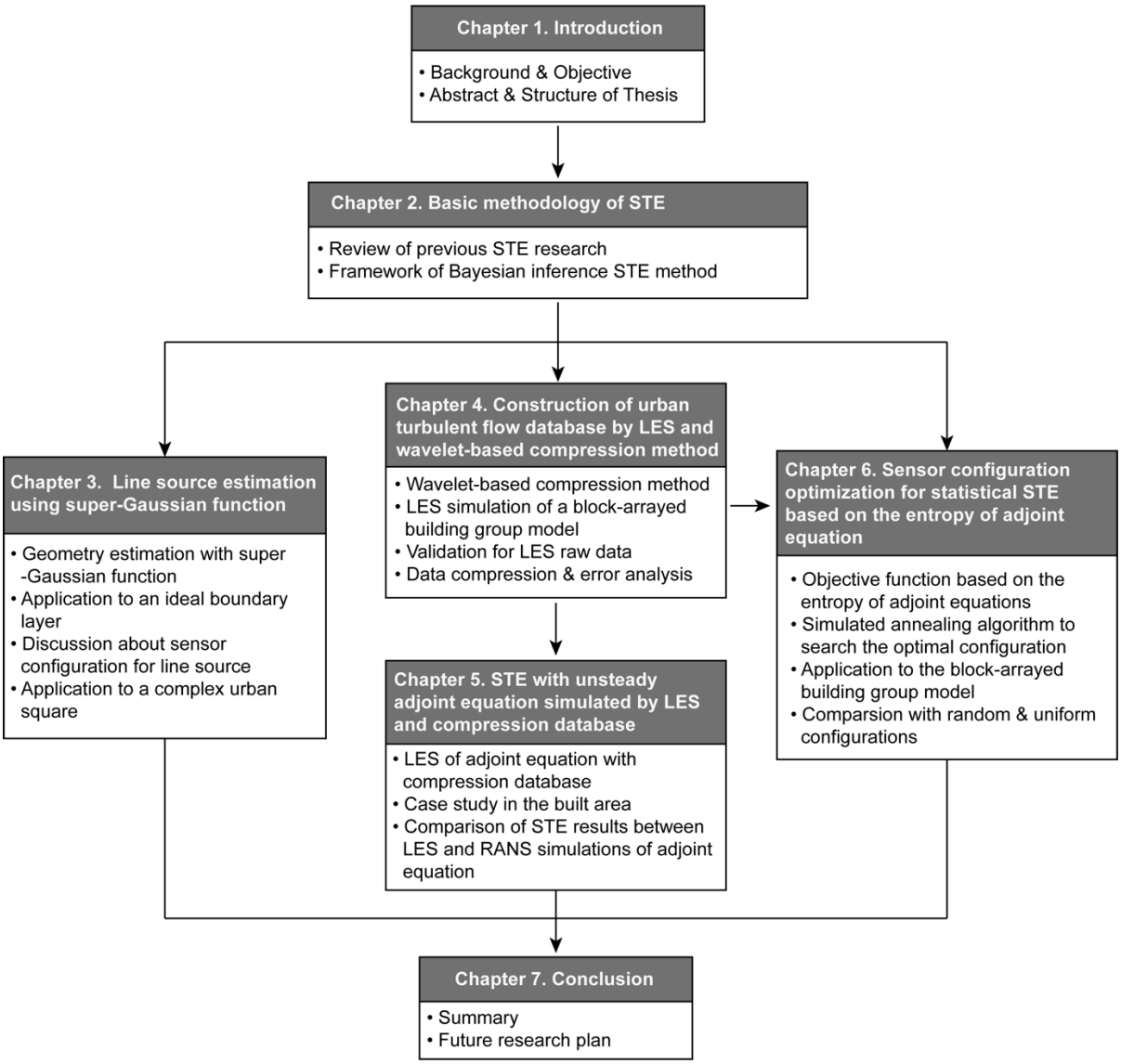


Figure 1.3. The structure of the dissertation.

Chapter 2
Basic methodology for source term estimation

Abstract

In this chapter, the basic methodology for source term estimation is introduced from the view of three elements: measurements, estimation algorithm, and source-receptor relationship. In each part, a short literature review about recent progress will be provided first. Then, for complicated urban applications, appropriate methods for each element are selected and described in detail.

2.1 Problem definition

The source term estimation (STE) problem is defined to identify unknown sources of pollutants in the urban environment. The focus is specified as the atmospheric pollutants here. Even so, the range of STE is still wide in the reality. It may involve different types of sources: single or multiple, fixed or mobile, point sources, or sources with geometry (volume). The emission patterns are also diverse. Some sources release the pollutants with constant strength, while some sources with varying strength. Besides, the properties of pollutants matter in STE. The inert gas can be regarded as the passive scalar which is easy to deal with, but some pollutants may be chemically active that they will react with other gases during dispersion. What's more, the pollution particles could experience buoyance, condensation, collision, and other phenomena during the dispersion. In addition, the meteorological condition is important to STE. In real applications, the incoming wind in the urban areas is unsteady. The wind speed and directions may fluctuate or even suddenly change during the dispersion emergency. It is inevitable to consider this possibility.

Therefore, STE is an inter-disciplinary research problem including numerous targets and complex dispersion situations. Each possible STE target mentioned above needs a series of research to solve, and even though intense research efforts have been devoted to STE, some issues remain unresolved. For instance, the forward modeling of pollution dispersion with chemical reaction is still in the early stage, let alone the STE for such pollution, which needs inverse simulation of such dispersion. Obviously, it is reasonable to infer that plenty of research is still necessary to comprehensively realize general STE in the future.

Since it is almost impossible to cover the total range of STE in a single research, and the current thesis is part of a series of studies to realize STE in the urban environment, it is necessary to limit the range of STE here by some assumption in order to focus on the main research target.

Considering that STE research is still rudimentary in the urban environment, estimations for multiple sources and mobile sources are too early. The STE target is limited to a single source $\boldsymbol{s}$ located at a fixed position $\boldsymbol{x}_s$. Here, $\boldsymbol{s}$ is a vector describing the information of the unknown source. $\boldsymbol{x}_s$ is the time-independent coordinate of the source in the domain. Because the source can have a geometry to estimate (**Chapter 3**), $\boldsymbol{x}_s$ includes all the coordinates that the source has.

As for the emission state, it is assumed that the pollutants are constantly released from the source with unchanged strength q_s. The temporal fluctuations of emission strength will not be considered here. After the emission, the pollutant is regarded as the passive scalar in the dispersion modeling. As mentioned before, the precise modeling techniques incorporating chemical reactions and other physical mechanisms are immature, which hinders STE for such pollution.

The meteorological condition also needs some limitations. Since the wind flow in the real-life cannot get rid of turbulence, turbulent inflow is taken into consideration. Meanwhile, it is expected that STE should be realized in a short time after the dispersion emergency happened. During this short time, it is appropriate to assume that the meteorological condition is statistically steady, which means that the wind direction will not change and the incoming turbulence owns a constant statistical property that can be described by several profiles.

According to these assumptions, the target is limited to a single source located at a fixed position in a statistically stationary flow field with a constant release strength, which can be described by a vector $\boldsymbol{s}$:

$$\boldsymbol{s} = (\boldsymbol{x}_s, q_s) \tag{2.1}$$

2.2 Measurement

After the pollutants were emitted from the source $\boldsymbol{s}$, the sensor network can monitor the information about the concentration and provide the measurement data. One of the most basic and common data types is the mean concentration over a certain period, which can be provided by low-cost sensors. In real applications, the time scale of averaging operation is influential to the measurements. At the beginning of the dispersion, because the spatial concentration distribution is developing, mean concentration may strongly fluctuate with the averaging time scale and start point. To avoid this, it is effective to start the averaging after the dispersion reached to statistically steady state and make sure the time scale is long enough.

With the development of sensing techniques, more information can be measured by sensors like time-series data and eddy covariance. Time series data can reflect the temporal change of concentration caused by multiple sources, moving sources, or unsteady emission

strength. Therefore, it is valuable in these scenarios. For instance, Wang et al. (2021) proposed a method for multi-point source identification based on the correlation of time-series measurements. Apart from that, eddy covariance measurement is also becoming a monitoring exercise rather than a purely scientific activity (Aubinet et al., 2012). The concentration flux can reveal the coupling effects of source and flow field on the sensor, which is valuable complementary information for STE. The flux has been well used in the modeling of the sensor's footprint (Hellsten et al., 2015), whose meaning is similar to S-RR which represents the measurement ability of the sensor and relationships between possible sources and a certain sensor. Thus, it is reasonable to deduce that the accuracy of STE could be improved further by adding flux measurements. However, it is necessary to predict flux measurements in advance in the source-receptor relationship (S-RR), which needs complicated simulation models. It still needs further research in the future.

Because the analysis of advanced measurement data is out of the interest of the thesis, the measurement data here is the time-averaged concentration data, which can be obtained by most sensors. The measurements provided by n sensors can be represented by a n elements vector $\boldsymbol{D}$.

2.3 Estimation algorithm

In the next step, it is necessary to estimate the source terms based on the measurements $\boldsymbol{D}$. The estimation algorithms based on static sensors can be divided into two frameworks: optimization and statistics.

2.3.1 Optimization framework

The optimization framework relies on a kernel objective function, which is constructed based on the difference between $\boldsymbol{D}$ and modeled concentration $\boldsymbol{R}$. The most common form is the residual sum square between $\boldsymbol{D}$ and $\boldsymbol{R}$. The true source is believed to be the one that minimizes the objective function in the vector space of $\boldsymbol{s}$. To effectively find the target, different types of methods are proposed for searching.

a. Gradient-based algorithm

The gradient-based methods find the optimal solution by decreasing the value of the

objective function. The descending direction is determined by the gradient of the objective function, which is approximated by the partial derivatives of the objective function over the source's parameters. Li and Niu (2005) used the gradient method to estimate the emission of volatile organic compounds in dry building materials. Sharan et al. (2012) applied the least square method to identify single and multiple point sources based on synthetic measurements and real measurements.

An extension of this method is called re-normalization (Issartel, 2003), which is a linear approximation for source terms based on measurements. It used the inverse linear relationship to estimate the source. To avoid artifacts generation in the inversion process, a renormalized weight function was introduced to limit the search space and can be updated in each iteration (Issartel, 2005a). The effectiveness of renormalization in STE was demonstrated in Kumar et al. (2015), where a point source in an urban-like environment was successfully reconstructed.

Overall, one drawback that needs to be noted in the gradient-based method is that the performance is vulnerable to the initial guess, and the optimization process may be stuck in a local minima without global searching.

b. Heuristics algorithm

Another searching strategy is the heuristics algorithms. Unlike gradient-based strategy, heuristics methods directly start searching without prior mathematical processing of objective function. They sacrifice precision or completeness for speed when there is no traditional way to efficiently find an accurate solution. The common heuristics searching methods in STE are pattern search, simulated annealing, and genetic algorithm.

Pattern search firstly defines the initial values for source parameters. Then, the algorithm changes each parameter by specified step length, which is called axis exploration, and the objective function value is calculated once again. The new value and original value are compared to decide whether a new estimation is accepted. If the objective value keeps unchanged in all directions, the step size is adjusted for new exploration, which is called pattern move. Zheng and Chen (2010) applied this method to locate a point source in a field experiment and found that it is limited to the local search, which needs to be improved.

Simulated annealing analogies the optimization process to the cooling process of a material, where the objective function is regarded as the thermodynamic energy of the system.

It also starts from an initial guess and generates new estimations by random disturbances. The special feature is that the cooling process analogy brings the concept of the temperature, which is used to adjust the acceptance probability of the new estimation. This feature enables the algorithm to jump out from the local minima and conduct a global search. More details can be found in **Chapter 6**. Thomson et al. (2007) used this method to locate an unknown point source in a desert environment.

Genetic algorithm is another popular global search method in engineering. At the beginning of the genetic algorithm, not a single guess but a random population of candidates are initialized. The most important feature is that the genetic algorithm focuses on the population rather than the parameter. It is inspired by the natural genetic evolution and finishes the optimization by selection, crossover, and mutation. Selection is to pick up the high-quality candidates to construct a gene pool by their values of the objective function. Crossover is to change part of the parameters of candidates in the pool to create the next generation. Mutation means that parameters may have random disturbances during the crossover to realize global search. The applications of the genetic algorithm in STE can be found in Allen et al. (2007) and Wang et al. (2018).

It is worth mentioning that the convergence of heuristic algorithms cannot be promised owing to their direct search property. It is difficult to judge whether the solution found by the heuristic is the global optimal or just a local peak. Many modifications have been proposed to improve its credibility like hybrid algorithms and combination with a good initial guess, but the effectiveness still needs more validations of real applications.

c. Limitation of optimization methods

Optimization methods frequently appear in the STE research, which demonstrates their effectiveness to the inverse problem. However, no matter in gradient-based algorithms or heuristic algorithms, only a single estimate, the final result, is provided to users. The complex, unavoidable noise in the real world cannot be separated from the result for further analysis. The substantial noise may make the optimization result totally deviate from the truth, thus users fail to identify the source.

2.3.2 Statistical framework

A better way to evaluate the noise in real applications is solving STE in a statistical framework, which estimates the probability of all candidates rather than a single result. The most popular statistical STE method is Bayesian inference. It treats each source term as an independent random variable. The posterior probabilities of these variables can be calculated by updating prior probabilities with sensor observations. Then stochastic sampling methods are used to attain the probability density function (PDF) of posterior probability for each parameter. As a result, users can obtain more information from PDF like the most probable estimate, the mean estimate, and credible interval when compared to optimization results. The uncertainty or noise during STE is reflected in PDFs. Keats et al. (2007) successfully identified a point source in a complex urban environment using Bayesian inference. Since their study, this method has been extended to estimate multiple point sources (Wade and Senocak, 2013a; Yee, 2012), mobile sources (Kopka et al., 2016).

Seeing that complex noise is inevitable in STE for the urban area, the Bayesian inference is selected as the main methodology in the thesis. Here a brief introduction is given.

a. Bayes' theorem and problem formulation

The definition of Bayes' theorem is:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.2}$$

where A and B are events and $p(B) \neq 0$. $p(A|B)$ is a conditional probability called posterior probability, which is the probability of event A occurring given that B is true. $p(B|A)$ is called likelihood function, which has the similar probability meaning with $p(A|B)$. Following the Bayes' theorem, the posterior probability that $\boldsymbol{s}$ is the true source given measurements $\boldsymbol{D}$ can be written as:

$$p(\mathbf{s}|\boldsymbol{D}, I) = \frac{p(\boldsymbol{D}|\mathbf{s}, I)p(\mathbf{s}|I)}{p(\boldsymbol{D}|I)} \tag{2.3}$$

Here, I is the background information, including meteorological and geographical data. $p(\boldsymbol{D}|\mathbf{s}, I)$ is the likelihood function. $p(\mathbf{s}|I)$ is the prior probability. $p(\boldsymbol{D}|I)$ is called the evidence. According to the definition of conditional probability, it is the probability of $\boldsymbol{D}$ can be measured given the background information. In this case, $p(\boldsymbol{D}|I)$ is independent of sources

and can be assumed as a constant when a certain I was imposed. It acts as a normalizing factor in Eq. (2.3), which does not affect the posterior probability distribution. Therefore, the posterior distribution can be estimated using a likelihood function and prior probability:

$$p(\mathbf{s}|\boldsymbol{D}, I) \propto p(\boldsymbol{D}|\mathbf{s}, I)p(\mathbf{s}|I) \tag{2.4}$$

b. Likelihood function

A likelihood function represents the probability that measurements, $\boldsymbol{D}$, can be obtained when the source is $\boldsymbol{s}$. This probability is evaluated based on the difference between the real measurements, $\boldsymbol{D}$, and the modeled concentration, $\boldsymbol{R}$, resulting from source $\boldsymbol{s}$. Considering the inevitable errors caused by measurement processes and modeling, the true concentration $\boldsymbol{D}_{true}$, $\boldsymbol{D}$, and $\boldsymbol{R}$ exhibit the following relationship:

$$\begin{aligned} \boldsymbol{D} &= \boldsymbol{D}_{true} + \boldsymbol{e}^d \\ \boldsymbol{R} &= \boldsymbol{D}_{true} + \boldsymbol{e}^m \end{aligned} \tag{2.5}$$

Here, $\boldsymbol{e}^d$ is the measurement error and $\boldsymbol{e}^m$ is the modeling error. In real-world applications, these errors play an important role in STE, though they are difficult to quantify (Yee et al., 2014). In the previous research (Keats et al., 2007a; Xue et al., 2017, 2018a), it is common to assume that the errors of each sensor follow Gaussian distributions with means of zero and variances of $\sigma_{d,i}^2$ and $\sigma_{m,i}^2$, where i is the index for the sensors. In this case, the likelihood can be calculated as:

$$p(\boldsymbol{D}|\boldsymbol{D}_{true}, I) \propto \exp\left[-\frac{1}{2}\sum\frac{(\boldsymbol{D}-\boldsymbol{D}_{true})^2}{\sigma_{d,i}^2}\right] \tag{2.6}$$

$$p(\boldsymbol{D}_{true}|\boldsymbol{s}, I) \propto \exp\left[-\frac{1}{2}\sum\frac{(\boldsymbol{D}_{true}-\boldsymbol{R}(\boldsymbol{s}))^2}{\sigma_{m,i}^2}\right] \tag{2.7}$$

$$\begin{aligned} p(\boldsymbol{D}|\boldsymbol{s}, I) &= \int p(\boldsymbol{D}|\boldsymbol{D}_{true}, I)p(\boldsymbol{D}_{true}|\boldsymbol{s}, I)d\boldsymbol{D}_{true} \\ &\propto \exp\left[-\frac{1}{2}\sum\frac{(\boldsymbol{D}-\boldsymbol{R}(\boldsymbol{s}))^2}{\sigma_{d,i}^2+\sigma_{m,i}^2}\right] \end{aligned} \tag{2.8}$$

c. Prior probability

A prior probability $p(\mathbf{s}|I)$ sets the probability distribution of source parameters before

the inference. The prior knowledge of an unknown source can be contained to make the inference faster or more certain. However, it is usually difficult to acquire credible information just after the emergency. It is usually assumed that no information except for the measurements was available before the inference. Therefore, the source can appear anywhere within the target domain and with any strength. Additionally, the source parameters are independent of each other, meaning that a uniform distribution was imposed on the prior probability.

$$p(\mathbf{s}|I) = constant \tag{2.9}$$

d. Posterior probability

The posterior probability can be obtained by combining Eq. (2.4), (2.8), and (2.9), as follows:

$$p(\boldsymbol{s}|D,I) \propto p(\boldsymbol{D}|\boldsymbol{s},I)p(\boldsymbol{s}|I) \propto \exp\left[-\frac{1}{2}\sum\frac{(\boldsymbol{D}-\boldsymbol{R}(\boldsymbol{s}))^2}{\sigma_{d,i}^2+\sigma_{m,i}^2}\right] \tag{2.10}$$

The only unknown information in Eq. (2.10) is the modeling concentration, $\boldsymbol{R}$, of source $\boldsymbol{s}$.

e. Sampling method

Although the target distribution can be explicitly formulated as Eq. (2.10), it is still difficult to determine its shape because the PDF lies in a multidimensional space composed of the elements of $\boldsymbol{s}$. Many solutions have been proposed to sample posterior distributions in Bayesian inference, among which the most commonly employed is the Monte-Carlo Markov chain (MCMC). However, the conventional MCMC often suffers a large rejection ratio in the high dimensional sampling, which diminishes the efficiency. In recent years, a hybrid MCMC method, the Metropolis–Hastings-within-Gibbs algorithm (MHMC) (Gilks et al., 1995; Hastings, 1970), was widely applied during the sampling process. MHMC introduces a statistical acceptance ratio adjustment based on the current sampling to mitigate this problem. This algorithm includes the following steps:

Step1 : Propose an initial guess of the source: $\boldsymbol{m}_1$

For i=1:*n*

Step 2 : Generate a new estimate $\tilde{\boldsymbol{m}}$ by a proposal distribution $q(\cdot)$: $\tilde{\boldsymbol{m}} \sim q(\tilde{\boldsymbol{m}}|\boldsymbol{m}_i)$

Step 3 : Evaluate the MH acceptance probability:

$$\text{Set } \alpha = min\left[1, \frac{P(\tilde{\boldsymbol{m}}|D,I)q(\tilde{\boldsymbol{m}}|\boldsymbol{m}_i)}{P(\boldsymbol{m}|D,I)q(\boldsymbol{m}_i|\tilde{\boldsymbol{m}})}\right]$$

Step 4 : Update the Markov Chain according to acceptance or rejection:

$$\boldsymbol{m}_{i+1} = \begin{cases} \tilde{\boldsymbol{m}} & if\ \alpha \geq N[0,1]; \\ \boldsymbol{m}_i & otherwise; \end{cases}$$

End For

Here $N[0,1]$ represents the uniform distribution bound between 0 and 1, and n is the sampling number.

On the other hand, some methods also have been proposed to accelerate MCMC with more complicated techniques, such as the application of variational inference (de Freitas et al., 2013), resampling (Gelfand and Sahu, 1994), and subspace construction (Constantine et al., 2016). However, their efficiency and robustness still need testification by practical applications. Heavy tails and local peaks may appear in the complex sampling process (Vrugt et al., 2009). Therefore, this thesis employed MHMC for the balance between speed and steadiness. The creditability of this method has been proved in the previous research of STE (Keats et al., 2007a; Xue et al., 2018a).

For each case in the dissertation, the Markov chain starts from an initial guess that is far enough from the true source. The total sampling number is set to 5.0×10^6, in which the first 5.0×10^5 samplings are discarded for the burn-in process. The rest 4.5×10^6 samplings are used to estimate the PDF of the posterior probability.

2.4 Source-receptor relationship

As shown in Eq. (2.10), the modeling concentration, $\boldsymbol{R}$, of each sensor corresponding to each possible source, $\boldsymbol{s}$, is needed to evaluate the posterior probability distribution. Apparently, it is necessary to calculate $\boldsymbol{R}$ for all possible sources and construct a database, which is exactly the S-RR. Because the number of possible combinations of parameters in vector $\boldsymbol{s}$ is so large, special simulation techniques including the Gaussian puff model, Markov chain method, and adjoint equation method have been proposed.

2.4.1 Gaussian puff model

One popular forward calculation method is the Gaussian puff model. The concentration distribution for the continuous release is formulated as Zheng and Chen (2010):

$$R(x,y,z) = \frac{q_s}{2\pi u\sigma_y\sigma_z} exp\left(-\frac{y^2}{2\sigma_y^2}\right) \times \left\{exp\left(-\frac{(z-H_e)^2}{2\sigma_z^2}\right) + exp\,(-\frac{(z+H_e)^2}{2\sigma_z^2})\right\} \quad (2.11)$$

Here, $R(x,y,z)$ is the modeling concentration at the location (x,y,z) in the downwind direction from the source following Cartesian coordinate. u is the wind speed. σ_y and σ_z are the dispersion coefficients for each direction. H_e is the effective height of the source.

It can be seen that Eq. (2.11) is an algebraic model. Given the source terms and sensor's position, the modeling simulation can be quickly calculated. In this case, even with a large number of possible sources, S-RR still can be easily constructed. However, the simplicity also brings with an important flaw that it can only describe an ideal dispersion in a homogenous flow field. Its accuracy cannot be promised when an obstacle appears in the domain. Until now, the literature using this model (Cui et al., 2019; Kormi et al., 2018; Ma et al., 2017; Wade and Senocak, 2013b; Wang et al., 2018; Zheng and Chen, 2010) also focus on the homogenous dispersion in the open field. As a result, the applicability of the Gaussian model in the urban environment is limited.

2.4.2 Markov chain model

Markov chain model is another forward calculation technique that was first proposed to solve indoor contaminant transport by Nicas (2000). The basic idea is to divide the dispersion domain into n zones or cells, and the concentration distribution of these zones is called states related to time. After the source appears at a certain state, the pollutant of each zone will move to an adjacent zone or remain still in the next step. Then, the transport behavior can be driven by a transition probability matrix:

$$\boldsymbol{P} = (p_{i,j})_{(n\times n)} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix} \tag{2.12}$$

In this matrix, $p_{i,i}$ is the probability that the pollutant will stay in the original cell, and $p_{i,j}$ represents the probability that pollutant will move from cell i to cell j in the next step. According to the research of Chen et al. (2015), the probability matrix can be calculated by the computational fluid dynamics (CFD) simulation of airflow. The concentration distribution at any time step is the result of iterating multiple of $\boldsymbol{P}$ to the initial state.

There are two drawbacks in this model. To decrease the computational burden, $\boldsymbol{P}$ is often kept constant or owns limited patterns. This treatment is acceptable when the flow field and the turbulent diffusion are steady. If strongly fluctuating turbulence exists in the target domain, the statistical properties of the flow field are difficult to be described only by several patterns. Furthermore, the number of divided cells or elements in $\boldsymbol{P}$ cannot be too large, which means the resolution of simulation and STE is restricted. As a result, the best application scenario of this method is the place where the flow field is almost unchanged and the low-resolution is allowable, such as a small indoor space with a fixed ventilation pattern (Li et al., 2020). Its applicability in the large-scaled outdoor space is in doubt and related literature is sparse.

2.4.3 Adjoint equation method

It can be stated that the accuracy of the two methods above is not enough in the complex urban area despite their computational simplicity. Speaking of accuracy, with the development of CFD techniques and high-performance computers, the dispersion behavior of atmospheric pollutants can be well predicted using turbulence models and transport equations. The related contents about turbulence modeling and CFD simulation of dispersion of atmospheric pollution are introduced in detail in **Appendix. A**. Some representative literature about applications in the urban environment can be referred to (Branford et al., 2011a; Coceal et al., 2014a; Tominaga and Stathopoulos, 2011a, 2013).

However, the forward CFD dispersion method needs massive computation cost and long calculation time, it is almost impossible to directly predict the dispersion of all possible sources.

To overcome this problem, Pudykiewicz (1998) introduced the adjoint equation method into S-RR simulation to decrease the extreme computational costs. This approach enables users to construct S-RR from the perspective of sensors using inverse simulations, rather than from possible sources via forward simulation. Since the number of sensors is finite, the use of adjoint equations makes the construction of S-RR much more convenient. A brief introduction of the adjoint equation is given below.

a. Adjoint equation theory

For any source, $\boldsymbol{s}$, that releases a passive scalar value in a spatiotemporal domain with a strength of q_s and location of $\boldsymbol{x}_s$, the conservation equation of passive scalar transport can be written as (also Eq. (A.3) in **Appendix. A**):

$$\frac{\partial C}{\partial t}+\frac{\partial u_j C}{\partial x_j}-\frac{\partial}{\partial x_j}\left(D_m\frac{\partial C}{\partial x_j}\right)=q_s\delta(\boldsymbol{x}-\boldsymbol{x}_s) \tag{2.13}$$

with the boundary conditions:

$$\begin{gathered}\nabla_{\boldsymbol{n}}C=0 \text{ at } \partial\Omega \\ C(\boldsymbol{x},t=0)=0\end{gathered} \tag{2.14}$$

Here, C is the concentration function of the passive scalar in a spatial domain of Ω and time of $[0,\mathbb{T}]$, D_m is the mass diffusivity, $\delta(\cdot)$ is the Dirac delta function, and $\nabla_{\boldsymbol{n}}$ is a directional derivative normal to the boundary. The left-hand side (LHS) of Eq. (2.13) can be regarded as the resultant transformation of linear operator $\boldsymbol{L}(\cdot)$ on function C in a Hilbert space, where $\boldsymbol{L}(\cdot)$ is defined as:

$$\boldsymbol{L}(\cdot)\equiv\frac{\partial(\cdot)}{\partial t}+\frac{\partial\left(u_j\,\cdot\right)}{\partial x_j}-\frac{\partial}{\partial x_j}\left(D_m\frac{\partial(\cdot)}{\partial x_j}\right) \tag{2.15}$$

and Eq. (2.13) turns into:

$$\boldsymbol{L}(C)=q_s\delta(\boldsymbol{x}-\boldsymbol{x}_s) \tag{2.16}$$

Then, we introduce the conjugate concentration field C^* and consider the following spatiotemporal integration.

$$\langle C^* \boldsymbol{L}(C) \rangle = \int_T \int_\Omega (C^* \cdot \boldsymbol{L}(C))\, d\boldsymbol{x} dt \tag{2.17}$$

For convenience, the spatiotemporal integration is expressed as $\langle \cdot \rangle$. If we expand the LHS of Eq. (2.17), we have

$$\begin{aligned}
\langle C^* \boldsymbol{L}(C) \rangle &= \langle C^* \left\{ \frac{\partial (C)}{\partial t} + \frac{\partial \left(u_j C\right)}{\partial x_j} - \frac{\partial}{\partial x_j}\left(D_m \frac{\partial C}{\partial x_j} \right) \right\} \rangle \\
&= \langle C^* \frac{\partial (C)}{\partial t} + C^* \frac{\partial \left(u_j C\right)}{\partial x_j} - C^* \frac{\partial}{\partial x_j}\left(D_m \frac{\partial C}{\partial x_j} \right) \rangle \\
&= \langle \frac{\partial (C^* C)}{\partial t} - C \frac{\partial (C^*)}{\partial t} + \frac{\partial \left(C C^* u_j\right)}{\partial x_j} - C u_j \frac{\partial (C^*)}{\partial x_j} \\
&\quad - \frac{\partial}{\partial x_j}\left(C^* D_m \frac{\partial C}{\partial x_j} \right) + D_m \frac{\partial C}{\partial x_j} \frac{\partial C^*}{\partial x_j} \rangle \\
&= \langle \frac{\partial (C^* C)}{\partial t} - C \frac{\partial (C^*)}{\partial t} + \frac{\partial \left(C C^* u_j\right)}{\partial x_j} - C u_j \frac{\partial (C^*)}{\partial x_j} \\
&\quad - \frac{\partial}{\partial x_j}\left(C^* D_m \frac{\partial C}{\partial x_j} \right) + \frac{\partial}{\partial x_j}(C D_m \frac{\partial C^*}{\partial x_j}) \\
&\quad - C \frac{\partial}{\partial x_j}(D_m \frac{\partial C^*}{\partial x_j}) \rangle \\
&= \langle \frac{\partial (C^* C)}{\partial t} \rangle \\
&\quad + \left\langle -C \frac{\partial (C^*)}{\partial t} - C u_j \frac{\partial (C^*)}{\partial x_j} - C \frac{\partial}{\partial x_j}\left(D_m \frac{\partial C^*}{\partial x_j} \right) \right\rangle \\
&\quad + \left\langle \frac{\partial \left(C C^* u_j\right)}{\partial x_j} - \frac{\partial}{\partial x_j}\left(C^* D_m \frac{\partial C}{\partial x_j} \right) + \frac{\partial}{\partial x_j}(C D_m \frac{\partial C^*}{\partial x_j}) \right\rangle
\end{aligned} \tag{2.18}$$

If the integration happened at the time when C and C^* reach steady state or the integration was conducted following the periodicity of C and C^*, the first term is 0. According to the divergence theorem, the last term can be transferred into the surface integration of the space, which is called the boundary term. If the spatial integration is conducted in the space far away from the source, the boundary terms are almost 0. Therefore, Eq. (2.18) changes into

$$\langle C^* \boldsymbol{L}(C) \rangle = \left\langle -C\frac{\partial (C^*)}{\partial t} - Cu_j \frac{\partial (C^*)}{\partial x_j} - C\frac{\partial}{\partial x_j}\left(D_m \frac{\partial C^*}{\partial x_j}\right) \right\rangle = \langle \boldsymbol{L}^*(C^*) C \rangle \tag{2.19}$$

Here, $\boldsymbol{L}^*$ is the adjoint operator of $\boldsymbol{L}$,

$$\boldsymbol{L}^* \equiv -\frac{\partial (\cdot)}{\partial t} - u_j \frac{\partial (\cdot)}{\partial x_j} - \frac{\partial}{\partial x_j}\left(D_m \frac{\partial (\cdot)}{\partial x_j}\right) \tag{2.20}$$

Similarly, the transformation of $\boldsymbol{L}^*(\cdot)$ in C^* can be regarded as the governing equation of an adjoint tracer dispersion equation, which can be expressed as:

$$\boldsymbol{L}^*(C^*) = q_m \delta(\boldsymbol{x} - \boldsymbol{x}_m) \tag{2.21}$$

with the boundary conditions:

$$\begin{gathered} \nabla_{\boldsymbol{n}} C^* = 0 \text{ at } \partial\Omega \\ C^*(\boldsymbol{x}, t = \mathbb{T}) = 0 \end{gathered} \tag{2.22}$$

In contrast to the dispersion behavior of source $\boldsymbol{s}$, where the passive scalar is transported by $\boldsymbol{u}(\boldsymbol{x}, t)$ from time 0 to time $\mathbb{T}$, the physical meaning of Eq. (2.21) is the dispersion transported by $-\boldsymbol{u}(\boldsymbol{x}, t)$ from time $\mathbb{T}$ to time 0, where the source is located at $\boldsymbol{x}_m$ and has the strength of q_m. The importance of the adjoint equation can be revealed after Eq. (2.16) and Eq. (2.21) are substituted into Eq. (2.19):

$$q_s \langle C^*(\boldsymbol{x}_s) \rangle = q_m \langle C(\boldsymbol{x}_m) \rangle \tag{2.23}$$

If we impose the location of a sensor to $\boldsymbol{x}_m$ and unit to q_m, the right-hand side (RHS) of Eq. (2.23) is the modeling concentration $\boldsymbol{R}$ of the sensor, which is equal to the LHS, and the time-averaged concentration of tracers at source $\boldsymbol{s}$ emitted from the sensor. Therefore, in this method, the number of dispersion equations that must be solved declines remarkably from the number of possible sources, $\boldsymbol{s}$, to the number of sensors applied. Owing to the convenience of this calculation, adjoint equations have been applied to model S-RR in several previous studies (Efthimiou et al., 2018a; Kumar et al., 2015a; Rajaona et al., 2015; Xue et al., 2017, 2018a) for outdoor STE.

b. Simulation of the adjoint equation

According to Eq. (2.23), the modeling concentration $\boldsymbol{R}$ can be obtained after the adjoint concentration field C^* was calculated. Because Eq. (2.21) is in the form of a partial differential equation like the transport equation of Eq. (2.13), it can be solved by numerical simulation. In the previous research, it was commonly simulated by Reynolds averaged Navier-Stokes model based on the Reynolds-averaged form.

$$-\overline{\boldsymbol{u}}\frac{\partial(\overline{C^*})}{\partial \boldsymbol{x}}-\frac{\partial}{\partial \boldsymbol{x}}\left([D_t+D_m]\frac{\partial(\overline{C^*})}{\partial \boldsymbol{x}}\right)=q_m\delta(\boldsymbol{X}-\boldsymbol{X}_m) \tag{2.24}$$

where $\overline{\boldsymbol{u}}$ is the mean velocity field obtained from forward simulation and $\overline{C^*}$ is the simulated mean distribution of the adjoint concentration. D_t is the turbulent diffusivity used to model the turbulent scalar fluxes, which is an extra term produced by Reynolds averaging. The molecular diffusion coefficient D_m was set as $1.5\times10^{-5}m^2/s$ in this thesis, which corresponds to the diffusion of C_2H_4, a common gas used in the dispersion experiment. It is true that different gases have different D_m values, but most of them lies in the range of 10^{-6} to $10^{-5}m^2/s$ (Wilke and Lee, 1955). When compared with the D_t, this small order difference would not cause too much effects on the dispersion behaviors.

Meanwhile, the wind speed determines the Reynolds number of the flow field in the domain. For the flow field with a high Reynolds number, which is common in the atmospheric boundary layer, the turbulence is well-developed, and its effects on the dispersion are large enough to make the molecular diffusion neglectable. Therefore, in the microscale or mesoscale of the atmospheric environment, where the ratio between the wind speed and diffusion coefficient is relatively large, the dispersion is dominated by turbulence rather than molecular diffusion.

Furthermore, for a certain dangerous pollutant with a special diffusion coefficient that needs attention, the adjoint equation can be specially simulated with this coefficient to build the S-RR, which would be applied later in the Bayesian inference when sensors monitor this pollutant.

Before the simulation of Eq. (2.24), a forward numerical simulation is necessary to obtain the mean velocity field $\overline{\boldsymbol{u}}$ of the target domain. One of the important factors in this forward simulation is the inflow boundary condition for $\overline{\boldsymbol{u}}$, which should correctly reflect the wind

characteristics of the dispersion emergency. It was assumed that the coming wind is statistically stationary during the dispersion process of pollutants. In fact, because adjoint concentration databases can be constructed long before emergencies, there is enough time to simulate them with different wind speeds and directions to cover the possible meteorological situations of the target domain. When the dispersion emergencies happen, as important meteorological information for STE, the properties of wind can be measured and included in the prior information. Then, the adjoint concentration corresponding to the wind properties would be selected for the Bayesian inference. In this case, because the adjoint equation was simulated under the same inflow as reality, the wind effects on the source-receptor relationship can be correctly built in the proposed method.

Because the simulation of the adjoint equation requires a preparatory forward simulation for the flow field information, theoretically, its calculation cost is a little bit higher than the Gaussian puff model and Markov chain model. Despite that, the complex turbulence dispersion is included in the adjoint equation to realize much higher accuracy, it is the most promising method for S-RR simulation in the outdoor STE among these three and consequently is applied in this thesis.

2.5 Conclusion

In this chapter, three core parts of STE: measurements, estimation algorithm, and S-RR were introduced. The commonly used methods in three parts were reviewed first. Then, considering the characteristics of STE in the urban environment, suitable methods for the thesis were selected to construct the basic structure of STE. The measurements are time-averaged concentrations measured by discrete sensors. These measurements are inputted into Bayesian inference to estimate posterior probability distributions of source parameters based on the comparison with modeled concentrations. The adjoint equations are used to predict the modeled concentrations before emergencies. In the following chapters, all the improvements and STEs will be conducted based on this basic structure.

Symbols

C : the concentration distribution caused by a source

C^* : adjoint concentration distribution

$\boldsymbol{D}$: measurements vector

D_m : the mass diffusivity of the pollutant ($1.5 \times 10^{-5} m^2/s$)

D_t : turbulent diffusivity

$\boldsymbol{D}_{true}$: the true concentration at the sensor

$\boldsymbol{e}^d$: measurement error of the sensor

$\boldsymbol{e}^m$: modeling errors in the modeled concentration

H_e : effective height of the source

I : background information for Bayesian inference

$\boldsymbol{L}(\cdot)$: linear operator for the conservation equation of passive scalar transport

$\boldsymbol{L}^*(\cdot)$: the adjoint operator of $\boldsymbol{L}(\cdot)$

$\boldsymbol{m}_i$: the ith sampling in the MHMC

$min[\mathrm{a}, \mathrm{b}]$: the minimum value between a and b

$N[\mathrm{a}, \mathrm{b}]$: the uniform distribution bound between a and b

$p(A)$: the probability of event A

$p(A|B)$: conditional probability of event A occurring given that B is true

$\boldsymbol{P}$: transition probability matrix in the Markov chain dispersion model

$p_{i,j}$: the probability that pollutant will move from cell i to cell j in the next step

q_m : release strength of the sensor in the adjoint equation

q_s : release strength of the source

$\boldsymbol{R}$: modeled concentration vector

$R(x,y,z)$: the modeling concentration at the location (x,y,z) in the downwind direction from the source following Cartesian coordinate in the Gaussian puff model

$\boldsymbol{s}$: a vector representing the unknown source

t : time from 0 to $\mathbb{T}$

$\boldsymbol{u}$: wind velocity

u_j : wind velocity in j direction

$\boldsymbol{x}_m$: coordinates of the sensor

x_j : coordinate in j direction

$\boldsymbol{x}_s$: coordinates of the source

α : acceptance probability for a new sampling in MHMC

$\delta(\cdot)$: Dirac delta function

$\sigma_{d,i}^2$: the variance of error in the measurement of the sensor with index i

$\sigma_{m,i}^2$: the variance of error in the modeling concentration for the sensor with index i

σ_y : dispersion coefficients for y direction in the Gaussian puff model

σ_z : dispersion coefficients for z direction in the Gaussian puff model

$\nabla_{\boldsymbol{n}}$: a directional derivative normal to the boundary

Ω : a spatial domain for dispersion and adjoint dispersion

$\langle \cdot \rangle$: spatiotemporal integration for domain Ω and time t

$\bar{}$: Reynolds average operator

Chapter 3
Line source estimation using super-Gaussian function

Abstract

This chapter tries to extend the standard statistical source term estimation method to be capable of line source estimation, which includes more geometric information. Firstly, the proposed solution based on the super-Gaussian function is introduced. To justify its effectiveness, it is applied to the estimation of a line source in an ideal urban boundary layer with simulated measurements. Based on the estimation results, the effects of different sensor configurations on results and special requirements of line source on the configuration are discussed. After that, the performance of the proposed method is confirmed further by identifying a line source in a complex urban square with measurements in a wind tunnel experiment. By comparing with the conventional point source estimation method, the necessity and effectiveness of the proposed method are demonstrated.

3.1 Introduction

According to **Chapter 1&2**, most prior research aimed to estimate the parameters of a single point source, assuming that sources can be represented by a point without geometry; however, not all sources can be regarded as a point and most have shapes or volumes that cannot be neglected. The point assumption may result in noisy or even wrong estimation of these sources. Additionally, the source's geometric information is important for risk management and evaluation of seriousness. The recent public controversy concerning the Amazon forest fire (BBC, 2019) is a typical result of a lack of geometric information. Therefore, extending the source term estimation (STE) method for the estimation of sources with geometry is meaningful.

In addition, the line is a common and elementary geometry for pollution sources in the atmospheric environment (Gromke et al., 2008; Meroney et al., 1996; Salim et al., 2011a, 2011b). Traffic pollution in a street canyon has been extensively studied as a typical line source. Ideally, a method of source estimation handling any geometry should be possible; however, owing to its complexity and the numerous different kinds of geometry, this would require a series of research beyond the scope of one thesis. Correctly estimating the line source terms, which are not only position and strength but also length, width, and inclined angle is a good starting point for research.

The objective of this chapter is to extend the current STE method from point source estimation to line source estimation, which includes more geometric information. The basic structure is a combination of the super-Gaussian function for the source geometry and the Bayesian inference method. The method has the potential to estimate several common shapes: point, line, rectangular, and ellipse. The effectiveness of the method is evaluated by estimating a line source in two cases: an ideal urban boundary layer with simulated measurements and a complex urban square with discrete measurements, obtained via a wind tunnel experiment. The results of the new method are also compared with the conventional method's performance with a point assumption in the second case to demonstrate the necessity of the proposed method and geometry estimation.

3.2 Line source model

Eq. (2.23) has been widely used in previous research for point source estimation due to

convenience. During the Bayesian inference calculation, the modeled concentration $\boldsymbol{R}$ for any possible point sources $\langle C^*(\boldsymbol{x}_s)\rangle$ can be picked up simply by their coordinates. However, the situation changes when the sources have various shapes or volumes instead of being a point. Considering the speed and efficiency of the sampling algorithm, the shape estimation of sources should be realized not by combining points randomly, but by some controllable functions that can be adjusted directly with several coefficients. For line source estimation, this chapter proposes that SGF is a good choice. This section introduces the super-Gaussian function as below.

3.2.1 Ordinary Gaussian function

A general format of the multivariate Gaussian (or normal) function is

$$f_{\boldsymbol{x}}(x_1,\dots,x_m)=\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{3.1}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of variable $\boldsymbol{x}$, and $\boldsymbol{\mu}$ is the mean value matrix of variable $\boldsymbol{x}$. When $m=3$, this function corresponds to any ellipsoid or sphere in a given 3-dimensional space by changing the correlation coefficients in $\boldsymbol{\Sigma}$ and mean value in $\boldsymbol{\mu}$. Without loss of generality, the bivariate case can be transformed to a more intuitive format:

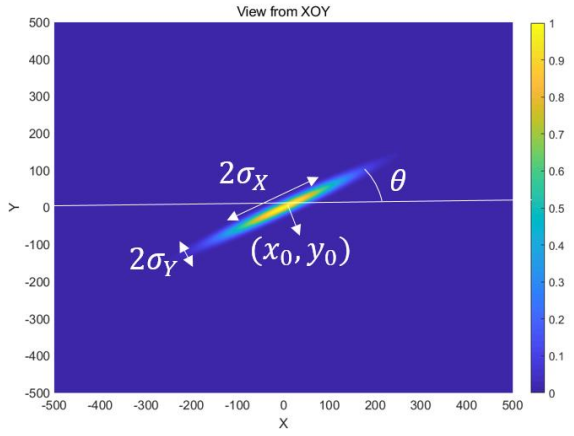
$$f(x,y)=A\exp[-(a(x-x_0)^2+2b(x-x_0)(y-y_0)+c(y-y_0)^2)] \tag{3.2}$$

where

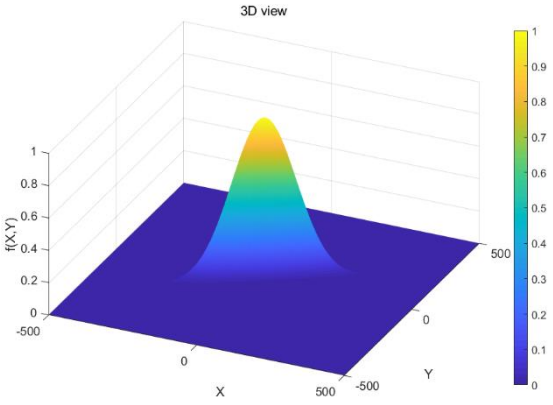
$$\begin{aligned} a&=\frac{\cos^2\theta}{2\sigma_X^2}+\frac{\sin^2\theta}{2\sigma_Y^2}\\ b&=-\frac{\sin 2\theta}{4\sigma_X^2}+\frac{\sin 2\theta}{4\sigma_Y^2}\\ c&=\frac{\sin^2\theta}{2\sigma_X^2}+\frac{\cos^2\theta}{2\sigma_Y^2} \end{aligned} \tag{3.3}$$

A is a normalized factor. In this case, the Gaussian function can be adjusted directly through five coefficients: x_0, y_0, θ, σ_X, σ_Y, which determine the geometric shape independently. If we let the ratio σ_X/σ_Y become very large or very small, the shape is very close to a line as shown by **Fig. 3.1(a)**. Meanwhile, (x_0, y_0) is the coordinate of the middle point of the line, θ represents the inclined angle, $2\sigma_X$ and $2\sigma_Y$ are the width and length of the line.

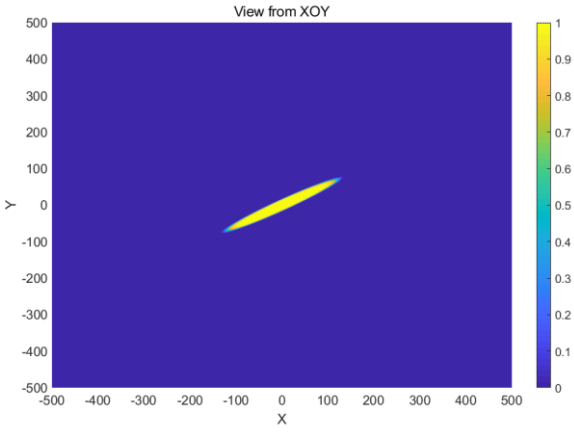
More importantly, this method is not limited to the line source only. It also has the potential to estimate sources with shapes of ideal point, ellipse, and rectangular. During the STE process, the Bayesian inference will tune the coefficients automatically according to the measurement information to find the most appropriate shape: very small σ_X and σ_Y when the source is a point, a large ratio σ_X / σ_Y when the source is a line and normal σ_X, σ_Y when the source is ellipsoid or rectangular. In other words, it is not necessary to know the shape of the source beforehand, which is impossible in practical application. This model can estimate the approximate shape using measurement information only, as long as the source's shape is close to a sphere, ellipse, and line. If the source is an ideal point like the previous research, this method will regress into the conventional process automatically and yield the same result.



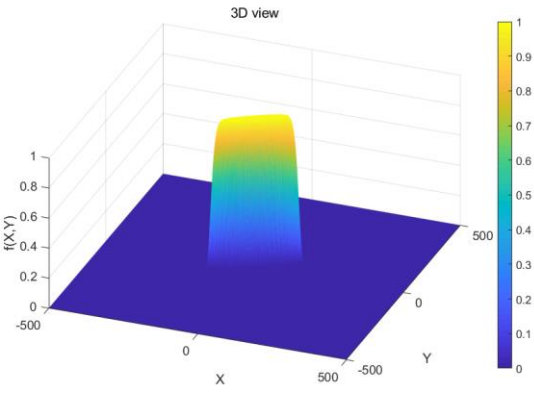
(a) Horizontal plane (ordinary, $\lambda = 1$)



(b) 3-D view (ordinary, $\lambda = 1$)



(c) Horizontal plane (Super, $\lambda = 8$)



(d) 3-D view (Super, $\lambda = 8$)

Figure 3.1. Schematic of ordinary and super Gaussian function

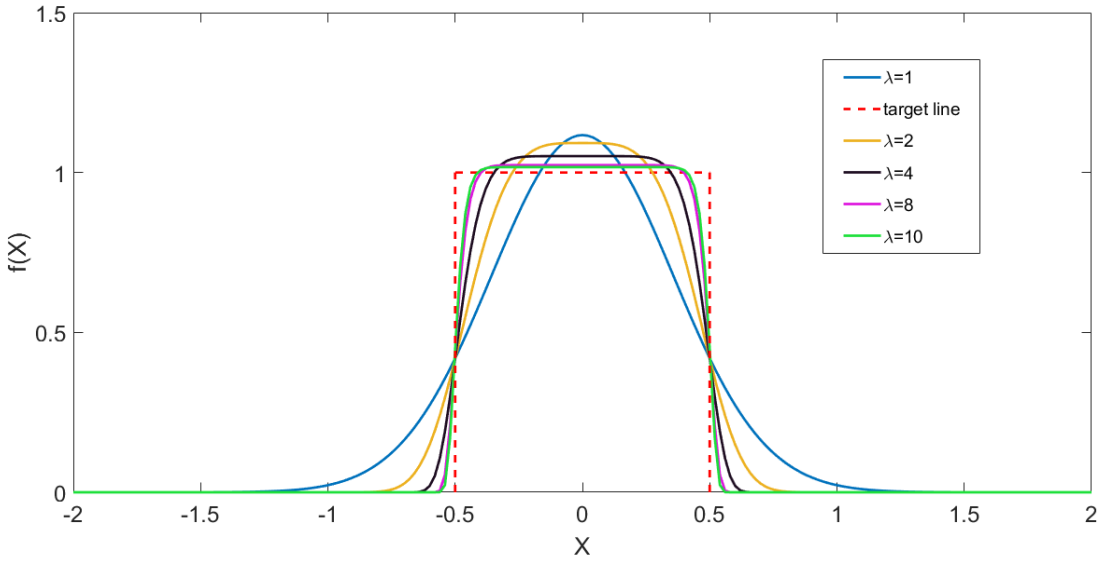


Figure 3.2. The shape of the super-Gaussian function with different powers

3.2.2 Super-Gaussian function

Nonetheless, the ordinary Gaussian function may not be the most appropriate one because of nonuniformity and tail error. Simply, if we use a one-dimensional normal distribution with the variance σ_X^2 to approximate the target line segment, it can be obviously noticed that the shape deviates from the target since the value is not uniform in the middle area (in **Fig. 3.2**). Besides, there are residual errors in the two tails where the value should be 0. Thereby, this model will bring extra error, unwanted in the inference. To minimize this error, this paper applied the super-Gaussian function (Parent et al., 1992) instead of the ordinary one.

$$f_{\boldsymbol{x}}(x_1, \ldots, x_m) = A \exp\left[-\left(\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{\lambda}\right] \tag{3.4}$$

This equation modifies the shape with a flat-top and Gaussian fall-off by raising the content of the exponent to a power λ. It can be observed in **Fig. 3.2** that when λ increases, the slope becomes precipitous and the flat top becomes longer, which makes the shape closer to the target shape. In this research, λ was set to eight to balance between error control and calculation simplicity. The two-dimensional super-Gaussian function is

$$f(x,y) = A\exp\left[-(a(x-x_0)^2 + 2b(x-x_0)(y-y_0) + c(y-y_0)^2)^\lambda\right] \quad (3.5)$$

and the shape is shown in **Fig. 1(c)(d)**.

In real urban areas, most line sources can be regarded as a line in the horizontal plane, which is close to the surface. Moreover, the right-hand side of Eq. (2.23) involves huge matrixes product in the numerical calculation, if the position of the line source has no restraint. Therefore, this paper only deals with a line source in a horizontal plane with a fixed height above the surface. The height of the line is included in the background information I and does not need to be estimated, due to which Eq. (3.5) is adequate for this research. In spite of that, it is worth mentioning that this does not mean that the super-Gaussian function is incapable of estimating the height or vertical inclined angle of other lines. In fact, that kind of estimation can be acquired quickly by increasing the dimension k to three and adding two more correlation coefficients in $\boldsymbol{\Sigma}$. Hence, the parameters vector of the line sources can be written as

$$\boldsymbol{s} = \left(x_s, y_s, \theta, \sigma_X, \sigma_Y, q_s\right) \quad (3.6)$$

In the application, these six parameters will be estimated.

3.3 Case I: numerical experiment of urban boundary layer

To evaluate the applicability of the proposed method, this case uses it to estimate the parameters of a line source in a simple urban area, based on computational fluid dynamics (CFD) simulation data. One of the most elementary dispersion cases, a steady-state urban boundary layer without obstacles on the ground, is analyzed. This setting is appropriate to explore the basic properties of the proposed method.

3.3.1 Numerical simulation

It is necessary to run a forward CFD simulation first to synthesize the measurements, then adjoint equation simulation is run based on the forward flow field.

The calculation domain of the forward simulation, with a size of $1500\,m(x) \times 1000\,m(y) \times 500\,m(z)$, is shown in **Fig. 3.3**. In the x and y directions, the cubic grid is

uniform, with 10 m length. To ensure that the line source has a sufficiently high resolution in the vertical direction and to simulate the dispersion behavior precisely near the bottom surface, the finest mesh is generated from the bottom, with a 0.2-m length. The mesh becomes rougher along with the height at a 1.02 ratio. To generate the effects of the urban boundary layer, a half channel flow is produced by mapping the flow properties of the outlet to the inlet. After each step, the pressure gradient is adjusted to keep the mass flow rate constant, at $\overline{U} = 4\ \mathrm{m/s}$. At the bottom wall, the wall function proposed by Blocken et al. (2007) is applied. k_s is set as 45 m, which corresponds to a Davenport roughness of 1.5 m, a common case for urban areas. The turbulence is modeled via a standard $k-\varepsilon$ model. The details of other configuration are listed in **Table 3.1**.

Table 3.1. Numerical schemes and boundary conditions for Case I

Time marching	Steady state, Semi-Implicit Method for Pressure-Linked Equations (SIMPLE) method.
Spatial discretization	Advection term: total variation diminishing (TVD) scheme;
Inlet	U, k, ε, D_t: mapped Outlet; C: constant (= 0); C^*: zero-gradient
Outlet	Flow: zero-gradient; C: zero-gradient; C^*: constant (= 0)
Top wall	Slip;
Bottom wall	Flow: generalized logarithmic law; Roughness height: 1.5 m Davenport Roughness; C, C^*: zero-gradient
Source	C, C^*: Constant injection rate (= 1 1/s)

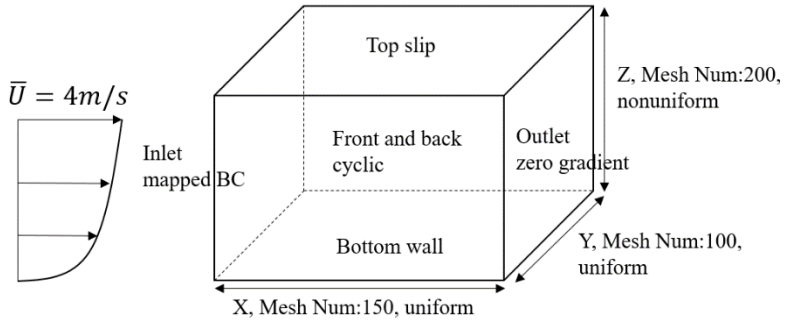


Figure 3.3. The calculation domain and boundary condition of the numerical simulation in Case I

The $100\sqrt{2}$-m-long line source is settled on the horizontal plane, 10 m above the bottom wall. The coordinate of the middle point is (750 m, 500 m, 10 m). The inclined angle is $\pi/4$ from the positive *x*- and *z*-axes directions. In the numerical simulation, this line source is resolved by discrete cells that are connected along the vertical edge. With the increase of iterations, the boundary layer gradually became steady. The wind profile along the vertical edge in the middle of the outlet settled down eventually after 90000 iterations as shown in **Fig. 3.4**. In this research, the data of iteration 130000 is used to make sure the flow is in a steady state. The steady dispersion field of the line source is presented in **Fig. 3.5**. It can be confirmed from the concentration profile that the influence of the dispersion reaches up to 200 m height in the boundary layer.

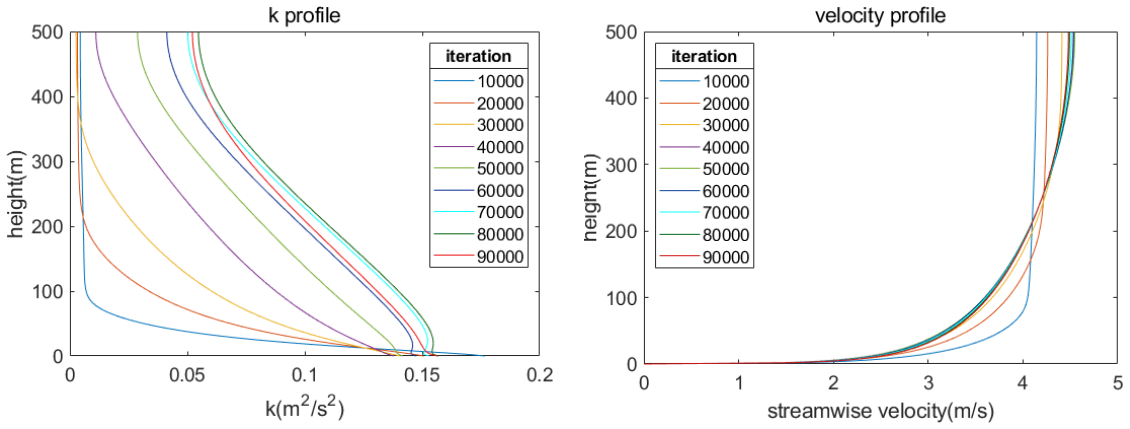
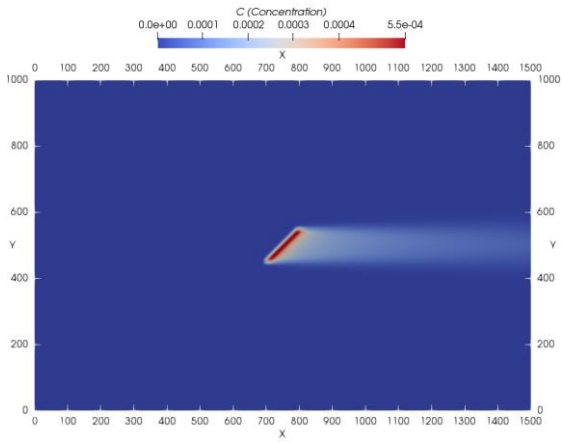
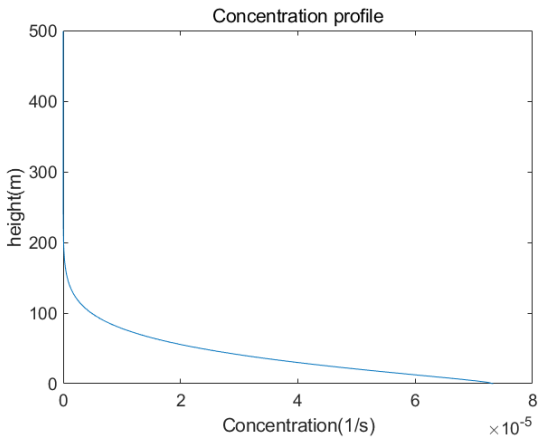


Figure 3.4. Profile along the vertical line in the middle of the outlet

(k: turbulent energy)



The dispersion field in the horizontal plane (height = 10 m)



The concentration profile along the vertical line in the middle of the outlet

Figure 3.5. Steady dispersion field of the line source

For the sensor configuration, a common setting in STE research is the uniform distribution in the domain (Keats et al., 2007b; Xue et al., 2018b). However, the distribution of stationary sensors cannot be uniform among complicated urban terrain in real life. Therefore, a random walk was added to the location of each sensor with a uniform configuration (**Fig. 3.6**); this could also test the robustness of the proposed method. The coordinate was changed from (x_m, y_m) to $(x_m + 10 \times \lfloor N[-10, 10] \rfloor, y_m + 10 \times \lfloor N[-5, 5] \rfloor)$. $N[\mathrm{a}, \mathrm{b}]$ represents the uniform distribution bound between a and b. The sensor heights are also different to simulate the real application. The adjoint concentration field was calculated in the same domain with a reverse velocity field for each sensor. One particular adjoint transport field from a sensor is shown in **Fig. 3.7**.

The simulated concentration at each sensor is regarded as measurement $\boldsymbol{D}$ in this case. As these concentrations are synthesized via the simulation, random noises $N[-0.5, 0.5] \times D_i$ are added to them to ensure that this case study is closer to reality.

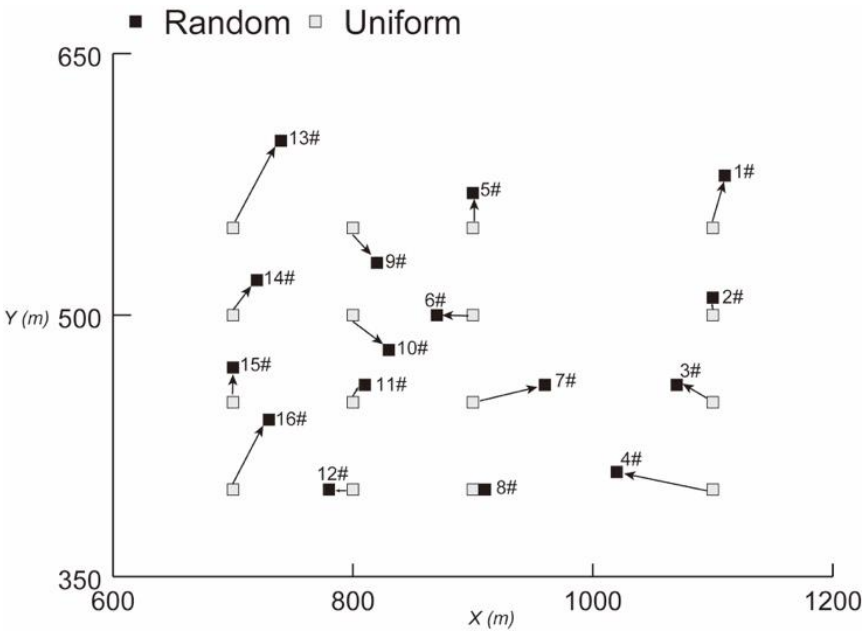


Figure 3.6. The sensor configuration in Case I

1

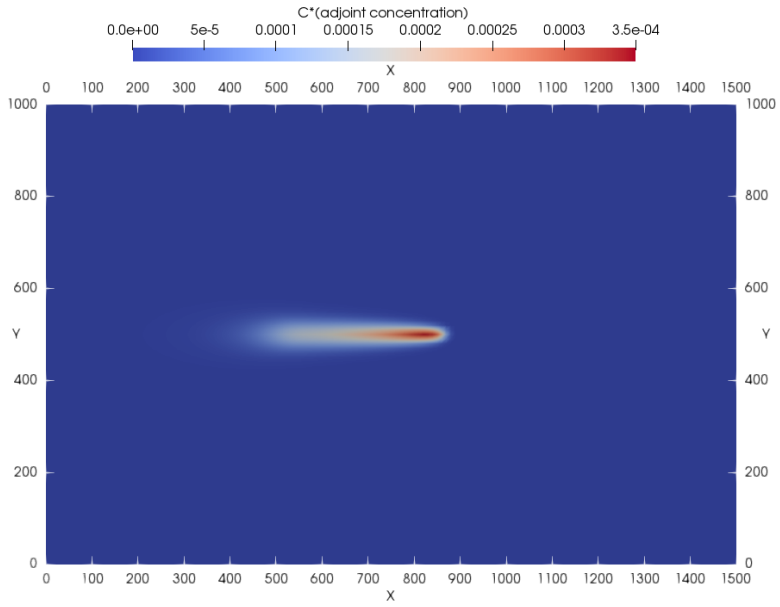


Figure 3.7. Adjoint transport field from a sensor in a horizontal plane (height = 10 m)

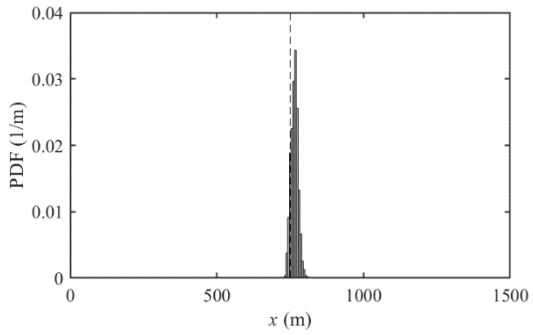
2

3.3.2 Bayesian inference settings

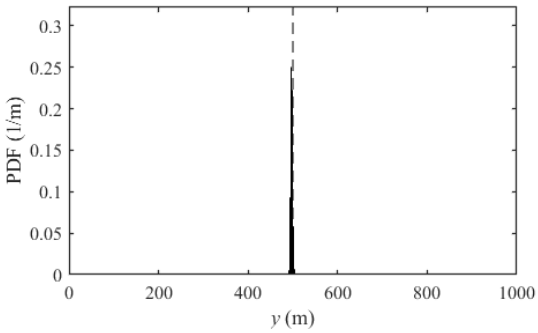
Bayesian inference was conducted after the simulation, as described in **Chapter 2**. In this process, one of the most critical factors is the error variance $\sigma_{d,i}^2 + \sigma_{m,i}^2$ in Eq. (2.8). The value of this factor subtly affects the effectiveness and accuracy of the estimation procedure. Conversely, the optimal value of this factor is difficult to determine and still in controversy. Keats et al. (2007b) assigned the values manually for each sensor, according to the observed trends. These values lie between 10% and 270% of the mean measurements. Xue et al. (2018b) applied the variance of the measurements of each sensor. Since the determination of this factor still requires further study and is not the core part of this paper, this research assigns values in the same way as Keats et al. (2007b); the values lie between 10% and 250% of the mean forward measurements. Meanwhile, because a minor variance could cause unsteadiness in the calculation of Eq. (2.8), the variances of the sensors that could barely measure concentration were assigned a fixed value. The Bayesian inference of STE was executed with MATLAB on a personal computer with Intel® Core™ i7-6700 CPU @ 3.4GHz and 16GB of RAM. The averaged computational time for estimating one source is about 42 min.

3.3.3 Estimation results

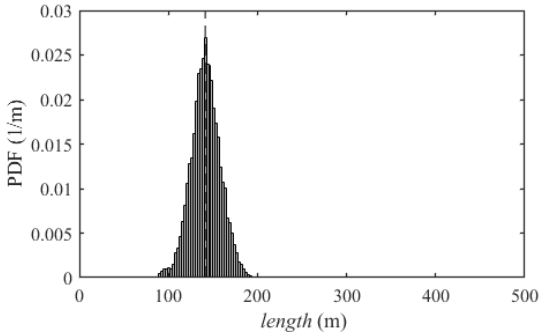
Fig. 3.8 shows the estimation results as the marginal probability distribution of each term. When the sensor measurements are combined with the super-Gaussian function, Bayesian inference can provide an ideal estimation for the line source, regardless of the initial guess provided to the Markov chain. All the parameters are correctly estimated. The coordinates, length, and strength estimations agree well with the true values. The peak value of width is no more than 20 m from the true value. Considering that the resolution of the simulation data is 10 m rough, the proposed method is considerably accurate. The discrepancy for the angle is also smaller than 10 degrees. Considering that the posterior probability density functions (PDFs) in the results are usually highly skewed, like **Fig. 3.8(d)**, only one estimator may not be able to objectively represent the distribution of estimations, we select both the 50th percentile and standard mean values of PDFs as the parameter estimators, the comparison between the true values and estimations are summarized in **Table 3.2**.



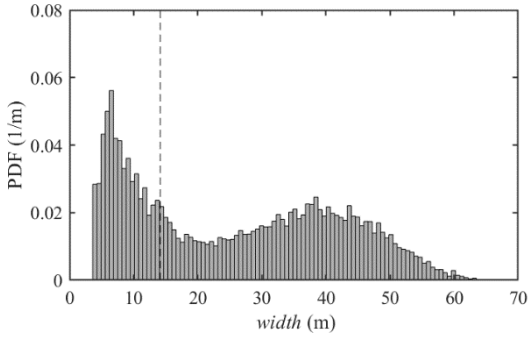
x-coordinate (middle point)



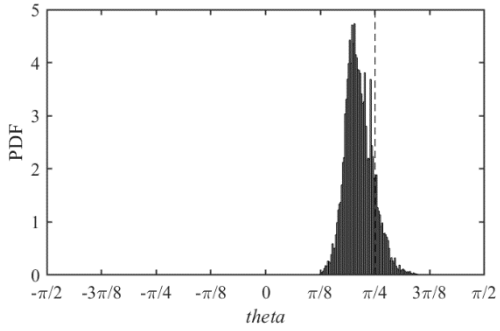
y-coordinate (middle point)



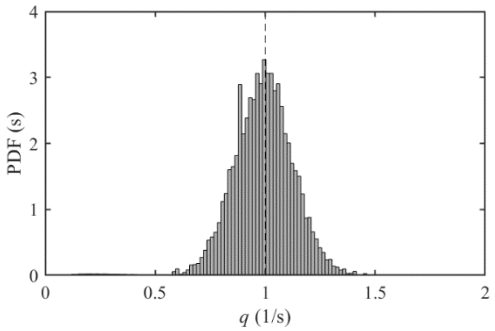
length of the line



width of the line



inclined angle of the line



strength of the source

Figure 3.8. Posterior PDFs of source parameters in Case I (dotted line: true value)

The expectation of line source geometry $\int p(\boldsymbol{s}|\boldsymbol{D},I) \times f(x,y|\boldsymbol{s})d\boldsymbol{s}$ is shown in **Fig. 3.9**. It can be confirmed that the proposed method restricts the estimation results to a small area around the true line. The shape of the contour is also an ellipse, with a similar line source angle. Therefore, it can be concluded that under ideal conditions, without measurement and modeling errors, the proposed method successfully identifies the line source information.

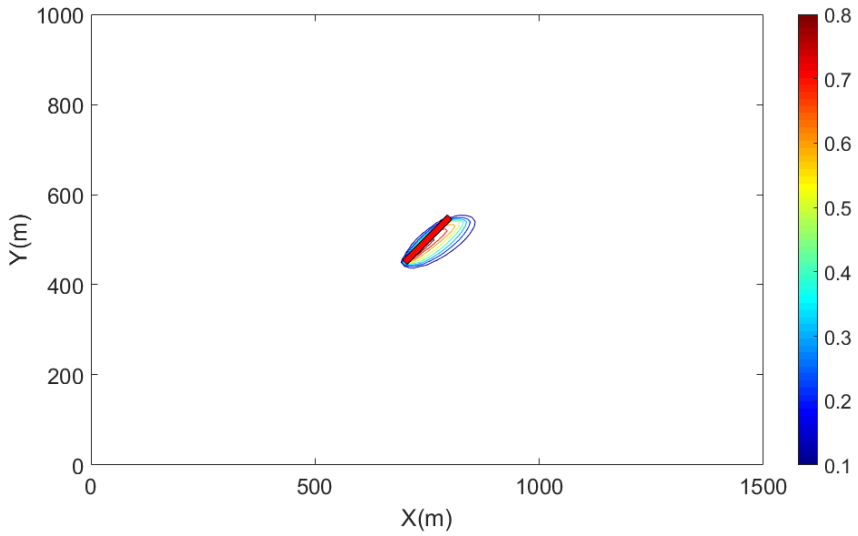


Figure 3.9. The expectation of line source geometry $\int p(\boldsymbol{s}|\boldsymbol{D},I) \times f(x,y|\boldsymbol{s})d\boldsymbol{s}$ estimated by the proposed method in Case I. (red patch: true line source)

Table 3.2. Summarized estimation results of the proposed method

Method	x_s(m)	y_s(m)	length(m)	width(m)	angle	strength(1/s)
True value	750	500	$100\sqrt{2}$	$10\sqrt{2}$	$\pi/4$	1
Estimations (50^{th} percentile)	765	496	142	25	$\pi/4.7$	0.99
Estimations (standard mean)	764	496	142	26	$\pi/4.7$	0.99

3.4 Discussion about sensor configuration

Sensor configuration is crucial for all STE methods as sensor measurements are essentially the only information on which the STE method is based. Considering that there are limited discussions concerning the sensor configuration for line source estimation in the literature, it is

necessary to study the basic properties of the line source's dispersion. Furthermore, the author wants to analyze the success and errors of line source estimation using different sensor configurations based on Case I. Notably, the contents in this section apply to all the line source identification methods based on stationary sensors, rather than to a specific design for the current method.

3.4.1 Uniform sensor configuration

In previous research regarding STE methods for point sources, a regular sensor network downstream of the source has commonly been used, yielding successful estimations. However, the situation has entirely changed in estimating the line source as it is more concerned with geometric information, which makes STE more difficult with a conventional sensor configuration.

In the first configuration, the sensor network is like the conventional ones used for the point source estimation in previous research studies (Keats et al., 2007b; Kumar et al., 2015b; Xue et al., 2018b). Only one sensor is in front of the source and all the others are at the downstream with regular positions. The schematic diagram is in **Fig. 3.10**.

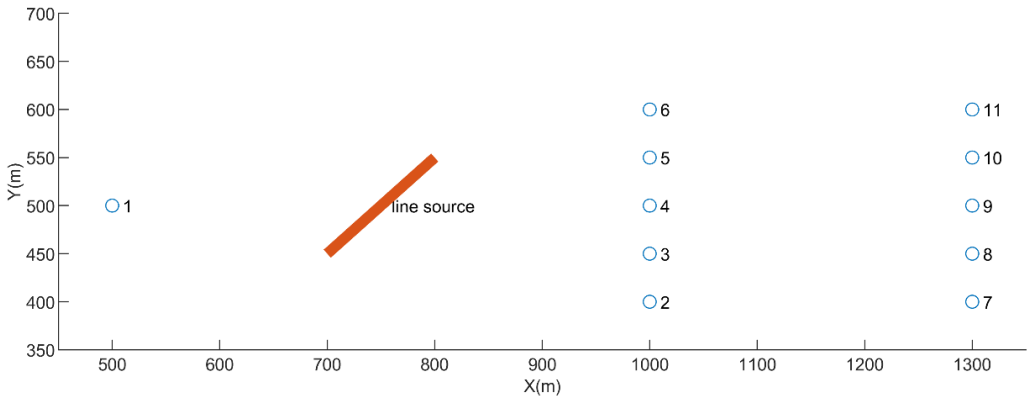
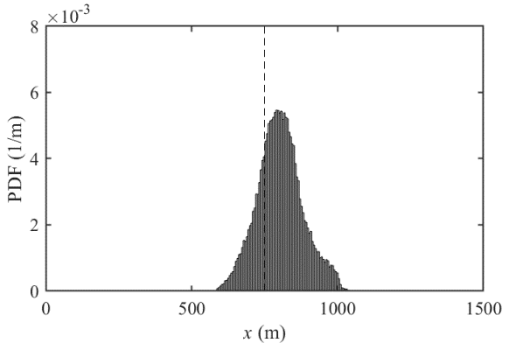


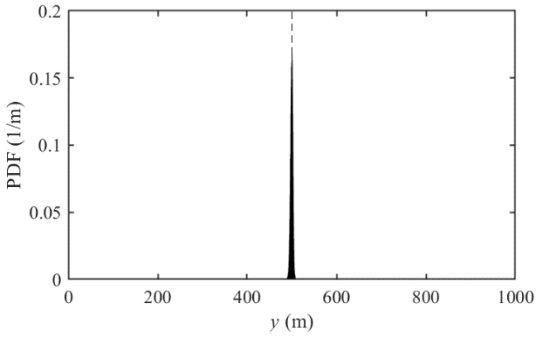
Figure 3.10. Schematic of uniform sensor configuration. (red patch: true line source)

Fig. 3.11 shows the marginal parameter distributions of the posterior probability. It can be noted that the estimation of the middle point coordinate, strength, and length is proper with small discrepancies from the true value. There is only one peak in each histogram. On the other hand, the angle and width results are totally distorted. It seems that the sampling algorithm

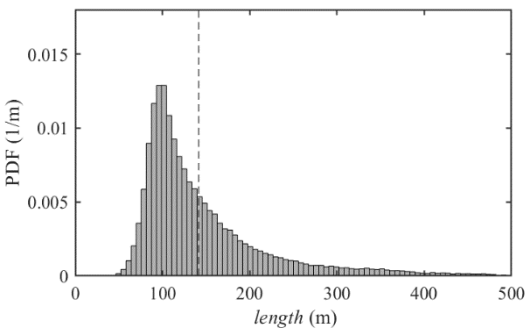
randomly travels in the width space, losing the ability to identify the true value. The accuracy is even worse in the angle estimation, in which the result is totally opposite to the true value. In consequence, the proposed method, coupled with the conventional sensor set, fails to estimate all the parameters of the line source, especially in estimating the angle and width.



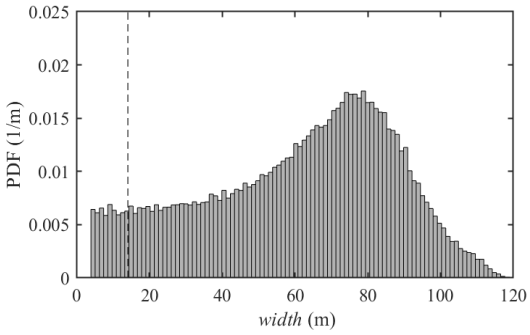
x-coordinate (middle point)



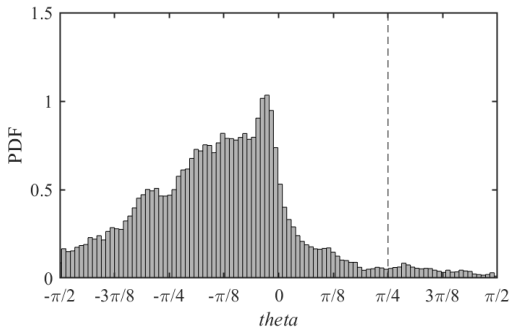
y-coordinate (middle point)



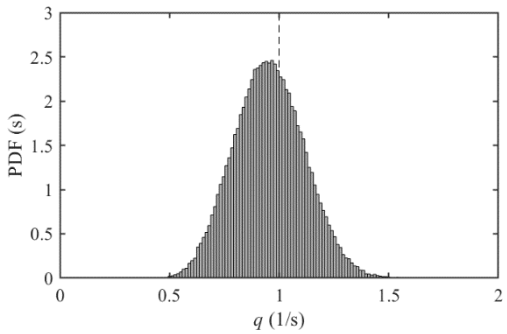
length of the line



width of the line



inclined angle of the line



strength of the source

Figure 3.11. Posterior PDF of source parameters with uniform sensor configuration.

(dotted line: true value)

Since the conventional configuration has been proved effective enough by previous research studies, and the proposed method has successfully identified the line source in **Section 3.3**, the problem should be in the different properties between the point source and line source. Actually, if we check the results of the conventional configuration in **Fig. 3.11**, even though the angle and width estimations are completely wrong, the coordinates of the middle point are still correctly estimated. This fact implies that the conventional configuration is capable of locating the position of the source, which is enough for a point source estimation. If geometric information is needed, like the inclined angle of the line, the conventional set cannot be relied on anymore. This can be illustrated by the three different line sources located at the same position as shown in **Fig. 3.12**. Although these three line sources are distinct from each other, their dispersion fields, caused by the same flow and the influence on the sensor downstream, are similar to a high degree. The only difference that can be captured by the downstream sensors is the concentration discrepancy in the span-wise direction, which is vulnerable under the noise of the model and measurement. Therefore, the inclination angle of the line source cannot be estimated through the measurement of the regular sensor configuration downstream of the source. For the same reason, the width is also free to change and is thus impossible to estimate. Accordingly, the estimation of the line source requires more geometric knowledge, which can be provided only by the sensors close to the circular area, whose diameter is the same as the length of the line source.

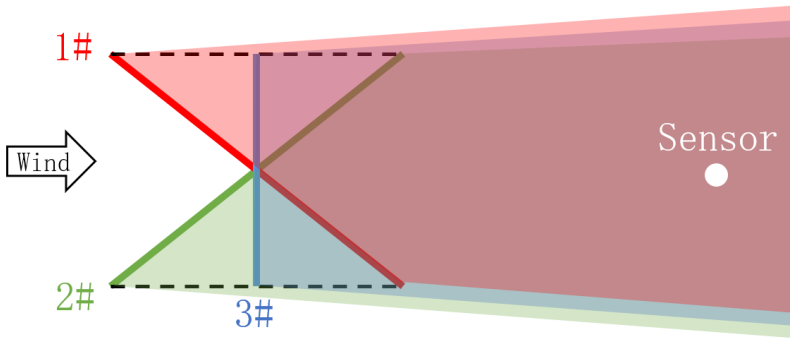


Figure 3.12. Three different line sources with similar concentration fields (areas with different colors)

The inclination angle and width were successfully estimated in Case I. It is reasonable to believe that the sensors around the line source (No. 9, 14, 15, 16) are the key factor. During the inference, the information provided by these sensors prevents the generated source from over-

rotating or expanding in width. Despite of that, the sources generated during sampling still fluctuated to some degree because the density of sensors is not high enough. It can be observed in **Fig. 3.8** that second peak appeared in the posterior PDF of width, and the estimation of angle also has a deviation. It means that a wider line source with smaller inclined angle would have similar impacts on the current sensor configuration, which is intuitively shown in **Fig. 3.9**.

3.4.2 Importance of null measurements

Apart from these indispensable sensors near the source, the effects of the "region of influence" should also be considered. This concept was proposed by Keats et al. (2007b), emphasizing the importance of sensors with null measurements to the point source estimation. The estimation of the line source has the same requirements. Considering the case shown in **Fig. 3.13**, whereby 9 sensors are used to estimate the red line source, the incorrect estimation, i.e., the green line, will destroy the inference because the red and green lines have almost equal influence on the sensors. The failure is essentially due to the lack of sensors in the upper-right area; although sensors in this area will measure none of the true red line, only the null measurement informs the inference of the length of the source and excludes the possibility that the green line is the true source. According to Keats et al. (2007b), this kind of sensor should also be classified in the 'region of influence'.

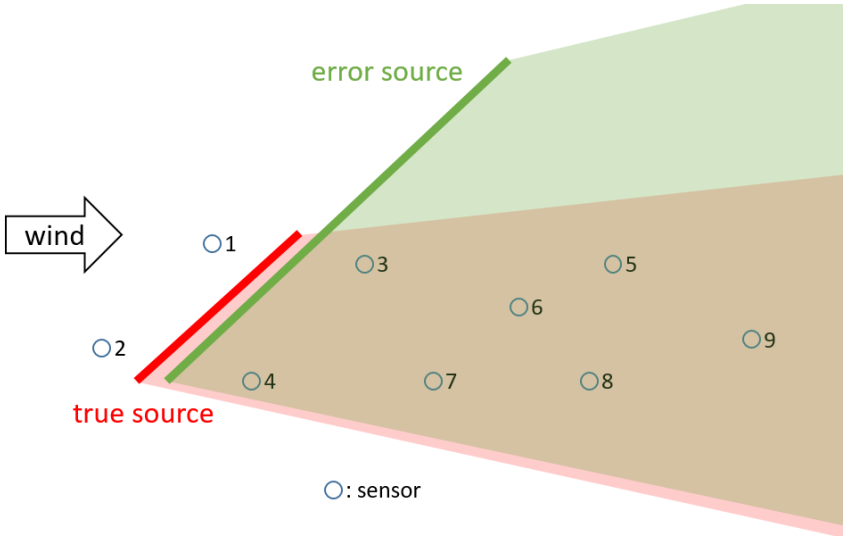


Figure 3.13. Diagram of typical estimation error caused by the lack of null measurement

This characteristic can be better demonstrated by the following trial with Case I. This

sensor configuration is totally produced by randomness. The horizontal distribution of sensors for the trail is shown in **Fig. 3.14**. The configuration has several sensors around the line source, which should improve the estimation accuracy when compared to the uniform configuration.

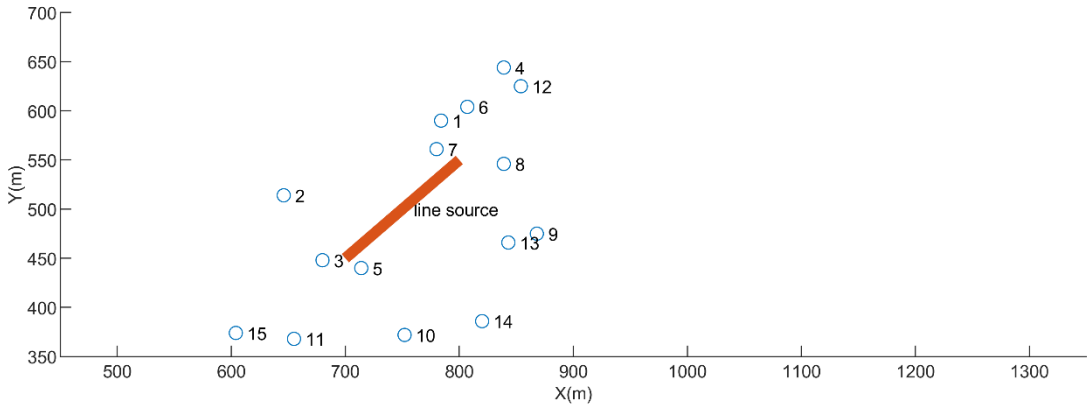
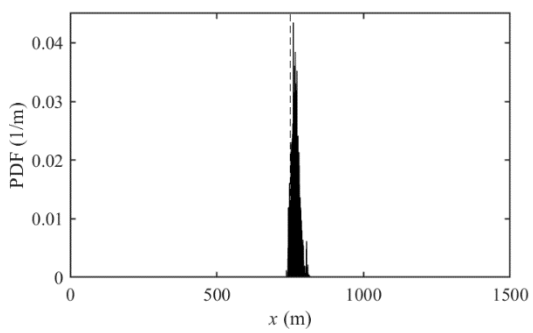


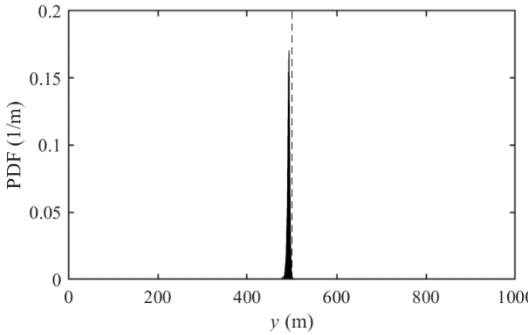
Figure 3.14. Schematic of sensor configuration (totally random)

The estimation result is shown in **Fig. 3.15**. From the first look, all the parameters are almost correctly estimated. The accuracy of the coordinates, length, and strength is the same or even better than that of conventional configuration. More importantly, the inference successfully narrows the width scale to under 25 m. Similarly, the estimation of the angle is improved significantly using the measurements from this configuration. The inference eliminates the possibilities of other angles and gathers only around the true value. It is reasonable to believe that the sensors around the line source (No. 3, 5, 7, 8) cause the improvement. During the inference, the information provided by these sensors prevents the generated source from rotating too much or expanding the width.

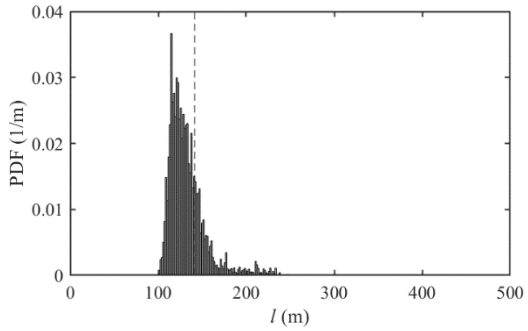
Despite this advancement over the conventional configuration, the Bayesian inference will be found unsteady and sensitive to the initial guess if different Markov chains with different start points are tested. Although for most of the time, the estimations are similar to the one in **Fig. 3.15**, one particular error in **Fig. 3.16** will appear repeatedly from time to time. In this particular error, wrong peaks show up in the x coordinate as well as the length and dominate over the true peak, which is unacceptable in practical use. Hence, the measurements provided by this configuration are still not enough to reach a steady and accurate inference.



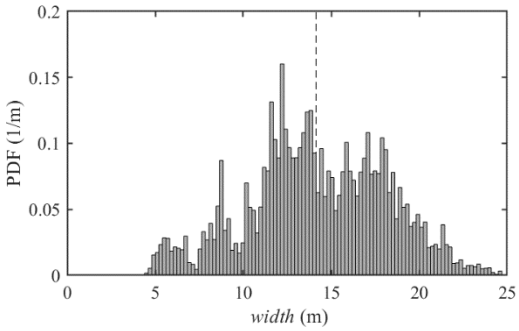
x-coordinate (middle point)



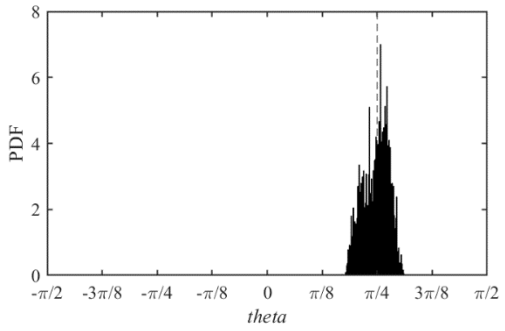
y-coordinate (middle point)



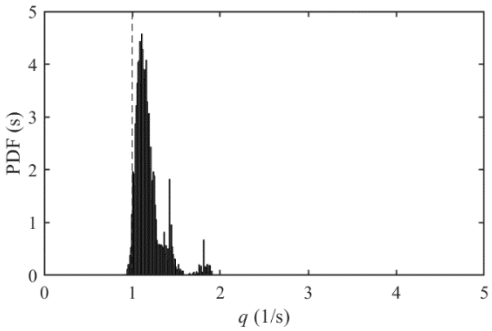
length of the line



width of the line



inclined angle of the line



strength of the source

Figure 3.15. Posterior PDF of source parameters with sensor configuration of **Fig. 3.14**.

(dotted line: true value)

1

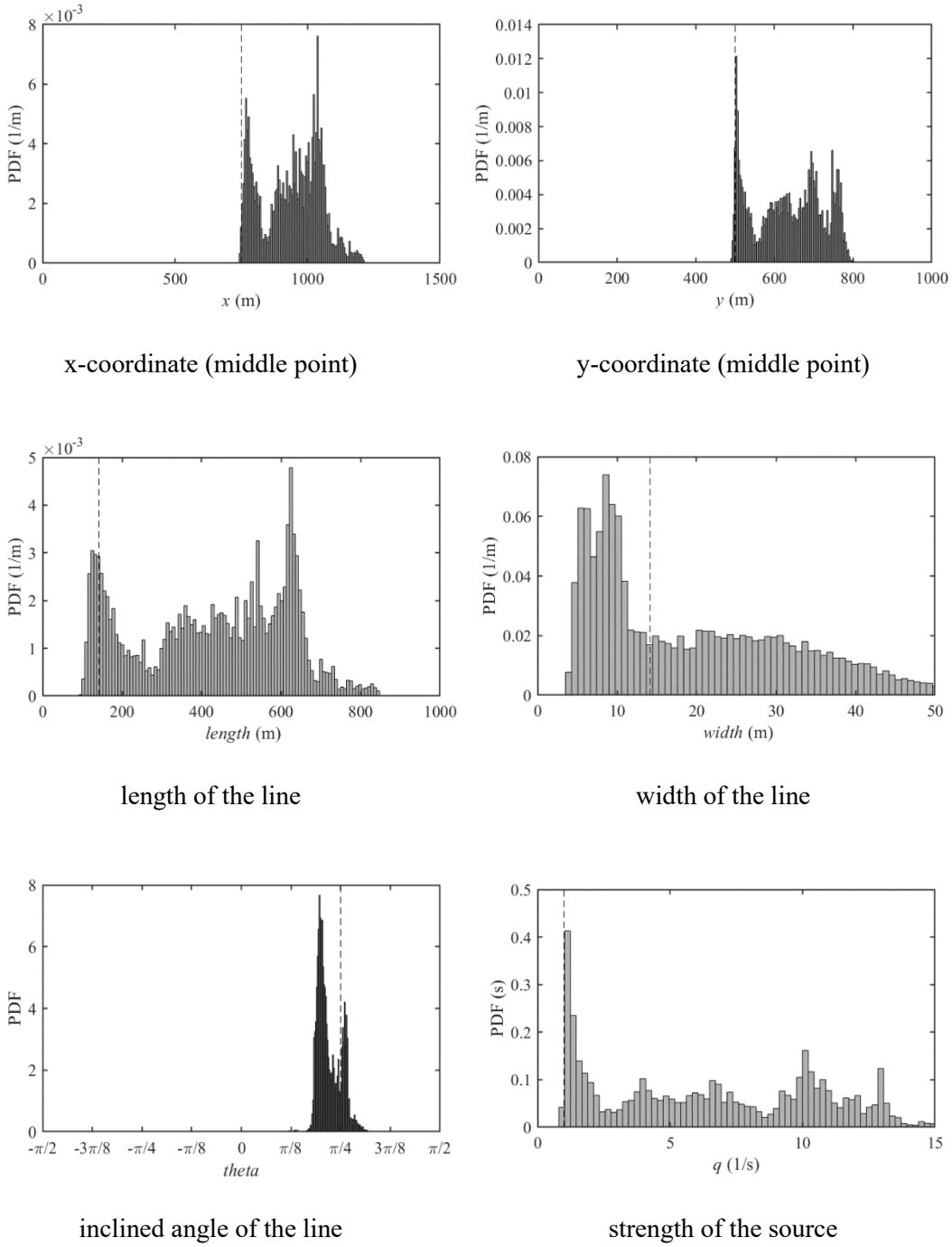


Figure 3.16. Posterior PDF of source parameters with sensor configuration of **Fig. 3.14**. (a particular wrong example) (dotted line: true value)

The failure of the length estimation is mainly caused by the lack of sensors at the upper right corner downstream to the line source. In this case, the line source can extend freely in that direction without significantly changing the influence on each sensor. One particular error is shown in **Fig. 3.17**. Consequently, apart from the sensors near the line source, we still need

some null-measurement sensors to prevent the error extension.

In **Section 3.3**, the length was well reconstructed. The zero measurements provided by sensors No. 1, 4, 5, 8, and 12 prevent error extension. Otherwise, the difference between the measured and modeled values would become significant, and the sampling algorithm (MHMC in **Chapter 2**) would abandon the sampling.

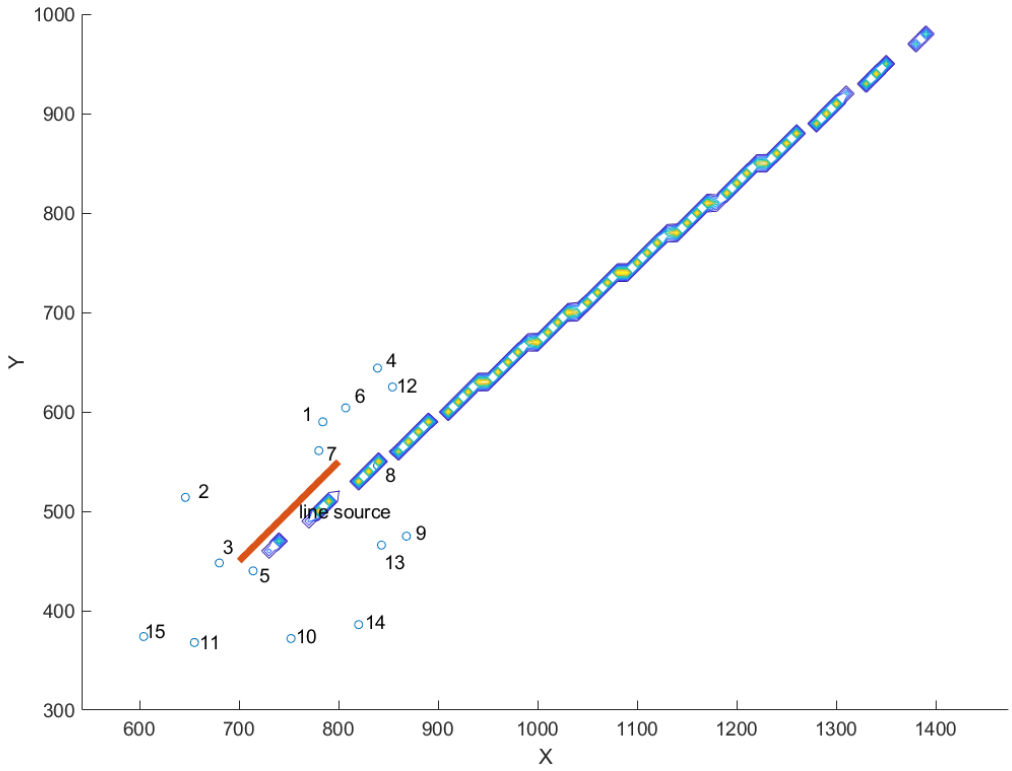


Figure 3.17. Wrong estimation with sensor configuration of **Fig. 3.14**.

(red patch: line source; blue line: wrong estimation)

3.4.3 Necessary sensors for line source estimation

It is natural to ask how can we know a set of sensors is enough for the estimation or not? Or, what is the minimum number of the necessary sensors for the line source estimation? Based on the analysis above, it is proposed here that at least six sensors located appropriately are enough for the purpose. The schematic diagram is in **Fig. 3.18**. Among them, three sensors around the source are mainly for the angle and width estimation. The upper one and lower one in the downstream can restrain the length. It does not mean that each sensor is only in charge

of the estimation of one parameter. In fact, all the sensors are helpful to the estimation of several parameters, but certain sensors are more important to the estimation of certain parameters.

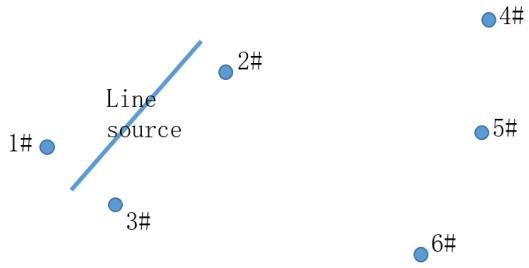
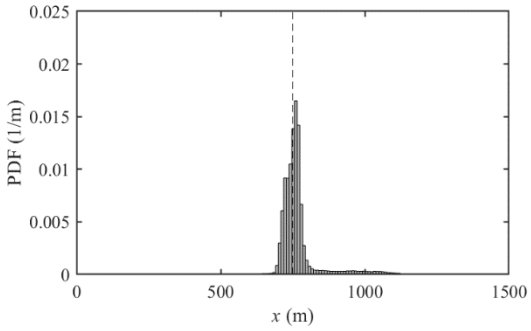
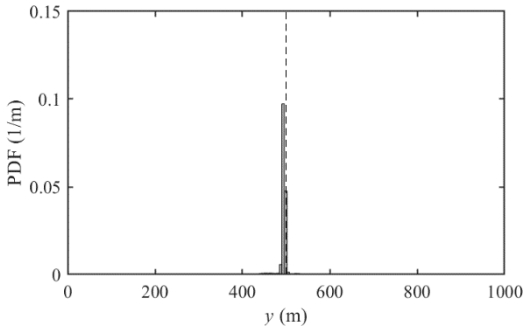


Figure 3.18. Schematic of the necessary sensors for the line source estimation

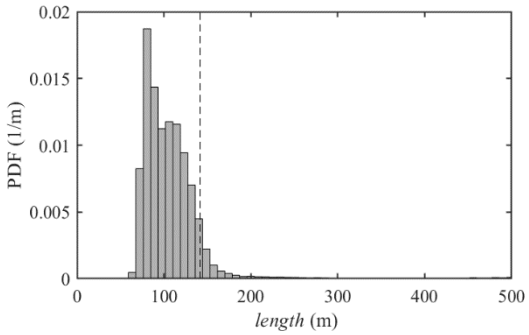
In order to prove the correctness of the proposal above, six sensors (No. 1, 3, 4, 9, 15, 16) are picked out from the configuration of Case I in **Section 3.3** according to **Fig. 3.18** to conduct a new STE. The results are summarized in **Fig. 3.19**. Apparently, the accuracy is almost consistent with that of Case I in **Fig. 3.8**. All the parameters are successfully estimated with tolerable error. Without a doubt, the samplings in **Fig. 3.8** gathered tightly around the true value while in **Fig. 3.19** more residual probability appears at wrong values. The redundant 10 sensors in Case I infuse more information to the inference so that the true value is captured effectively. This fact is also well illustrated by the cumulative probability density of six sensors in **Fig. 3.20**. Thereby, we can conclude that, in order to estimate the line source term, both for position and geometry, six sensors with proper location are indispensable. Any sensor network including these six sensors can perform well in line estimation cooperated with the super-Gaussian function method. Apart from these six sensors, the more sensors we have, the more effective and accurate will the inference be.



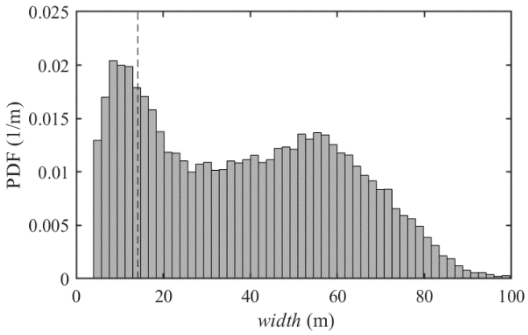
x-coordinate (middle point)



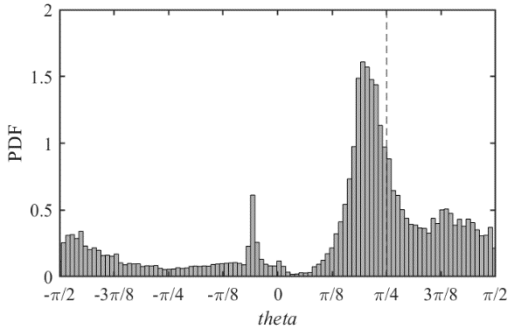
y-coordinate (middle point)



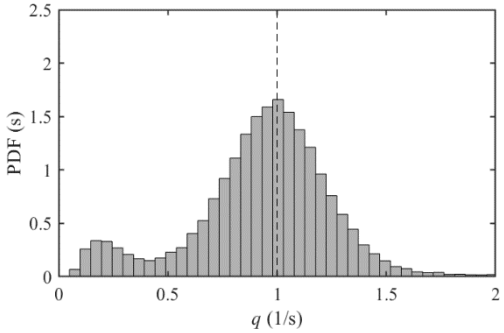
length of the line



width of the line



inclined angle of the line



strength of the source

Figure 3.19. Posterior PDF of source parameters (selected 6 sensors from Case I).

(dotted line: true value)

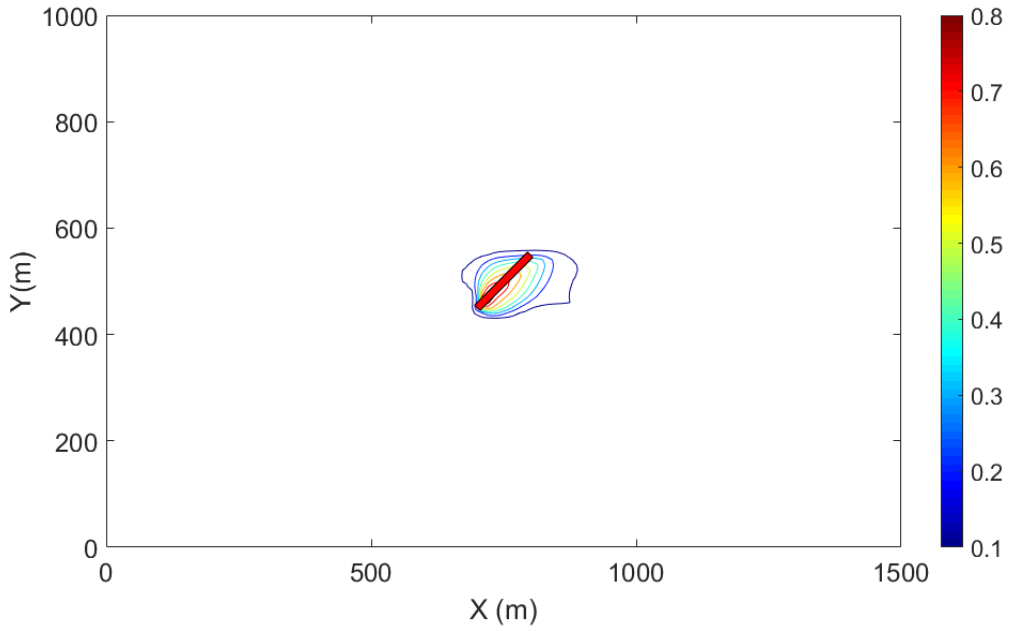


Figure 3.20. The joint posterior probability density distributions $P(x_s, y_s, \theta, \sigma_X, \sigma_Y | D, I)$ of selected 6 sensors. (red patch: true line source)

3.5 Case II: wind tunnel experiment of the urban square model

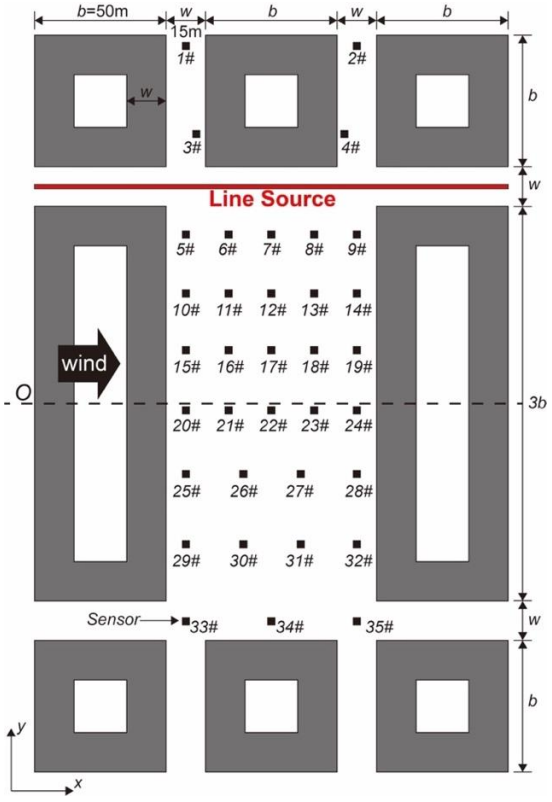
After the success with the elementary case presented above, it is necessary to examine the robustness of the proposed method under the complicated measurement and modeling errors within the practical case. Although the line source's dispersion has been frequently studied in the literature, there are very few experimental databases accessible online for validating the inversion methodology. Balczó and Lajos (2015) measured the dispersion of a line source in a simplified urban square, as shown in **Fig. 3.21(a)**. Their case includes buildings of different sizes, and the configuration is representative of real urban neighborhoods. In addition, they performed a comprehensive measurement in the space between buildings, which makes the STE possible. Hence, we demonstrate the feasibility and necessity of the proposed method based on this practical case.

Meanwhile, to illustrate the necessity of geometry estimation, we also conducted the STE via the conventional method, with the ideal-point assumption. The results are compared with those obtained using the proposed method.

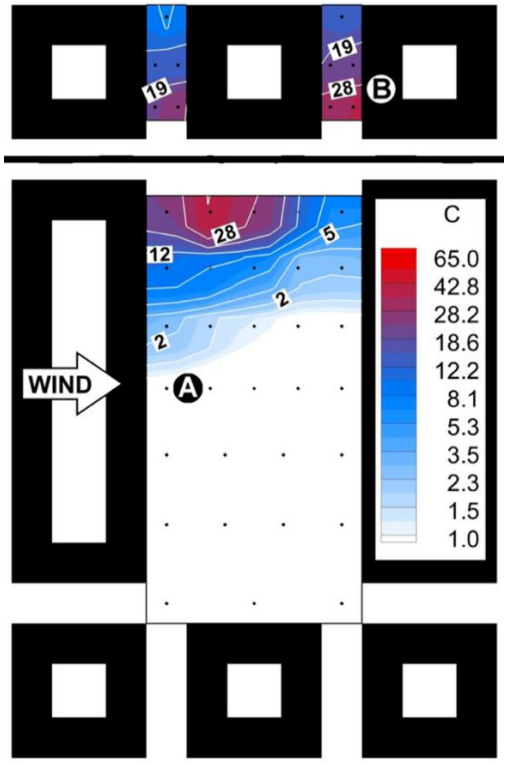
3.5.1 Wind tunnel experiment for measurements

In this case, measurements $\boldsymbol{D}$ were obtained via the wind tunnel experiment in Balczó and Lajos (2015), conducted in an Eiffel-type wind tunnel with a closed test section of $0.5\ m \times 0.5\ m$. All buildings had the same height. The model had a scale of 1:650, which means the building height was $H = 30$ m at full scale and $h = 46$ mm at the model scale. The turbulent boundary inflow generated by roughness moved from left to right (**Fig. 3.22**). The reference velocity U_r was measured at the building height.

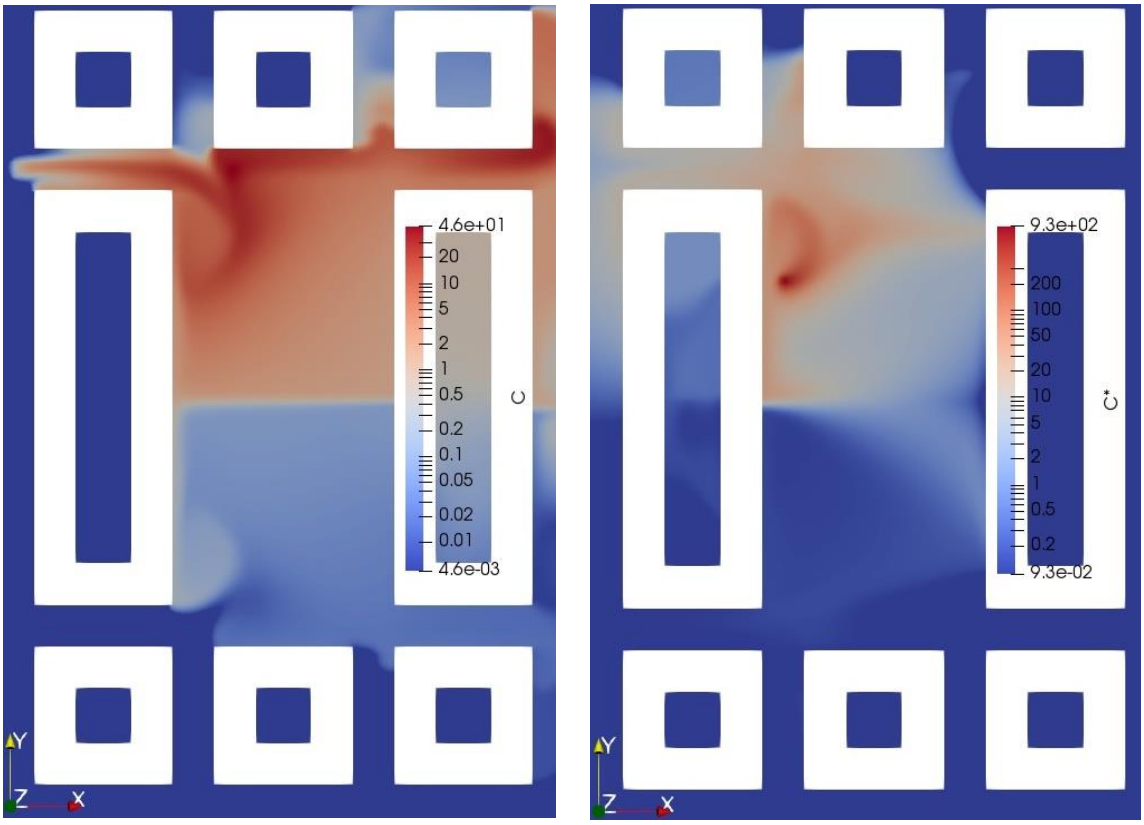
The line source was located on the bottom wall in the upper part of the building area, as shown in **Fig. 3.21(a)**. The gas emitted from the line source was measured using slow flame ionization detector sensors located at 35 positions in a horizontal plane with a height of $H/10$. The measured concentration was non-dimensionalized via the reference concentration $\bar{C} = \frac{Q}{U_r h^2}$, where Q[g/s] is the source strength. The non-dimensionalized concentration distribution in the horizontal plane of sensors is shown in **Fig. 3.21(b)**.



The basic settings for the simplified urban square model



The measured distribution of concentration emitted from the line source in the WTE



The simulated distribution of concentration emitted from the line source

The simulated adjoint tracer field emitted from the No. 15 sensor

Figure 3.21. Schematic of the wind tunnel experiment and simulation in Case II (horizontal plane with $z = H/10$): (b) is referred to Balczó and Lajos (2015).

3.5.2 Numerical simulation

A CFD simulation was conducted to calculate the flow field and adjoint tracer field. The calculation domain was $26H(x) \times 10.67H(y) \times 6H(z)$. The building area was $10H$ away from the inlet and outlet and $5H$ from the top boundary. The experimental model was close to the side walls in the small wind tunnel and thus the distance between the building area and lateral boundaries was also set small, as $0.67H$, to simulate the effects of side walls. In the building area, the cubic mesh was uniform, with a $H/18$ (1.67 m) length in all directions, then expanded along with the coordinates at a 1.08 ratio.

The boundary conditions were set according to the AIJ basic guidelines (Tominaga et al., 2008b). The streamwise velocity and turbulence intensity profiles were prescribed on the inlet

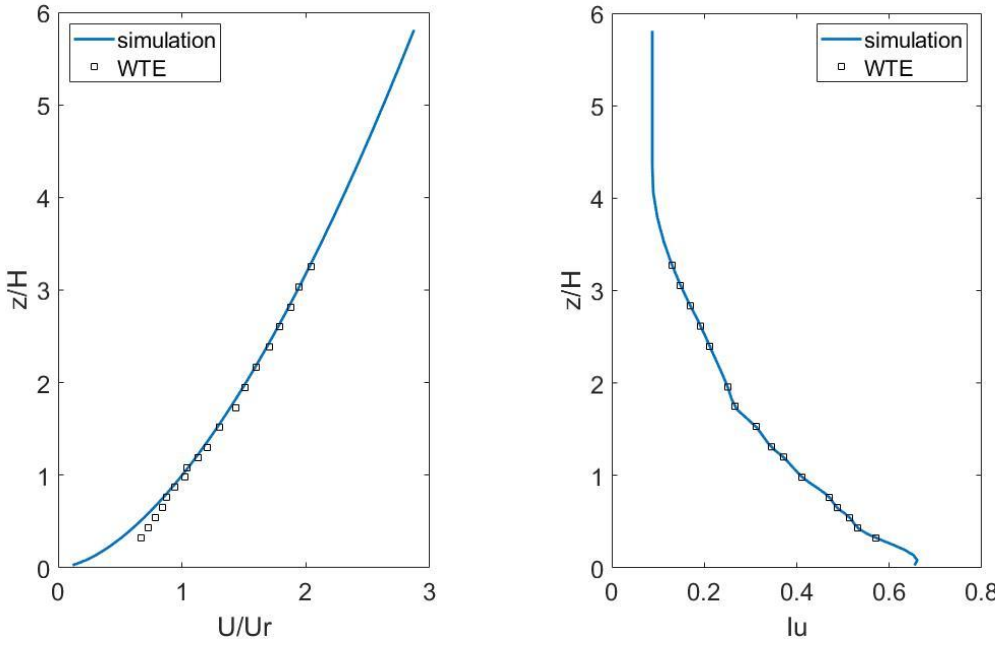
based on the experimental data, as shown in **Fig. 3.22**. At the bottom and side walls, a no-slip condition with the Spalding wall function was applied. The turbulence was modeled using the renormalization group $k-\varepsilon$ model (Yakhot et al., 1992). The details of other configurations are listed in **Table 3.3**. The adjoint transport field was calculated in the simulation domain, with a reverse velocity field for each sensor in the wind tunnel experiment. One adjoint transport field from the No. 15 sensor is shown in **Fig. 3.21 (d)**.

Table 3.3. Numerical schemes and boundary conditions for case II

Time marching	Steady state (SIMPLE method)
Spatial discretization	Advection term: TVD scheme;
Inlet	$\overline{U}, k, \varepsilon$: experimental profile; Pressure: zero-gradient C: constant (= 0); C^*: zero-gradient
Outlet	Flow: zero-gradient; Pressure: fixed value 0; C: zero-gradient; C^*: constant (= 0)
Top wall	Slip;
Bottom & side walls	Flow: Spalding wall function; Pressure: zero-gradient; C, C^*: zero-gradient
Source	C, C^*: Constant injection rate

In the numerical simulation, apart from the hypothetical tracers, the dispersion from the line source was also predicted to reveal the model error. A comparison between the measurements and simulation can be observed in **Fig. 3.21(b) & (c)**. The general distribution was predicted accurately via simulation; the pollutant mainly accumulated in the upper part of the square, while the concentration in the lower part was very low. The high pollutant concentration downstream of the building area was not reflected in the measurement, owing to the limited number of sensors. One evident simulation error occurred in the wake region behind

the buildings in the upper part of the domain. The simulation failed to predict the convection of pollutants in this area, which may lead to noticeable modeling errors during Bayesian inference. This concentration underestimation was also noted by Balczó and Lajos (2015), suggesting that the steady simulation based on Reynolds averaged Navier-Stokes equations still needs improvement to accurately predict dispersion in complex urban areas.



Streamwise velocity Turbulence intensity

Figure 3.22. Vertical profiles of streamwise velocity and turbulence intensity for inflow in Case II

3.5.3 Bayesian inference settings

As in Keats et al. (2007b), the values of $\sigma_{d,i}^2 + \sigma_{m,i}^2$ lie between 10% and 250% of the mean sensor measurements in the wind tunnel experiment. The variance of the sensors with minimum concentrations was set as a small fixed value to maintain steady Bayesian inference. For prior information, this research applied a common assumption in STE that the source is in the main area $x[-10m, 190m] \times y[-150m, 150m]$ and will not appear in the buildings.

3.5.4 Estimation results

The estimation results of the proposed method for a single term are shown in **Fig. 3.23**. In

general, for each term, except length, the peak probability distribution value is close to the true value. The probability distribution for coordinate x_0 of the middle point demonstrates that the reconstructed source moves unsteadily along the *x* direction. The result for y_0 appears better as the bulk of the probability mass concentrates in several grid cells around the true source. The zero measurements provided by sensors No. 17 ~ 35 limit the sampling of y_0 to move downward; if not, the difference in the measured and modeled value would be too large to invalidate the sampling. Samplings in θ tightly surround the true value, with small discrepancies. Besides, the peak value of the width is no more than 2 m away from the true value. Again, the estimation is considerably accurate since the resolution of the simulation data is approximately 1.67 m. The sensors around the line source (No. 3~9) help the algorithm successfully estimate the angle and width.

There is a deviation in the estimation of source strength. It should be mentioned that the scale of the horizontal axis was set small in the results of width and strength to clearly show the distributions. Hence, the uncertainty seems to be amplified. The estimation for length has a wide distribution range, and the estimated peak value has a relatively large bias compared with the true value. Combining the fact that x_0 estimation also moves along the *x*-direction, it can be concluded that the reconstructed source stretches and moves unsteadily in the *x*-direction during inference to find the most probable sampling. The reason for this behavior could be the lack of sensors along with the *x*-axis upstream and downstream of the source, which is exactly the null-measurements discussed in **Section 3.5.2**. These sensors are sensitive to the change in length; thus, their measurements can provide accurate information on length for the estimation process.

Combining the estimation performance in two cases, it can be concluded that line source estimation has a stricter requirement for sensor configuration than point source estimation, to obtain the geometric information. Sensors near the source and sensors in the 'region of influence' with zero measurements are all necessary. As a result, in practical applications, the density of the sensors in urban areas should be high enough to capture all possible unknown line sources, which could appear anywhere.

Admittedly, the model error of steady CFD simulation for the source-receptor relationship, especially the concentration difference in the wake region behind buildings (sensors No. 1~4), affects the accuracy of estimation. Hence, it is necessary to improve the

adjoint concentration simulation to ensure that the STE method is accurate enough in complicated real applications. Even with model error, the proposed method restricts the sampling between 0.5 and 2 times the true value, which is acceptable.

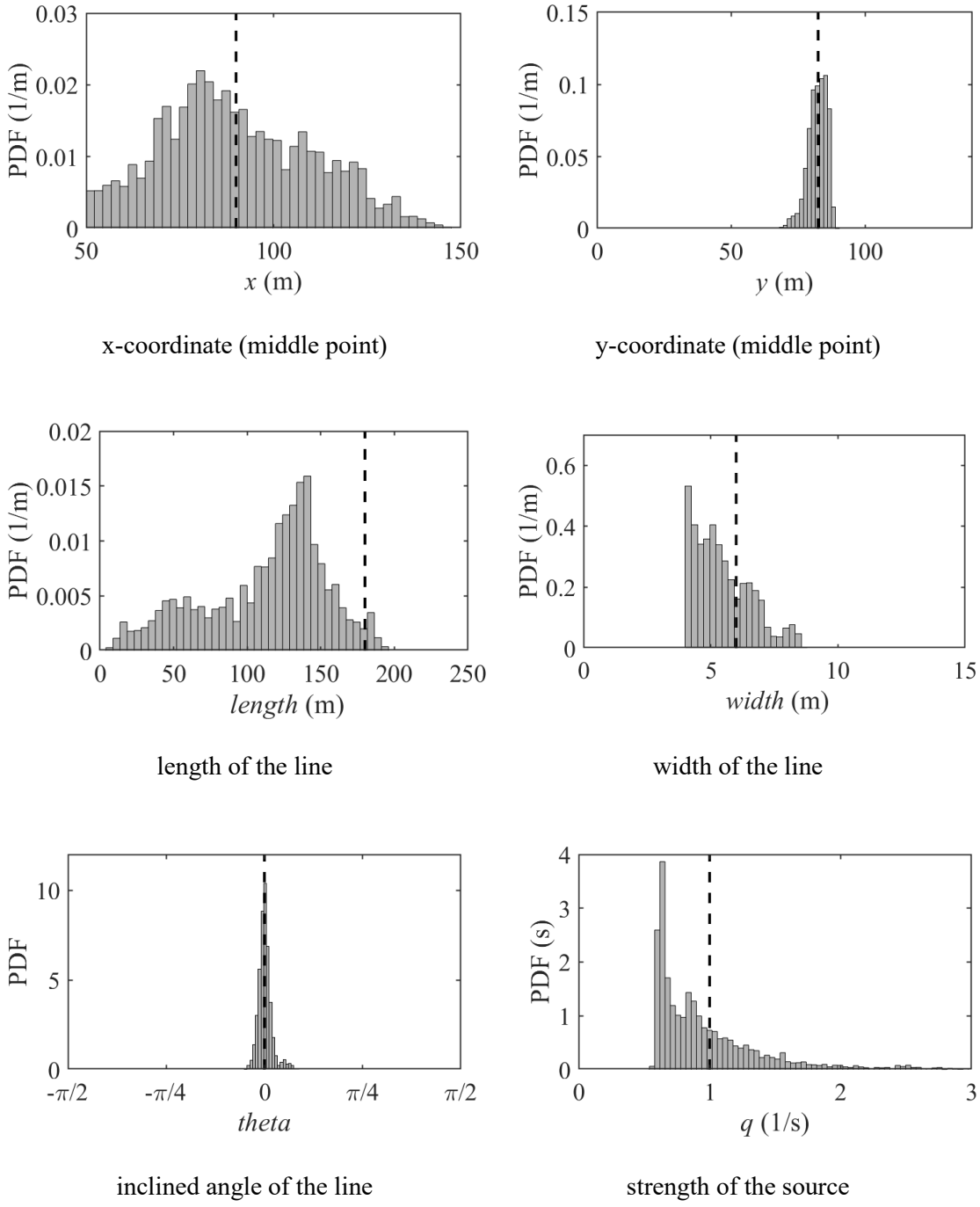
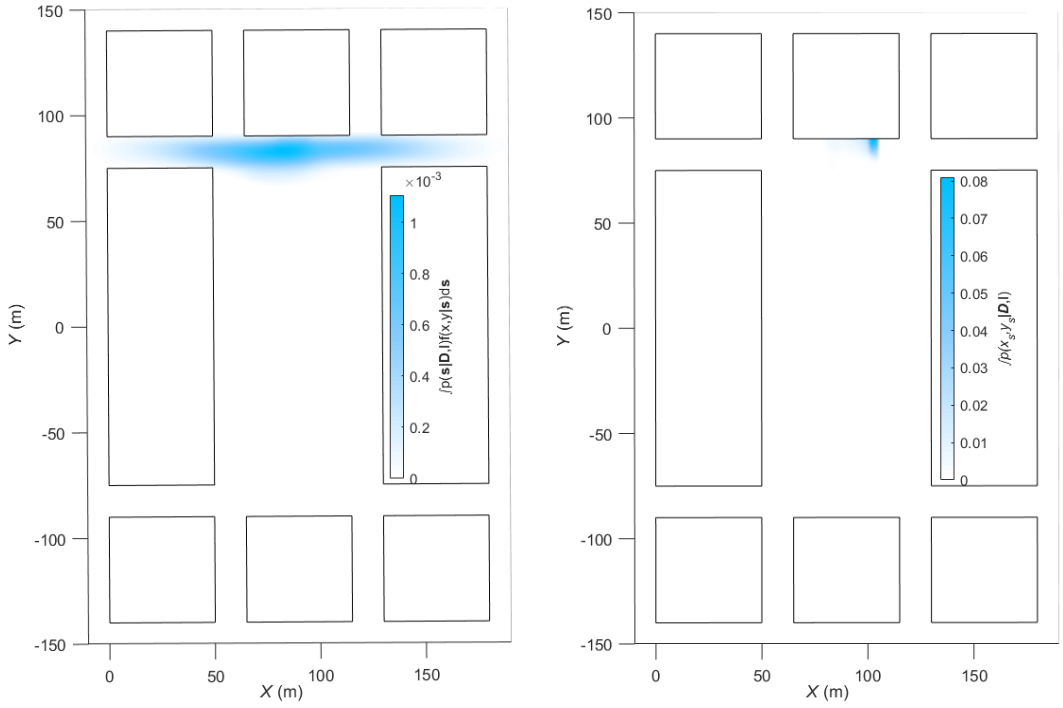


Figure 3.23. Posterior PDF of source parameters estimated by the proposed method in Case II.

(dotted line: true value)

The power of the proposed method can be better illustrated by the estimated line source geometry expectation $\int p(\boldsymbol{s}|\boldsymbol{D},I) \times f(x,y|\boldsymbol{s})d\boldsymbol{s}$, as shown in **Fig. 3.24 (a)**. Although the estimations for a single term are not perfect, the joint probability distribution clearly reflects that the target is a line-shaped source, with a geometry that is very similar to the actual geometry. The proposed method constricts the distribution into an area around the true line while the possibility of other areas is eliminated by inference. Therefore, it can be confirmed that the proposed method effectively estimates the geometry information of the source in this practical case.



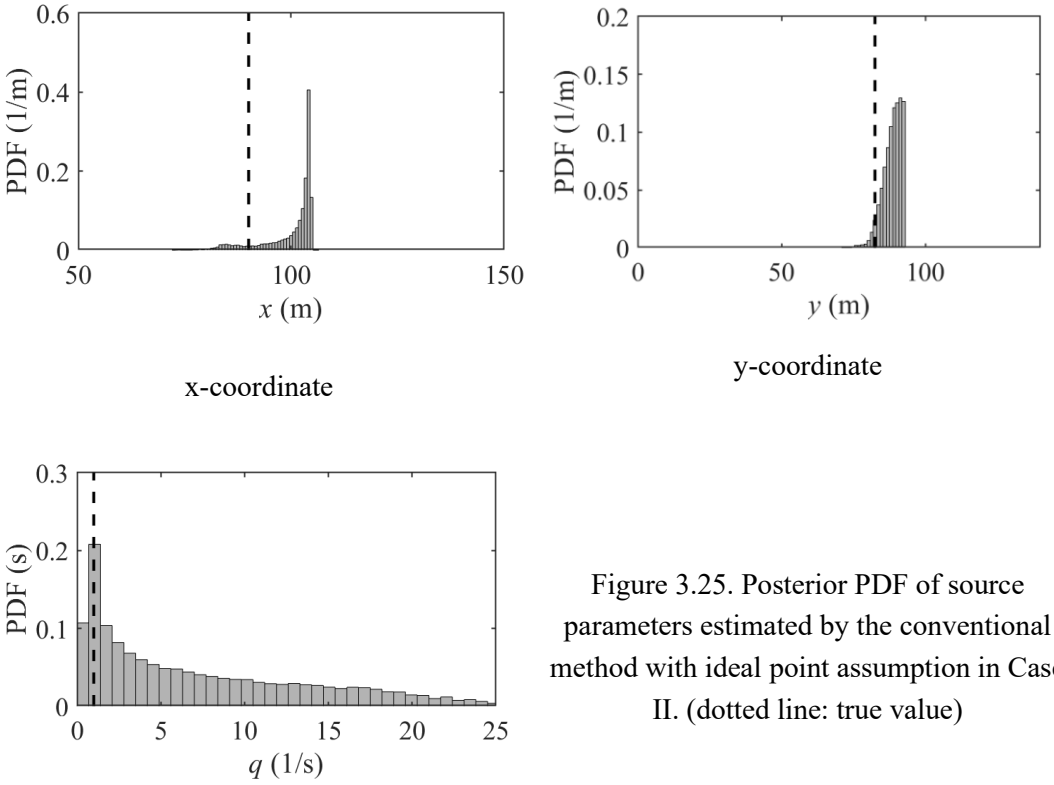
$\int p(\boldsymbol{s}|\boldsymbol{D},I) \times f(x,y|\boldsymbol{s})d\boldsymbol{s}$ obtained from the proposed method

$p(x_s,y_s|\boldsymbol{D},I)$ obtained from conventional method with point assumption

Figure 3.24. The estimation results of the proposed method and conventional method in Case II.

3.5.5 Comparison to conventional method with point assumption

To prove the necessity of geometry estimation in the proposed method, we also conducted STE using the conventional method introduced in **Chapter 2**, in which the source is always assumed to be an ideal point without any geometry. All the settings for Bayesian inference and measurement data are the same as those used above.



x-coordinate

y-coordinate

strength of the source

Figure 3.25. Posterior PDF of source parameters estimated by the conventional method with ideal point assumption in Case II. (dotted line: true value)

The results are summarized in **Fig. 3.25**. Since the conventional method does not include the geometry estimation function, it can only provide the samplings for coordinate and strength. The coordinate estimation diverges from the true value of the middle point of the line source. Although the peak value of strength PDF coincides with the true value, it is notable that the sampling range spreads up to 25 times the true value while the proposed method constrains this range by a factor of less than 2. In actual applications, this wide range may confuse the risk management process. It is also possible that the complicated dispersion in real urban may produce a wider range or a wrong peak. Again, regarding the 50^{th} percentile and standard mean values of PDFs as the parameter estimators, the comparison between the true values, estimations of the proposed method, and estimations of conventional methods are summarized

in **Table 3.4**. The proposed method can not only provide valuable information about the geometry of the unknown source but also perform better than the conventional method in the aspects of locations and strength.

The joint posterior PDF $p(x_s, y_s|D, I)$ is shown in **Fig. 3.24 (b)**. Compared with the results of the proposed method, the entire distribution is limited to a small area. The main body of the line source was excluded in the probability distribution. The result cannot provide enough information about the location or geometry of the target source, suggesting that the STE with the conventional method fails in estimating a source that is not an ideal point.

Table 3.4. Summarized estimation results of the proposed method and conventional method

Method		x_s(m)	y_s(m)	length(m)	width(m)	angle	strength(1/s)
True value		90	82.5	140	6	0	1
Estimations (proposed method)	50th percentile	87	83	123	5	0	1.7
	standard mean	88	82	113	5.5	0	1.1
Estimations (point assumption)	50th percentile	103	89				5.7
	standard mean	100	88				7.5

3.6 Conclusion

In this Chapter, a method was proposed for the identification of a line source with the application of the super-Gaussian function. The main conclusions are as below.

The applicability of the method was first evaluated through Case I: a numerical experiment involving an ideal urban boundary layer. The numerical measurements were simulated using the Reynolds averaged Navier-Stokes model. According to the estimation results, all the parameters were correctly identified with almost none or acceptable errors under ideal conditions, without measurements and modeling errors. The proposed method successfully inferred that the source's shape is line-like, by automatically adjusting the coefficients in the super-Gaussian function, without any prior knowledge of the source geometry.

After that, the effects of different sensor configurations on the estimation results based on Case I were discussed to reveal the measurement requirements of a line source. Because the line source contained more geometric information (such as length, angle, and width) compared with the point source, it cannot be identified correctly by the conventional sensor network, in which all the sensors are placed regularly downstream. This chapter summarizes that, in order to collect enough information about the line source, at least six sensors are indispensable, which includes the sensors near the source and null-measurement sensors. In the real application, it was noted that the density of sensors should be sufficiently high to cover the "region of influence" and handle any possible unknown sources.

Next, to examine the robustness of the proposed method under the complicated errors caused by measurements and numerical modeling, a published practical case of a simplified urban square was utilized. In this case, the measurements were obtained from discrete sensors in wind tunnel experiments rather than through simulation. The estimation results showed that, in the practical situation, the proposed method could still estimate the line source parameters efficiently, with few and simple coefficients that were tractable for Bayesian inference. More importantly, the estimation was accomplished without any prior geometric information.

Furthermore, with the experimental measurements in Case II, the performance of the proposed and conventional methods with ideal points was compared. The results implied that the ideal point assumption makes the conventional method incapable of providing any information about the source geometry. This simplified assumption also resulted in significant errors while estimating location and strength. Consequentially, the ideal point assumption failed in estimating the unknown source with geometry, which is a common scenario in real life. The proposed method is an important improvement as it can estimate the geometry of the source and reduce the error caused by the point assumption.

Admittedly, the current model may still bring a tail error into the length and width estimation as the super-Gaussian function cannot be the same as a line no matter what the value of λ is, though it has demonstrated that the error is acceptable.

It should also be noted that this is an early trail to estimate an unknown source with geometry. Few related research, either the validation database or comparable algorithm, is available in the literature. Most of previous research still concentrates on the point source estimation. Therefore, in the following chapters, the focus will change from the geometry

estimation to point source estimation to utilize the existing database and benchmarks, and the super-Gaussian function will not be used.

1 Symbols

A : normalization factor of Gaussian function

a, b, c : coefficients of Gaussian function

C : the concentration distribution caused by a source

C^* : adjoint concentration distribution

$\boldsymbol{D}$: measurements vector

D_i : the measurement of the sensor with index i

D_t : turbulent diffusivity

$f_{\boldsymbol{x}}$: the Gaussian function for variables $\boldsymbol{x}$

$f(x, y)$: the bivariate Gaussian function

H : the building height in the real life of Case II

h : the height of the building model in the experiment of Case II

I : background information for Bayesian inference

I_u : the turbulence intensity in the streamwise direction

k : turbulent kinematic energy

k_s : roughness coefficient for the wall function applied in Case I

m : the dimension of the Gaussian function

$N[\mathrm{a}, \mathrm{b}]$: the uniform distribution bound between a and b

$p(A|B)$: conditional probability of event A occurring given that B is true

Q : the source strength in the experiment of Case II

q_s : release strength of the source

$\boldsymbol{R}$: modeled concentration vector

$\bar{U}$: the spatially averaged wind velocity of the inlet in Case I ($4\ \mathrm{m/s}$)

U_r : the reference velocity measured at the building height in the experiment of Case II

x_0 : x coordinate of the middle point of Gaussian function

x_m : x coordinate of the sensor

x_s : x coordinate of the middle point of line source

y_0 : y coordinate of the middle point of Gaussian function

y_m : y coordinate of the sensor

y_s : y coordinate of the middle point of line source

$\boldsymbol{\Sigma}$: the covariance matrix of variable $\boldsymbol{x}$ in the Gaussian function

$\boldsymbol{\mu}$: the mean value matrix of variable $\boldsymbol{x}$ in the Gaussian function

θ : inclined angle of the Gaussian function or the line source

$\sigma_{d,i}^2$: the variance of error in the measurement of the sensor with index i

$\sigma_{m,i}^2$: the variance of error in the modeling concentration for the sensor with index i

σ_X : the covariance of Gaussian function in the x direction

σ_Y : the covariance of Gaussian function in the y direction

λ : power coefficient of the super-Gaussian function

ε : turbulent energy dissipation

1

2

Chapter 4
Construction of urban turbulent flow database by large-eddy simulation and wavelet-based compression method

Abstract

According to **Chapter 2**, the prediction of the source-receptor relationship relies on the simulation of the adjoint equation, in which the hypothetical tracer emitted from each sensor is transported in the inverse spatiotemporal flow field. As shown in **Chapter 3**, an adjoint equation is a partial differential equation that is similar in form to a dispersion equation and can be simulated by computational fluid dynamics (CFD) model like Reynolds averaged Navier-Stokes (RANS). However, the prediction accuracy of RANS models for the time-averaged flow fields around buildings has been shown to be insufficient when compared with the large eddy simulation (LES) model (Tominaga et al., 2008a). Besides, the turbulence diffusion is approximated based on the mean field in the RANS model. These limitations undermine the accuracy of adjoint equation simulations and source term estimation (STE) with RANS. It is believed that LES may improve the prediction accuracy of adjoint equation simulation and enhance the reliability of statistical STE in urban applications.

Until now, the LES of the adjoint equation has been regarded as impractical because the time-series flow field data of the entire domain must be produced by forward simulation and stored in advance to realize the inverse simulation. One important challenge here is that the volume of data acquired by forward LES is too large for practical application. This research proposed to use the wavelet-based compression method to mitigate the storage pressure. The LES flow field can be compressed into a portable database for later usages like adjoint equation simulation, new dispersion simulation, and validation for new models.

Before the detailed introduction of LES for adjoint equation and STE in **Chapter 5**, this chapter constructed a compression turbulent flow database for a block-arrayed building group model simulated via LES. The objective has three aspects: (1) Evaluate the accuracy of compression and the applicability of compressed database; (2) Construct a compressed database for the adjoint equation simulation in **Chapter 5**; (3) Prepare a raw dispersion field for the STE in **Chapter 6**.

More details of the last two aspects can be found in their own chapters. This chapter focuses on the compression method and database construction. The compression performance was analyzed from two viewpoints: single snapshot and time-series data.

4.1 Introduction

In environmental flow simulations, pollutant dispersions in complicated urban areas are often simulated together with the flow fields by computational fluid dynamics (CFD). To simulate complex flow fields accurately, sophisticated numerical models, such as large eddy simulation (LES) and direct numerical simulation (DNS), have been developed, and refined grids are involved. With the help of supercomputers or clusters of workstations, most of the details needed in different scales can be simulated by proven techniques with sufficient calculation time. For example, Kikumoto and Ooka (2012) simulated air pollutant dispersion with bimolecular chemical reactions in a street canyon by LES. Coceal et al. (2007), Jacob and Sagaut (2018) explored a dynamic wind environment in a real and an idealized urban area with LES and DNS. These cases often require more than a hundred calculation hours, even if the simulations are conducted in parallel by dozens of cores. However, when new sources and questions arise, these massive simulations have to be repeated many times. A more economical way is to construct a flow field database with accurate simulation. It can be used to simulate new dispersion cases and validate new simulation methods. One example is the Johns Hopkins Turbulence Database (http://turbulence.pha.jhu.edu) for DNS simulations of turbulence. This database has been used, fully or partially, in several research projects, contributing to more than 40 publications (Kanov et al., 2015). For instance, Graham et al. (2016) used this database to validate a new wall model for LES.

However, another issue has been brought to the attention of researchers by this trend: the resultant data of a highly resolved flow simulation by LES or DNS has an enormous volume (gigabytes or more) per time step. The handling of such big data has challenged the limitations of the storage space and I/O speed of computers. To alleviate the pressure on storage and promote sharing, new calculation, and postprocessing of databases, an appropriate data compression method is necessary. In addition, considering that there are strong correlations in turbulent structures, temporally and spatially, the information of the flow could be represented by fewer data given that some data can be reconstructed by others with these correlations. Consequently, the compression of CFD data is theoretically realizable.

Generally, there are two kinds of compression methods: lossless and lossy. Lossless compression methods are developed for situations in which any difference between the decompressed and original data cannot be tolerated. One important application is text

compression, where even a different character from the original text may change the meaning completely. Compression of confidential data, such as financial records, also requires strict accuracy. In this case, because all the data need to be preserved rigorously, lossless compression methods, such as the Lempel–Ziv–Markov chain algorithm, can reduce the file size to less than 20% (Hadjidoukas and Wermelinger, 2019).

However, data loss is allowed in lossy compression, so the compression ratio is correspondingly higher. It happens that most CFD data have robustness, to some extent, against proper loss. It is a common practice to save the data with lower accuracy than full precision considering that numerical error makes saving with full precision meaningless. Moreover, as discussed above, some data can be deleted in the compression and reconstructed later by physical correlations. Hence, lossy compression methods with a high compression ratio seem feasible. Still, it is worth mentioning that, since the compression is based on the correlated turbulence structures, for small turbulence that is highly random with minimum correlation, the effective application of compression still needs confirmation. Despite that small-scale turbulence seems vulnerable to even small errors, it is reasonable to believe that the compressed data is applicable when this small turbulence is not the dominant part.

One popular direction in the field of fluid mechanics is mode decomposition, in which the nonlinear, complicated flow field is decomposed into different modes and can be approximated by the reconstruction of a limited number of dominant modes. Several methods, such as proper orthogonal decomposition (Berkooz et al., 1993) and dynamic mode decomposition (Schmid, 2010), have been invented and have achieved success in application. These methods can be naturally utilized as lossy compression methods where the original time-series data are represented by the dominant modes and corresponding mode-temporal coefficients, whose volume is much smaller. Meanwhile, a loss is caused by neglecting the nondominant modes. The main drawback of these methods is that the decomposition process always involves convoluted eigen-calculation of a huge matrix. If a database involves with long time series and large number of grids, the matrix will occupy a massive memory space that is difficult to process for a common computer. The requirements of calculation ability and the time history of the flow confine their general application as compression tools. However, its merit is the convenience of decompression. Once the modes and coefficients are successfully decomposed, the reconstruction only needs simple algebraic operation, which is fast.

In contrast, another decomposition method, wavelet decomposition (WD), can produce a set of bases for the fluid with a much lighter calculation burden. Furthermore, it can be applied to one single snapshot without time history. The only demand for achieving a high compression ratio is that the data must change smoothly in the space (Schmalzl, 2003), which is guaranteed by the diffusion term in the governing equations of the flow field. Therefore, wavelet-based compression is a probable choice for CFD data. Zubair et al. (1992) introduced this idea first to decompose and compress turbulent signals. Wilson (2002) applied wavelets to compress the turbulent simulation data, and the compression ratio reached 256 for the maximum. Schmalzl (2003) pointed out that the existing image wavelet compression method provides an available, easily adopted tool for CFD data. The efficiency of different compression schemes, such as JPEG and MPEG, were evaluated. Sakai et al. (2013) combined the WD with quantization and entropy encoding to construct the basic structure of the wavelet-based compression method (WCM) and increased the compression ratio further. They verified the WCM by simulation data in a structured Cartesian mesh, which is a common situation in CFD. Following that method, Kolomenskiy et al. (2018) proposed an empirical equation for error control, which makes the WCM more controllable and user-friendly. They also implemented the method in different simulation cases with different scales and resolutions. The wavelet-based method has shown a high compression ratio and accurate decompression in general. Therefore, it is a promising solution to construct a small-sized CFD database by the WCM.

However, to the authors' knowledge, the WCM has not been applied to construct a CFD database including both the spatial distribution and time-dependent dynamics. In previous research, the verification of the WCM was limited to the compression of a single snapshot of the flow field. The major concern was the effects of compression of a snapshot on the postprocessing or restarting simulation from that snapshot, but the cumulative effects resulting from compression of time series data have not been reported. It is not clear whether the dispersion simulation can proceed successfully with such a compressed database. Moreover, studies from the perspective of the limitation of the compression ratio and source of errors have been few. There is little research in which the feasibility of compression in complicated flow field simulation, such as building clusters in urban areas, has been addressed. These aspects should be investigated further before the construction of the CFD database with WCM and applying it for inverse simulation later.

In this chapter, a small urban flow database was constructed by using the WCM, and the

applicability was verified. The original data for the database were obtained from an LES simulation of a block-array urban area. The LES data were compressed with different specified errors by the WCM to construct the database. The relationship between compression error and the compression ratio, as well as the limitation of the compression, are demonstrated. Furthermore, the effects of compression on a single snapshot and time-series data are discussed. To confirm the usability of the compressed database, it was used to re-simulate the dispersion of a passive scalar, which is compared with the dispersion field of the original LES and wind tunnel experiment.

4.2 Compression methodology

First of all, the compression method is introduced briefly. The basic idea of using WD to realize data compression for the CFD simulation was proposed by (Zubair et al., 1992) for turbulent signals. Subsequent research (Sakai et al., 2013; Schmalzl, 2003; Wilson, 2002) enriched the idea by adjusting the algorithm for general cases and combining it with other data compression concepts. Kolomenskiy et al. (2018) proposed an empirical formula for error control of the WCM and verified its accuracy in several possible application scenarios.

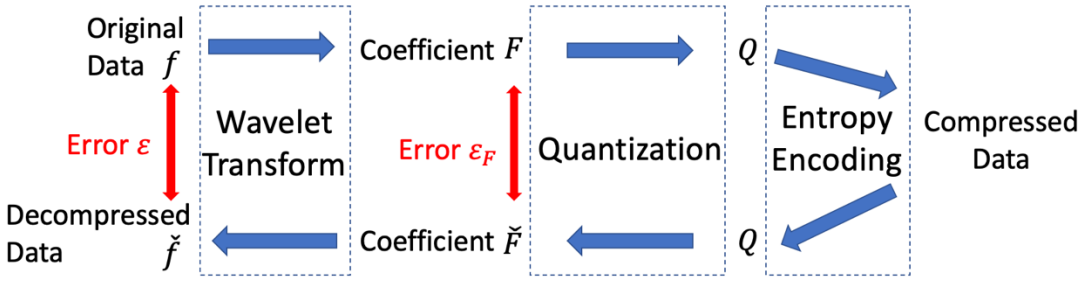


Figure 4.1. The structure of the compression method

Here, the method proposed by Kolomenskiy et al. (2018) was applied exactly because the structure of this method is relatively complete and, more importantly, the error control is explicit. There are three steps in this compression method: WD, quantization, and entropy encoding (**Fig. 4.1**). Decomposition can be simply realized through the inverse of the process. In the following, these three steps are introduced. The original data are assumed to constitute a three-dimensional scalar field obtained from the Cartesian indexing mesh grid $\{f_{i,j,k}\}$. The

index i, j, k indicates the coordinate of a grid in the x, y, z directions.

4.2.1 Wavelet decomposition

WD is popular in the engineering field and has shown its power in signal processing (Rioul and Vetterli, 1991), noise control (Qiu et al., 2016), and image compression (Usevitch, 2001). The essence of the WD is to decompose the original data into the approximation part with low frequency and the detailed part with high frequency. On the one hand, for the flow field, most of the energy is contained in the low-frequency structures, which means the decomposition coefficients are large for the approximation part and small for the detailed part. On the other hand, the approximation part only occupies part of the original saving space. Because the WD can be implemented on the decomposed one repeatedly, the resulting approximation part uses quite a small space to represent most of the information.

A two-dimensional WD was used as an example. **Fig. 4.2a** is an averaged streamwise velocity field around an isolated building on a horizontal plane. This flow field was decomposed once to yield **Fig. 4.2b** using the Cohen–Daubechies–Feauveau 9/7 (CDF9/7) (Daubechies and Sweldens, 1998) wavelet. It indicates that the decomposed flow field consists of four parts, which correspond to the low-low frequency part (1), high-low frequency part (2), low-high frequency part (3), and high-high frequency part (4), respectively, transformed by the wavelet in the horizontal and vertical directions. In the decomposition, wavelets work as a filter to convert the flow field to the wavelet coefficients. Large filters can extract large structures and small filters can extract small structures. The filtering process is conducted in each axis separately. In this 2-dimensional example, the horizontal decomposition puts the coefficients obtained from the large filters to the left and coefficients of small filters to the right. Similarly, the vertical decomposition puts the coefficients of large filters to the up and coefficients of small filters below. Therefore, the coefficients at the top-left are the ones corresponding to the large filters in both horizontal and vertical directions, representing the large-scale structures with low frequencies in both directions, and labeled with “low-low”. The other parts are labeled in the same way.

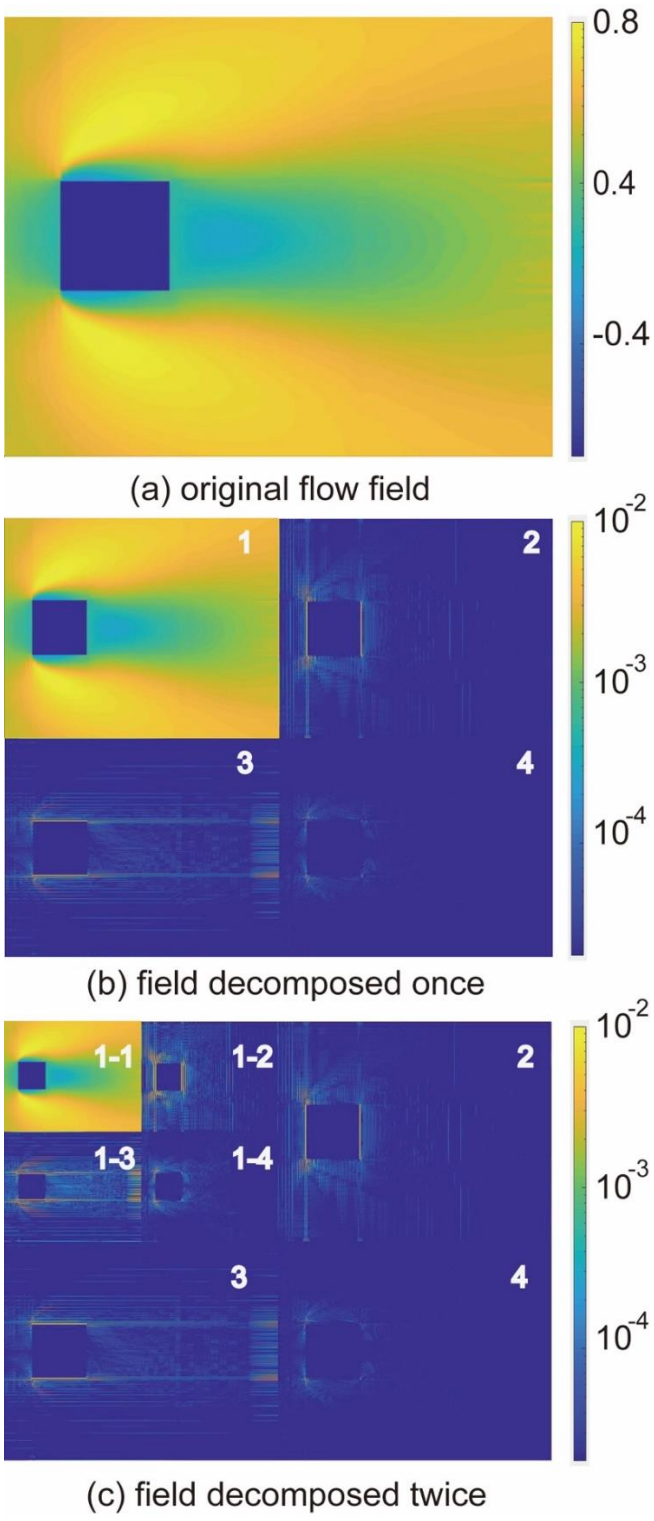


Figure 4.2. A 2-dimensional example of wavelet decomposition of flow field: (a) is the original mean streamwise velocity field around a building; (b) is the magnitude of wavelet coefficients obtained from one-time decomposition of (a); (c) is the magnitude of wavelet coefficients obtained from two-times decomposition of (a).
(parts 1 and 1-1 are the approximation parts and use the same color bar in (a), and others are detailed parts using the logarithm color bar to show the coefficients clearly)

Among these four parts, only the coefficients in part (1) have the same order of magnitude as the original field, which means this is the approximation part with the most energy. In contrast with part (1), the order of coefficients in the other three detailed parts is extremely small. These coefficients can be approximated to reduce the volume of the data considerably if the minor energy is not in the scope of interest. Another advantage of WD is that it can be repeated several times on the decomposed field. **Fig. 4.2c** shows the further decomposed field from **Fig. 4.2b**. In the second transform, parts (2)–(4) are kept unchanged while part (1) is further decomposed into four smaller parts. Similar to the first transform, these four parts represent different frequencies. Part (1-1) is the approximation part with major energy, while parts (1-2), (1-3), and (1-4) are the detailed parts with a small order of magnitude. After WD is done several times, the approximation part containing the dominant information occupies a limited storage volume in the upper left corner, while the rest of the detailed parts can be approximated further.

In this research, a three-dimensional CDF9/7 WD was performed four times on the original data. It was confirmed that CDF9/7 can reach a higher compression ratio than other low-order wavelets, such as Haar, Daubechies, and Symlets (Kolomenskiy et al., 2018). The basic wavelet like Haar is the low-order orthogonal type, i.e. the basis functions of wavelet are orthogonal to each other and have zero inner product. Although it is simple to use, it may produce too many wavelet coefficients, which are unfavorable to the compression. Therefore, the image processing community suggests the use of biorthogonal wavelets, which can transform the image into fewer coefficients because there is more flexibility to design and optimize the basis functions (Li, 2015). Among the biorthogonal wavelets, CDF 9/7 is confirmed to be the best one for image processing (Villasenor et al., 1995) and employed as the default wavelet in the JEPG2000 (Skodras et al., 2001). However, it has also been pointed out that the input coefficients of CDF 9/7 need more hardware resources and processing time because they are irrational and need high precision representation in the computer (Martina and Masera, 2005; Naik and Holambe, 2014). Several methods have been proposed by them to improve this point and their effects on turbulence compression still need further tests. Meanwhile, according to the author’s knowledge, research about the relationship between turbulence and wavelet compression is very limited. Systematic research concerning the best wavelets for turbulence decomposition is still in need. Therefore, in the current research, the most popular wavelet CDF

9/7 is utilized. The wavelet coefficients $\{F_{i,j,k}\}$ are processed further in the next two steps.

4.2.2 Quantization

Considering that most of the wavelet coefficients $\{F_{i,j,k}\}$ have a small order of magnitude, one brute-force compression algorithm changes all the detailed parts with 0 and only keeps the approximation part. Although this algorithm can ensure a high compression ratio, the resultant accuracy may be too poor because there are still relatively large coefficients in the detailed parts that are removed directly. Moreover, there is no selection leeway for users to control the error margin of the compression. Only one choice remains because all the details are abandoned. A better algorithm is quantization, in which the user can specify how much of the detail should be removed to satisfy the application requirements and error control.

In the computer system, eight 1-byte memory spaces are commonly used to represent the double-precision floating number $\{F_{i,j,k}\}$ in a lossless manner, in which one bit represents the sign, 11 bits represent exponent and 52 bits represent the fraction. To achieve a high compression ratio, one must quantize $\{F_{i,j,k}\}$ with fewer 1-byte memory spaces and accept the inevitable truncation errors ε_F. In the quantization system, the bits do not function in the conventional way but only create positions to quantize $\{F_{i,j,k}\}$. The algorithm will scan all $\{F_{i,j,k}\}$ to pick up the maximum and minimum values, and assign the minimum value to the first bit array and maximum value to the last bit array. The positions in the middle represent the values with an even difference between the maximum and minimum. In this case, if one allocates N 1-byte spaces for the quantization, there are $2^{8^N} = 256^N$ positions, and the data $\{F_{i,j,k}\}$ should be assigned to the corresponding positions. The true value is replaced by the rounded one nearest it, and the differences lying in the gap of the truncation error $\varepsilon_F = \frac{\max(F)-\min(F)}{256^N}$ are removed. In this case, the algorithm can control the width of the gap ε_F to decide how much of the detailed part in the wavelet coefficients is contaminated. Furthermore, the value with a large order of magnitude is added to the truncation error instead of being dumped directly. In the algorithm, the number of 1-byte memory spaces N was calculated by the truncation error ε_F.

$$N = log_{256}(\frac{\max(F) - \min(F)}{\varepsilon_F}) \tag{4.1}$$

Meanwhile, it is obvious that

$$\varepsilon_F \approx \max\left|F - \check{F}\right| \tag{4.2}$$

where $\check{F}$ is the quantized wavelet coefficients.

However, this truncation error is not directly correlated with the original data $\{f_{i,j,k}\}$ but rather with the wavelet coefficient $\{F_{i,j,k}\}$. It is necessary to figure out the effects of the truncation error on the original data to control the compression process effectively. Kolomenskiy et al. (2018) proposed that ε_F and the user-specified error ε for the original data have following relationship:

$$\varepsilon_F = \frac{\varepsilon \times \max|f|}{\eta} = \frac{\max\left|f - \check{f}\right|}{\eta} \tag{4.3}$$

Here, η is an empirical coefficient. Hence, in the compression process, the user can specify ε according to the requirement. The algorithm calculates the truncation error ε_F and necessary memory spaces N through Eqs. (4.1) and (4.3) for the quantization, which can ensure that the real compression error metric $\varepsilon_r = \frac{\max|f-\check{f}|}{\max|f|}$ is almost the same as ε. Moreover, η is proposed to be approximately 1.75 to realize the relationship according to several numerical cases. This value was examined also in this study.

According to the definition of ε_r and $\varepsilon \approx \varepsilon_r$, if ε is larger than 1, the largest compression error would be even larger than the largest flow field data, which makes the compression data too inaccurate. Meanwhile, if ε is too small, the compression ratio is so small that the compression is meaningless. Therefore, ε should be appropriately specified between 0 and 1. The attained coefficients $\check{F}$ are input in the entropy encoding.

4.2.3 Entropy encoding

Entropy encoding is a lossless compression technique to reduce the storage volume by representing subsequences of data with different symbols according to the times of their

appearance (Song, 2008). Although quantization does compress the data significantly, there is still redundancy in the data because the output of quantization does not occur with equal probability. This redundancy can be optimally removed by entropy encoding (Ruttimann and Pipberger, 1979). Specifically, in the algorithm, the appearance times of each value in the input data arrays are counted first. Then, values with large appearance times are represented by short symbols, while ones with small times are represented by long symbols, which reduces the volume of the input data further. In this research, the entropy encoding method called “range coding” developed by Martin (1979) was employed. The quantized wavelet coefficients are represented by different symbols and saved eventually.

4.3 Case description

In this part, the WCM is used to construct a compression database based on LES raw data. The details of the study case are as below.

4.3.1 Production of raw data with large-eddy simulation

The time series raw data for the database was first produced by CFD. To evaluate the compression ability of the WCM, the simulation case should have complicated flow fields and turbulence structures. Meanwhile, as a database for later source identification, the case should represent the common situation for dispersion in the urban area. In recent years, as a theoretically representative form of a real urban area, the block-arrayed idealized urban model has been widely used to investigate the wind field (Abd Razak et al., 2013; Ikegaya et al., 2017), pollutant dispersion (Branford et al., 2011b; Coceal et al., 2014b), and thermal effects (Uehara et al., 2000). Hence, the simulation case was set as regular building arrays.

LES was applied to simulate the turbulence flow because it can resolve most of the details of the flow. Its reliability and accuracy of simulating both the time-dependent dynamics and spatial distribution in wind engineering have been confirmed by multiple research (Blocken, 2018). The flow was simulated with a finite volume method implemented in OpenFOAM v1906. The standard Smagorinsky model was adjusted with the Smagorinsky constant $C_s = 0.12$, which has been confirmed in previous work (Tominaga et al., 2008a) for the flow past buildings.

The geometry of the calculation domain is shown in **Fig. 4.3**. The settings are the same as the wind tunnel experiment conducted at Tokyo Polytechnic University. The experimental data are open in the AIJ database (https://www.aij.or.jp/jpn/publish/cfdguide/index_e.htm). All blocks are the same cubes with edges $H = 60$ mm. They were arranged uniformly with a distance H between each other. The domain size was set as $16H\ (x) \times 14H(y) \times 4H(z)$. For the simulation of block-arrayed urban flow, Coceal et al. (2006) has explored the influence of domain height on the simulation with $4H$, $6H$ and $8H$. It was found that although $4H$ is too small to capture the largest turbulence structures, like the long streaky one, the differences in mean and turbulence statistics were negligible except in the vicinity of the top. Considering the calculation burden and the fact that flow around blocks is the main target, several research (Abd Razak et al., 2013; Claus et al., 2012; Ikegaya et al., 2017; Xie and Castro, 2006) applied $4H$ as domain height. Therefore, we set the same height following these research. Uniform structured meshes with a medium size ($H/20$) were applied. In previous research (Ikegaya et al., 2019; Xie and Castro, 2006), it was reported that the medium-sized mesh is sufficiently reliable to reproduce the mean flow distribution, turbulent kinematic energy, and large-scale flow dynamics. The total number of mesh cells was approximately 6.9 million. The bottom side of the domain and the surfaces of all blocks were defined as a nonslip boundary with a Spalding wall function. At the top of the domain, a free slip condition was imposed for velocity, and the zero-gradient Neumann condition was imposed for pressure. The four sides of the domain were all periodic boundaries. The turbulent inflow was generated by coupled inlet and outlet and was driven by a pressure gradient, which was adjusted at each step to ensure a fixed flux through the outlet.

The free stream speed u_r was measured at $(x, y, z) = (0,\ 7H,\ 3.33H)$ in both simulation and the experiment. The inflow moves along the positive x axis. The Reynolds number based on u_r and H is 1.68×10^4. The friction speed u_* was calculated by $0.07u_r$, which was proposed in previous work (Cheng and Castro, 2002). Approximately a $200T$ $(T = H/u_*)$ calculation time was spent for initialization to ensure that the simulation converges to a statistically steady state. Then, the time series data of $240T$ at all grid points were compressed and saved to construct the database. Each time step was 0.00075 s (approximately $T/240$).

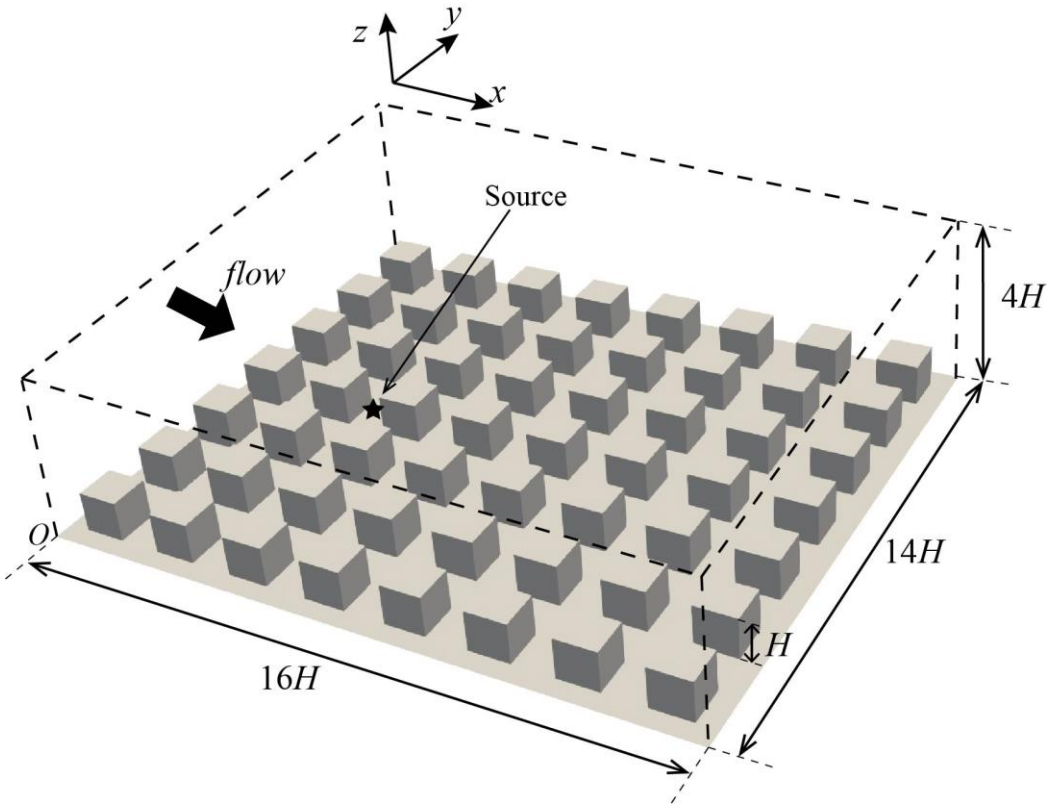


Figure 4.3. Schematic of simulation domain and the block-arrayed model.

4.3.2 Method to verify the compression database

Commonly, databases should satisfy usages in two time scales. The single snapshot is used to visualize the instantaneous flow fields and turbulent structures. The time-series data can provide statistical information and be applied later for inverse simulation or dispersion simulation of new sources. The compression should not affect these applications to protect the quality of databases.

Correspondingly, in this research, the instant effects on the single snapshot and cumulative effects on the time-series data of the compression process are confirmed. The spatial distribution of the flow field in one snapshot and statistics at several points in the compression database are analyzed in detail. To confirm that whether the compression database can be used for the dispersion simulation, the dispersion of the passive scalar emitted by the same source was simulated with both compressed databases and raw LES data. The original concentration distribution field of passive scalar was simulated online together with the raw LES while the concentration field of the same source transported by the compressed flow field was simulated

offline after the compression databases were constructed. The source was arranged at (x, y, z) = ($4H$, $7H$, 0) with the same injection speed in the wind tunnel experiment. The simulated concentrations in different cases are compared.

The governing equation for the transportation of the passive scalar in the LES can be expressed as

$$\frac{\partial C}{\partial t}+\frac{\partial \boldsymbol{U}C}{\partial \boldsymbol{x}}=S+\frac{\partial}{\partial \boldsymbol{x}}\left(D_e\frac{\partial C}{\partial \boldsymbol{x}}\right) \tag{4.4}$$

where C is the concentration in the grid scale, $\boldsymbol{U}$ is the velocity in the grid scale, S is the source term, and D_e is the effective diffusion coefficient including the molecular diffusivity D_m and the sub-grid scale (SGS) turbulent diffusivity D_{sgs}. Here, D_m is a parameter dependent on temperature and does not change in this simulation. Its value was given in **Chapter 2**. D_{sgs} is modeled as $D_{sgs}=\nu_{sgs}/Sc_{sgs}$. Here, the SGS turbulent Schmidt number Sc_{sgs} is assigned as 0.7, and ν_{sgs} is the eddy viscosity coefficient at the SGS. During the simulation, $\boldsymbol{U}$ and D_{sgs} are time-dependent. The database compressed these two fields at each time step with different error control ε: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, and 10^{0}. Then, the dispersion fields were repeatedly simulated with the decompressed $\breve{\boldsymbol{U}}$ and $\widetilde{D_{sgs}}$ data. The calculations restarted from the same time step in which the simulation reached statistical steadiness. A snapshot of each component of velocity field or effective diffusion coefficient field is about 100MB. The original data size of $240T$ sampling time is about 24 TB.

In this research, the implementation of WCM can be divided into the following procedures:

- LES: the raw data of $\boldsymbol{U}$ and D_{sgs} were produced at each timestep using OpenFOAM, and the original concentration field of passive scalar was calculated simultaneously.
- Compression: At each timestep, the raw data of $\boldsymbol{U}$ and D_{sgs} were inputted into WCM to be compressed and stored to construct the database.
- Decompression: The database was decompressed by WCM into $\breve{\boldsymbol{U}}$ and $\widetilde{D_{sgs}}$ data for each timestep.
- Dispersion analysis: OpenFOAM read $\breve{\boldsymbol{U}}$ and $\widetilde{D_{sgs}}$ to calculate the concentration field of passive scalar for each case.
- Post-processing and Comparison.

The compression source code is exactly the same one which is opened in Kolomenskiy et al. (2018). A script was made by the author to enable it to process the output of OpenFOAM in each processor in parallel automatically.

4.4 Results and discussion

The quality of the compressed database was examined by analyzing the effects of compression on the spatial distribution of the flow field in one snapshot and time-series statistics at several points. The compression ability of WCM was also examined in terms of the compression ratio and compression limitation. All the velocity results were nondimensionalized by u_r, and the concentration results were nondimensionalized by the standard concentration $C_r = C_{\mathrm{gas}} q/(u_r h_r^2)$. Here, C_{gas} is the emission strength of the source in the simulation and concentration of the source in the experiment, q is the gas flow rate, and h_r is the reference height $3.33H$.

4.4.1 Validation of large-eddy simulation

First of all, the raw data of LES results are validated in **Fig. 4.4 & 4.5**. Because the flow field is regular in the spanwise direction, we validated two rows in the streamwise direction: the profiles in the wake region of buildings (y=7H) and profiles in the open street region (y=6H). Samplings for $240T$ are used for the validation. It can be confirmed that the simulated flow field and concentration agree well with the experimental measurements. The mean, standard deviation of streamwise velocity, and the turbulent kinetic energy agreed well in most places.

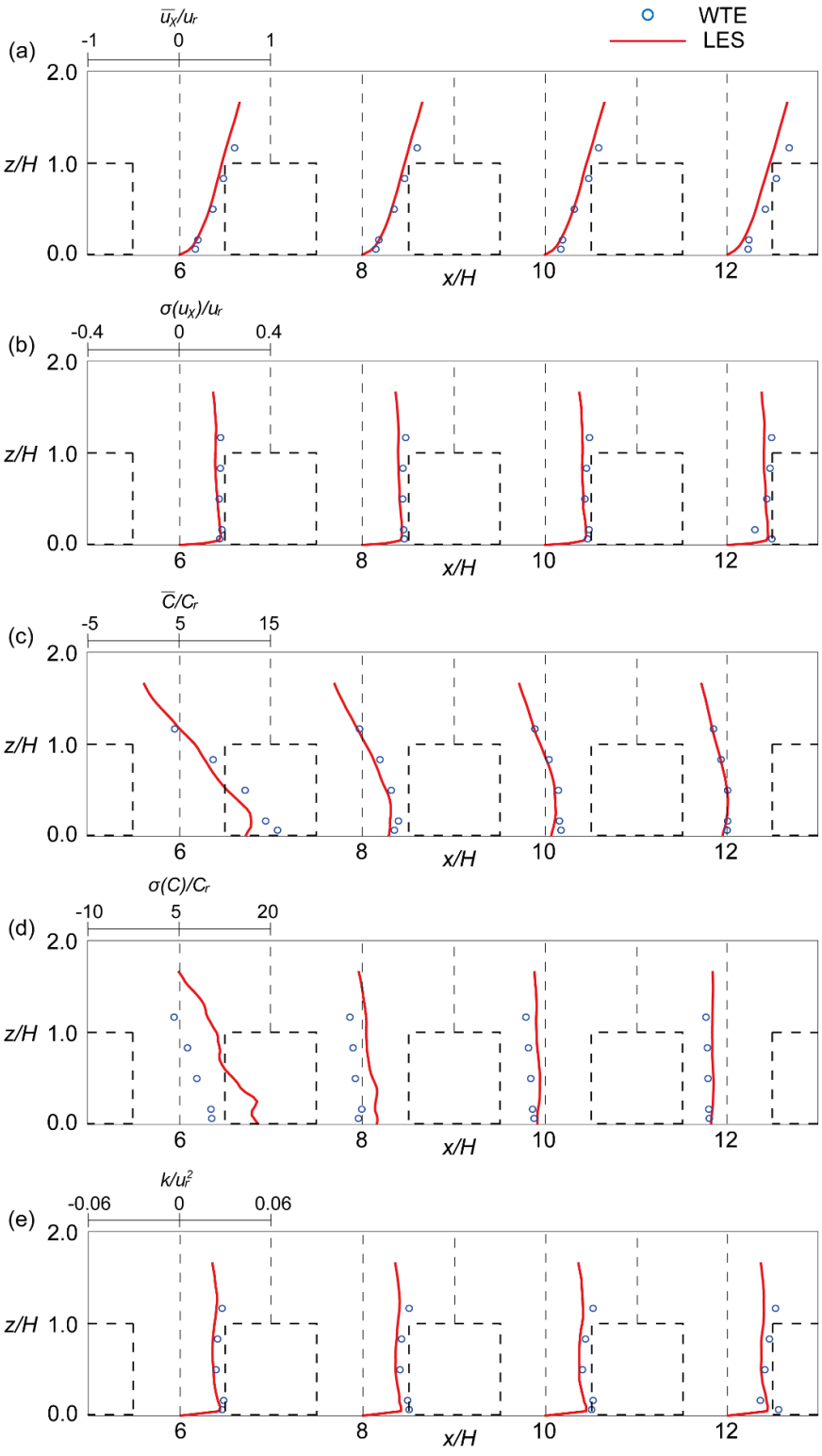


Figure 4.4. Validation of the velocity and concentration fields predicted by LES based on the experimental measurements in the open street region ($y = 6H$). (a) Time-averaged streamwise velocity; (b) standard deviation of streamwise velocity; (c) time-averaged concentration; (d) standard deviation of concentration; and (e) turbulent kinetic energy.

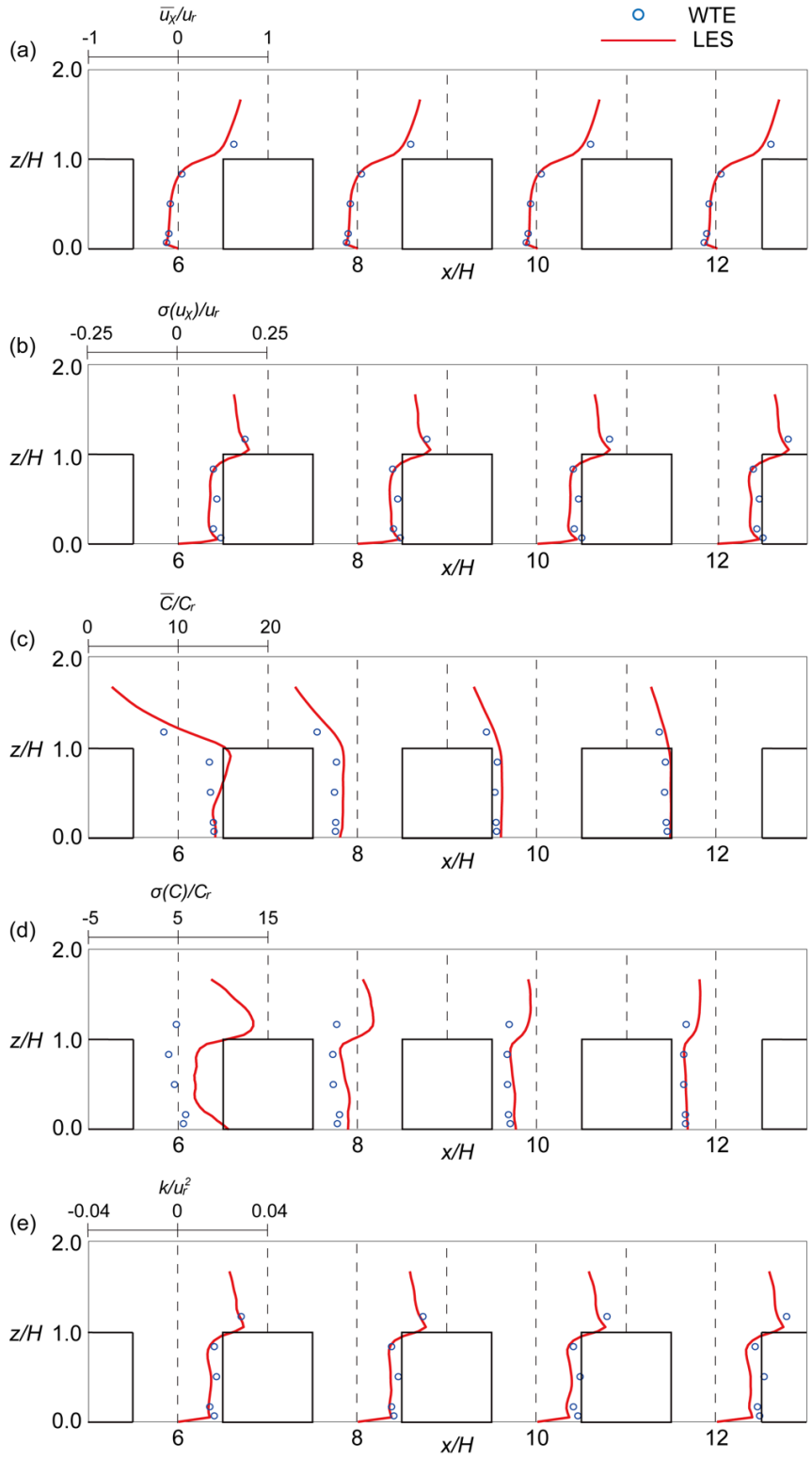


Figure 4.5. Validation of the velocity and concentration fields predicted by LES based on the experimental measurements in the wake region ($y = 7H$). (a) Time-averaged streamwise velocity; (b) standard deviation of streamwise velocity; (c) time-averaged concentration; (d) standard deviation of concentration; and (e) turbulent kinetic energy.

4.4.2 Compression ability

The compression ability of the WCM was explored. All the following discussions in this section are based on the compression result of the effective diffusion coefficient D_e in a single snapshot. The velocity field compression has similar results and is discussed later.

First, it is necessary to determine whether $\eta = 1.75$ in Eq. (4.3), proposed by Kolomenskiy et al. (2018), is an appropriate empirical coefficient for the error control. It is indicated in **Fig. 4.6** that the prescribed error control ε and resultant real error $\varepsilon_r = \frac{\max|f-\check{f}|}{\max|f|}$ are almost the same in the compression method using $\eta = 1.75$. Only a small discrepancy appears after ε is larger than 10^{-1}. The reason why the error control fails afterward is the limitation of the compression algorithm, which is addressed specifically later in this section. As a consequence, the error control with the empirical coefficient is accurate for complicated urban flow.

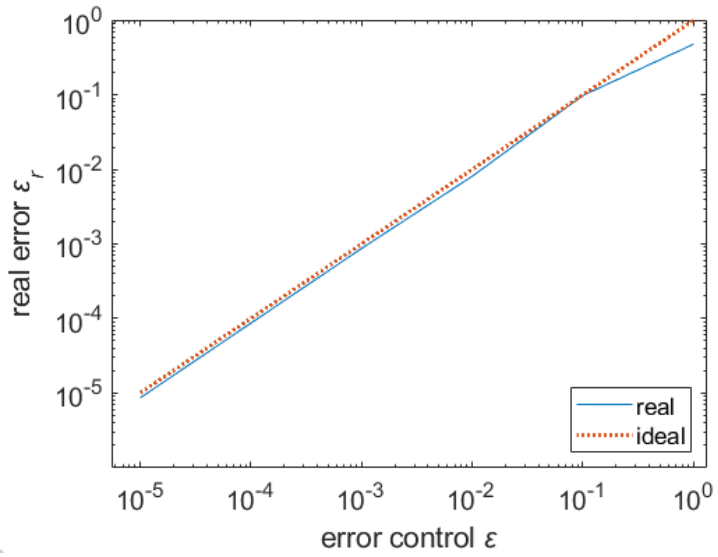


Figure 4.6. Relationship between the error control ε and real error ε_r: the red dotted line is the ideal one where $\varepsilon = \varepsilon_r$, and the blue line is the real situation (compression of effective diffusion coefficients)

Fig. 4.7 shows the relationship between the compression ratio $r = V_o/V_c$ (where V_o is the storage volume of the original data, and V_c is the storage volume of the compressed data) and the error control ε. When ε is set small, r increases with ε exponentially. After ε

reaches approximately $10^{-0.8}$, the increasing tendency starts to change. Large oscillations of the compression ratio appear in the red area, which indicates the compression limitation of the WCM. It is shown that r reaches its largest value when ε is larger than 1, where the largest compression error is almost the same as the largest value in the original field. The dotted line denotes the compression ratio realized by a conventional compression method, in which the flow data was represented by single-precision numbers and then compressed with the standard ZIP algorithm. It can be confirmed that the current method outperforms the conventional one in the compression ratio and provides error control to users.

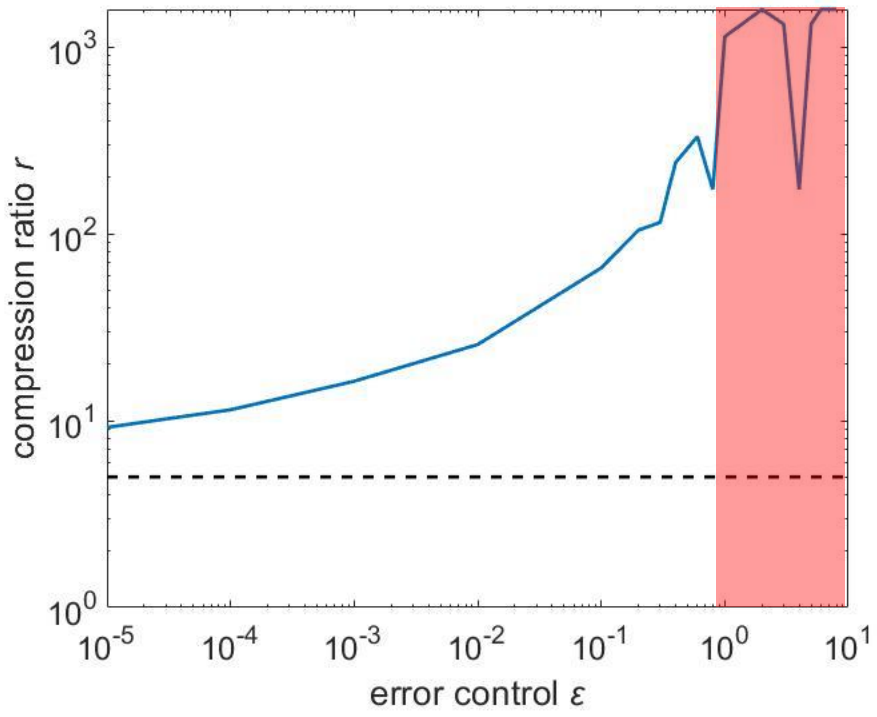


Figure 4.7. Relationship between the error control ε and the compression ratio r (compression of effective diffusion coefficients). The dotted line marked the compression ratio realized by a conventional method: single-precision representation + standard ZIP.

The reason for the appearance of the limitation was studied here. Because the compression method in this research retains all the wavelet coefficients for the quantization and the entropy encoding causes no error, the main error results from the truncation in the quantization process. **Fig. 4.8** shows the probability density function of the wavelet coefficients $\{F_{i,j,k}\}$ of the D_e field in one snapshot compression. Although the coefficients ranged to more than 3×10^{-3}, most of the information lies in the area between -1×10^{-4} and 1×10^{-4}. Because the

coefficient difference that is smaller than the truncation error ε_F is discarded in the quantization process, if ε_F is larger than 2×10^{-4} in this case, the main information is lost, and the whole structure of the wavelet coefficient is destroyed. In Eq. (4.3), if ε_F equals 2×10^{-4} and the true value of $max|f|$ is substituted, the limitation of the user-specified error ε can be calculated as

$$\varepsilon \approx \frac{0.0002 \times 1.75}{0.00029} = 1.2 \tag{4.5}$$

This value is close to the one where the compression ratio confronted the limitation. Consequently, the main reason for the compression limitation is that the truncation error is too large so that most wavelet coefficients are removed by the quantization. Although the error control ε that reaches the limitation of the compression method can change for each case, it is approximately 1 in the condition of this study. There is no universal value suitable for all cases, but the limitation can be calculated by the range of the wavelet coefficient to ensure that the core information is not damaged.

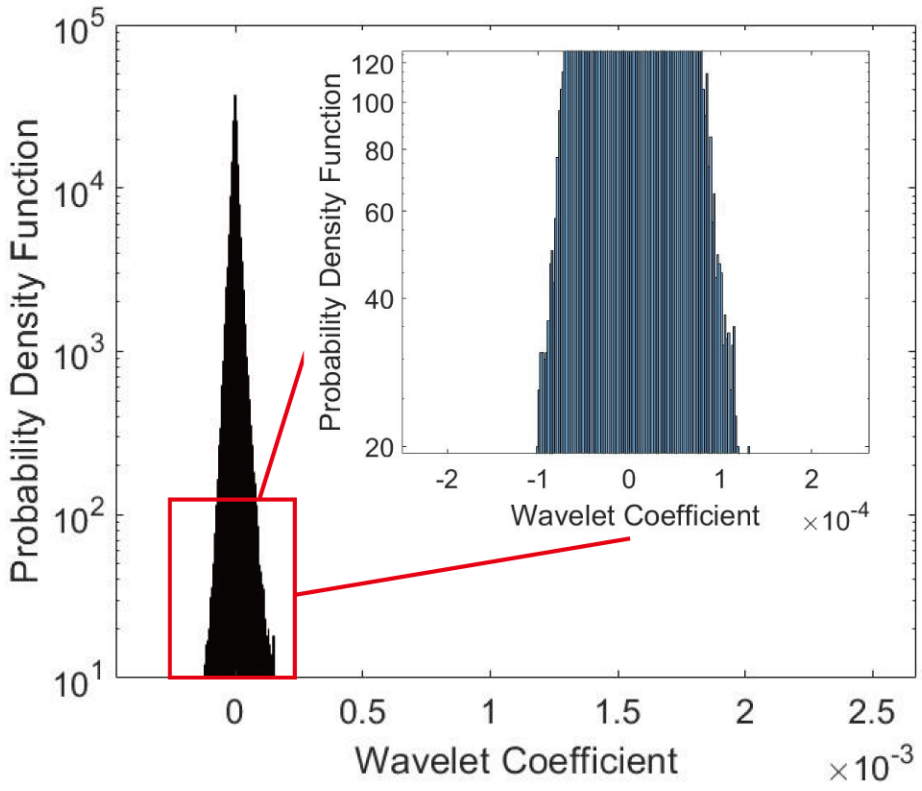
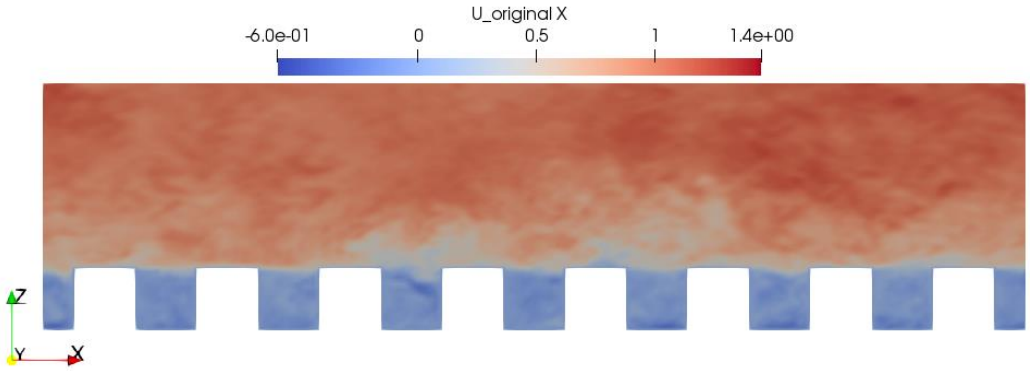


Figure 4.8. Distribution of wavelet coefficients of effective diffusion coefficients in one snapshot

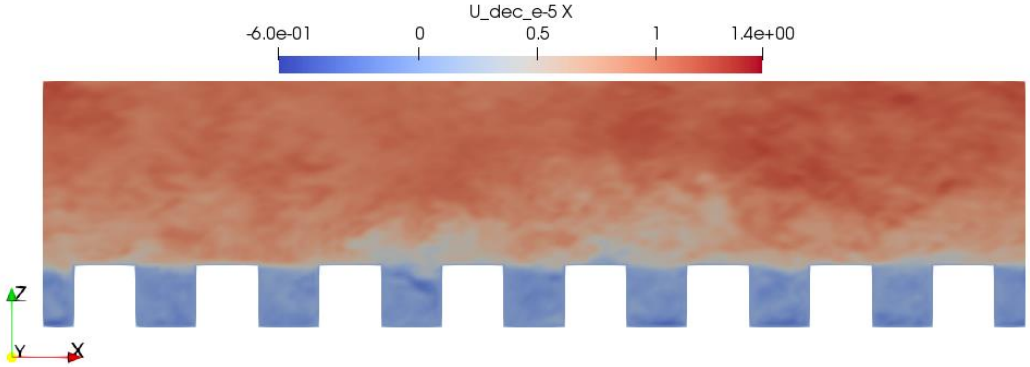
4.4.3 Compression effects on a single snapshot

In this section, the focus is on the compression effects on a single snapshot, which is important for the postprocessing and flow visualization with the database. Because the WCM compressed each timestep independently, a snapshot was arbitrarily selected and analyzed.

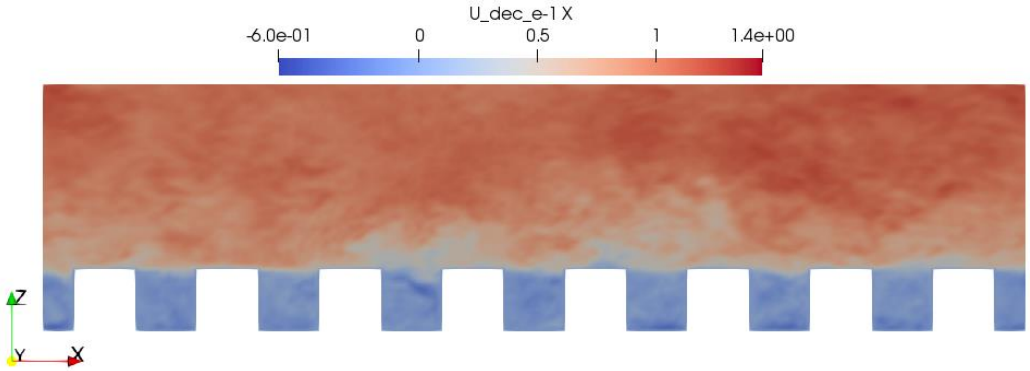
Fig. 4.9 compares the instantaneous streamwise velocity field U_x in the vertical plane in the middle of the y direction between the original LES and compressed databases. The difference between these fields can be checked in **Fig. 4.10**. When the error control ε is smaller than 10^{-1}, the compression effects are so small that the visualization cannot identify the difference. When ε reaches 1 (10^0), the situation changes so that discrete large values appear and mottle the velocity field.



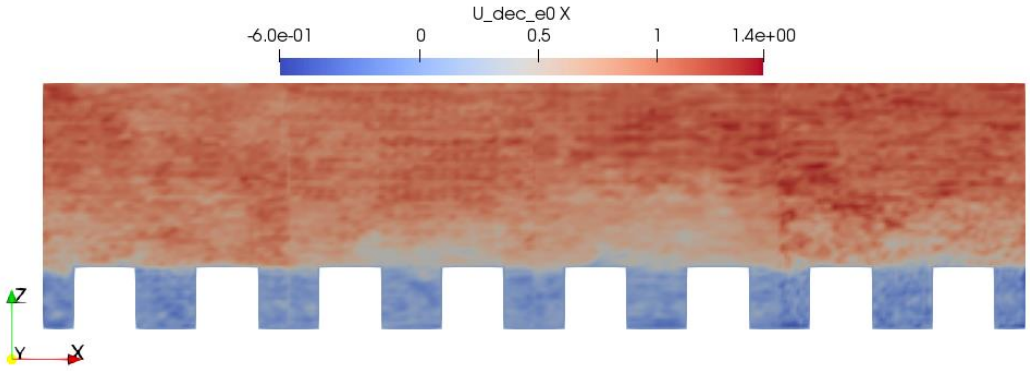
(a) Original LES



(b) Compression database with $\varepsilon = 10^{-5}$



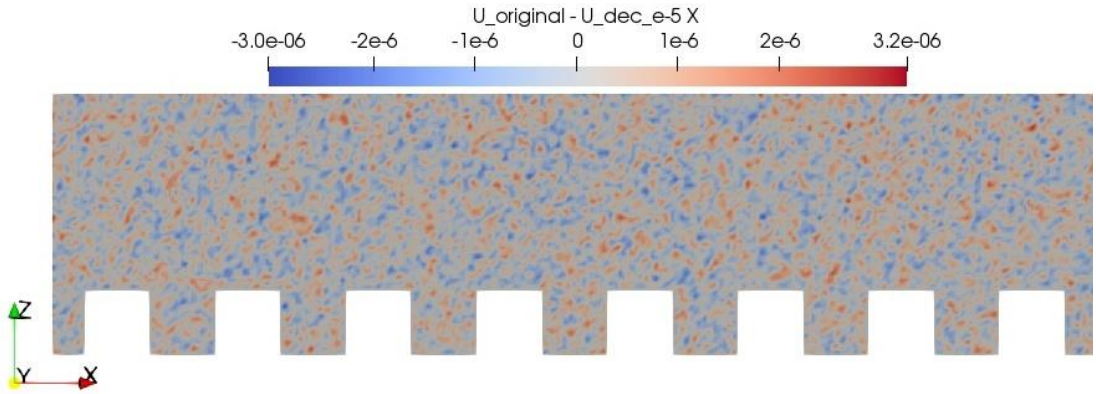
(c) Compression database with $\varepsilon = 10^{-1}$



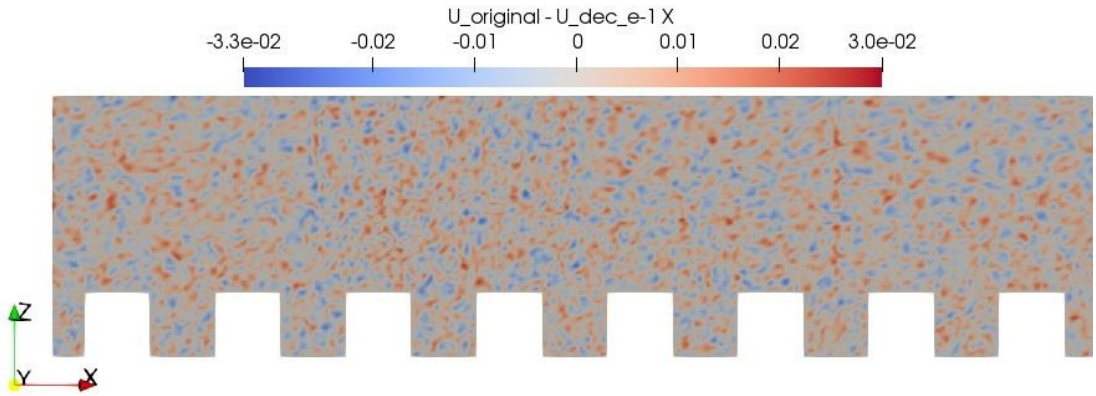
(d) Compression database with $\varepsilon = 10^{0}$

Figure 4.9. Streamwise velocity field in a vertical plane ($y = 7H$)

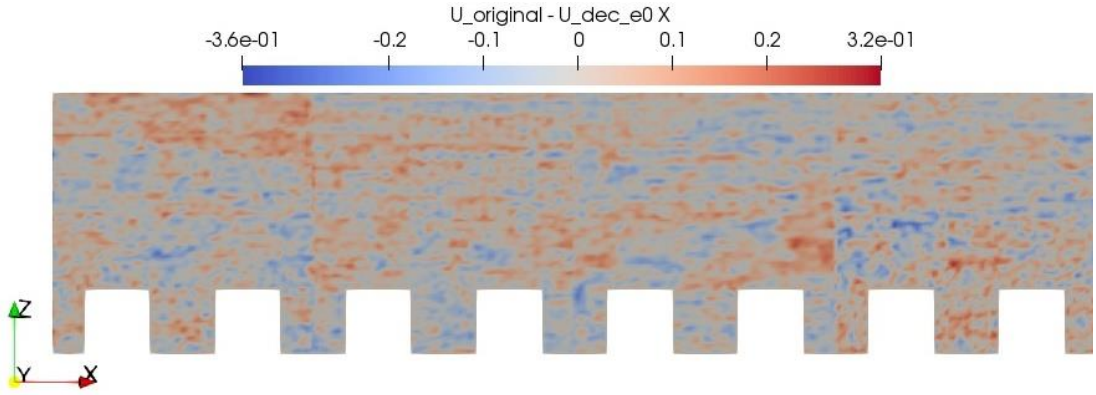
1



(a) Difference between original LES and compression database with $\varepsilon = 10^{-5}$



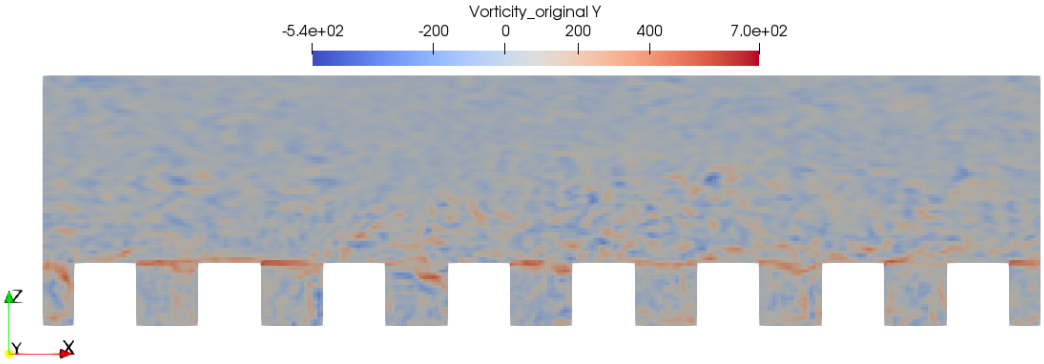
(b) Difference between original LES and compression database with $\varepsilon = 10^{-1}$



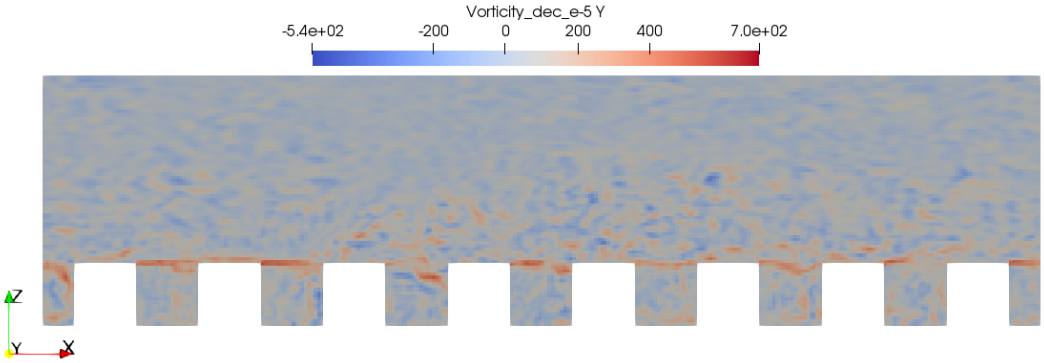
(c) Difference between original LES and compression database with $\varepsilon = 10^{0}$

Figure 4.10. Streamwise velocity difference field between original LES and compression databases in a vertical plane ($y = 7H$)

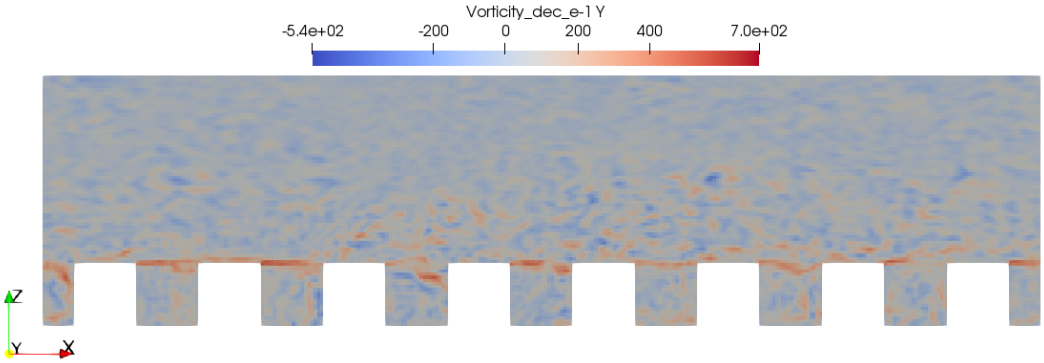
1



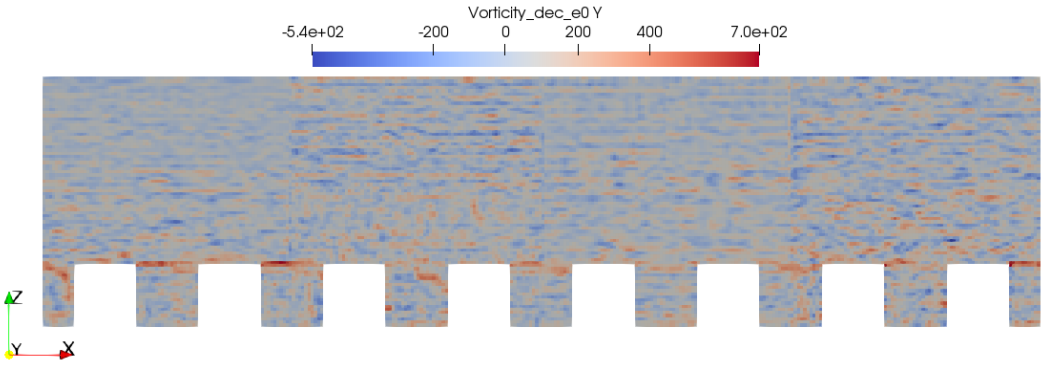
(a) Original LES



(b) Compression database with $\varepsilon = 10^{-5}$



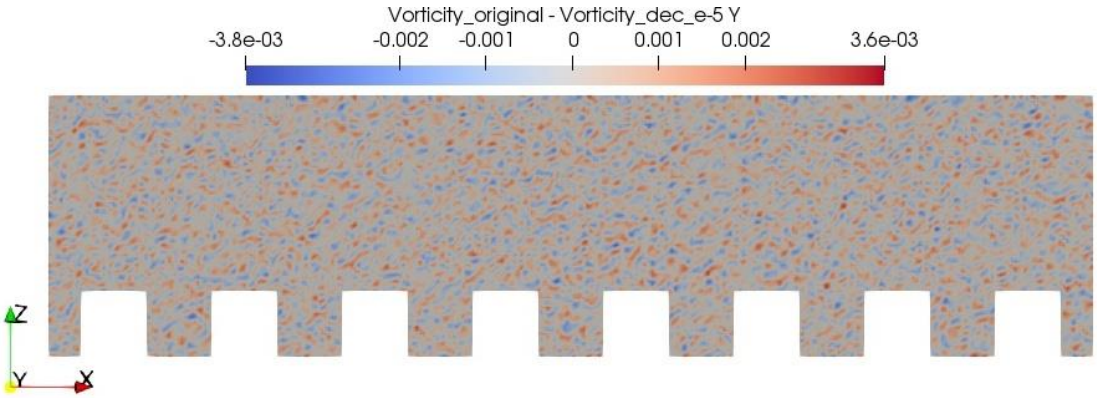
(c) Compression database with $\varepsilon = 10^{-1}$



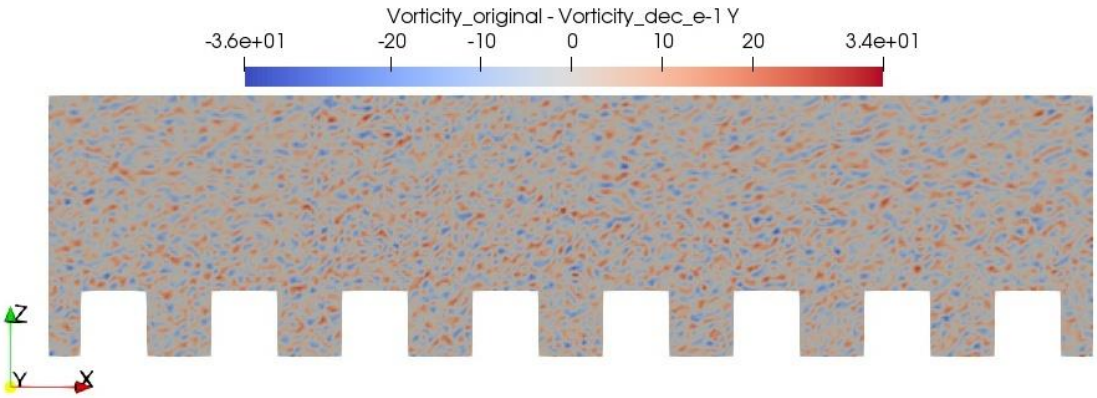
(d) Compression database with $\varepsilon = 10^{0}$

Figure 4.11. Vorticity field of spanwise direction in a vertical plane ($y = 7H$)

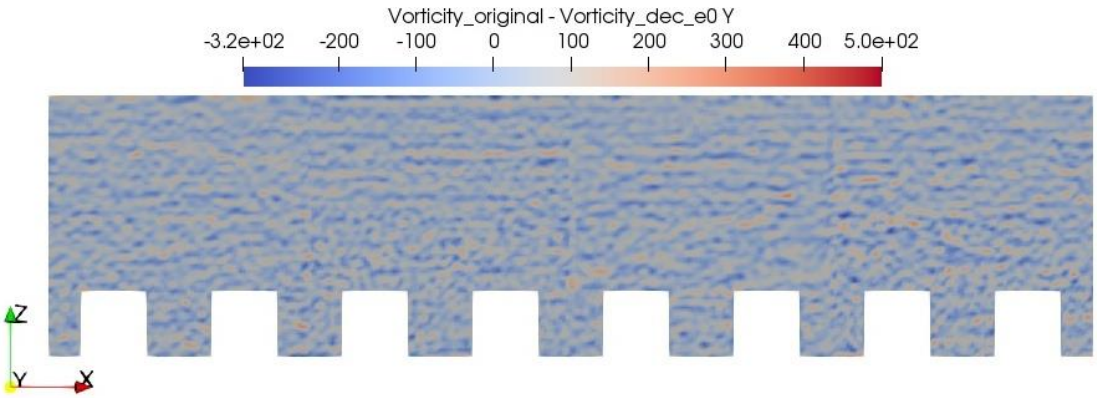
1



(a) Difference between original LES and compression database with $\varepsilon = 10^{-5}$



(b) Difference between original LES and compression database with $\varepsilon = 10^{-1}$



(c) Difference between original LES and compression database with $\varepsilon = 10^{0}$

Figure 4.12. Spanwise vorticity difference field between original LES and compression databases in a vertical plane ($y = 7H$)

1

As for the vortex structures, the vorticity field in the spanwise direction of the same plane in **Fig. 4.11** shows the same results. The difference fields between the original LES and compression databases can be checked in **Fig. 4.12** in the supplementary material. When ε is smaller than 1, the vorticity field is well maintained in the compressed database. The separation region and wake vortex are clearly seen. While $\varepsilon = 1$, many small vorticities are produced in the whole plane by the compression error, contaminating the original field.

A comparison of the velocity and vorticity field reveals that the compression database with $\varepsilon = 10^{-1}$ can be applied for the postprocessing and visualization of a single snapshot because the compression error effects are relatively small. In this case, the compression ratio can reach approximately 100 times according to **Fig. 4.7**. However, when $\varepsilon = 1$, the compression reaches the threshold where significant errors appear. This threshold where most information in the wavelet coefficients has been removed by the quantization process was confirmed in **Section 4.4.2**. In this case, the decompression cannot reconstruct the flow field properly. Considering that $\varepsilon \approx \frac{\max|f-\check{f}|}{\max|f|}$, the largest compression error is almost the same as the largest value at this threshold. Hence, there are many discontinuous large values in the velocity field, further resulting in unphysical, small vortexes, as shown in the vorticity field.

Another important issue about the velocity field is continuity, which is one of the most fundamental properties in fluid mechanics. Since the simulation case here is incompressible flow, the divergence of the velocity field in one snapshot is a suitable metric. In the ideal status, the divergence should be 0 everywhere in the flow when the continuity is promised, but the value will not be 0 in one snapshot of the numerical simulation due to the differential process and truncation error. **Fig. 4.13** compares the probability density functions of divergence of the velocity of all cells in a snapshot between original LES and three compression cases. Because the curves are too close to the original LES when $\varepsilon < 10^{-2}$, they are not shown here. The interpolation scheme was set as Gauss Linear in the OpenFOAM. As a reference, the average of the absolute value of velocity divergence in the original LES case is about 10($1/s$).

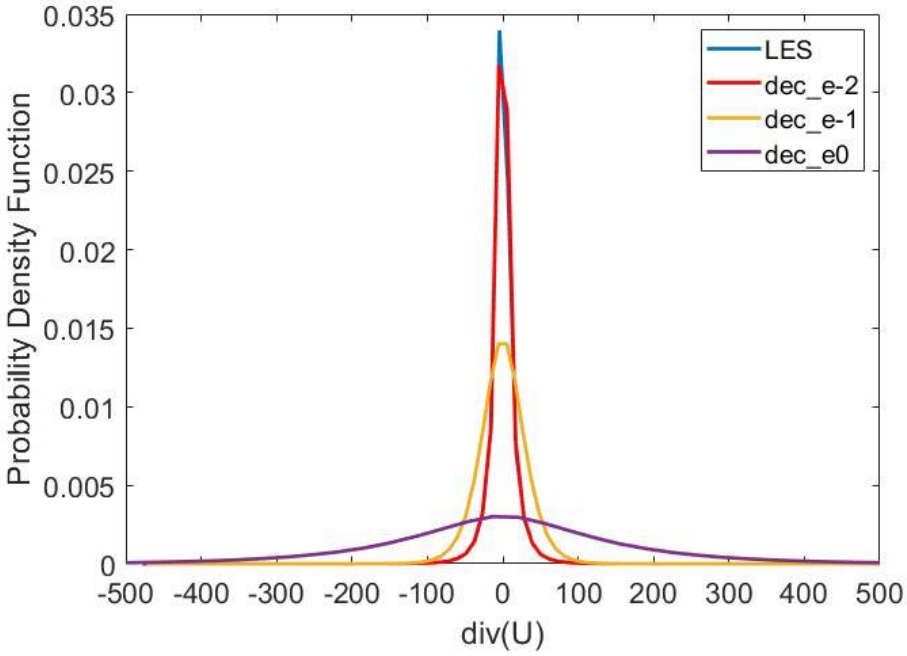


Figure 4.13. Comparison of probability density functions of velocity divergence in one snapshot between original LES and several compression databases: $\varepsilon = 10^{-2}, 10^{-1}, 10^{0}$

From **Fig. 4.13** it can be observed that the curve of the compression database with $\varepsilon = 10^{-2}$ is similar to that of the original LES, where the divergence concentrates near 0. The continuity only suffers small damage due to the compression. As ε becomes larger, the curve becomes wider and lower, which means the continuity in more areas was disturbed. When the compression error ε is imposed to the value 10^{-1}, the standard deviation of the velocity divergence changes from 17 (1/s) to 31 (1/s) holding the same mean value compared with the original LES. It is important to note that the cure changes significantly when ε turns to 10^{0}. The flattening curve with a large range shows that the continuity of the whole field has been corrupted. This result agrees well with the above one that the compression reaches its threshold when $\varepsilon = 1$.

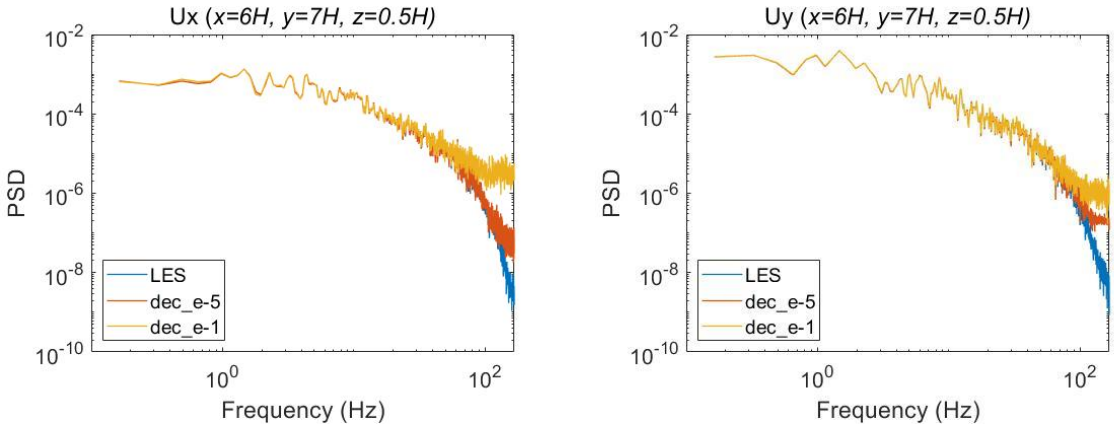
4.4.4 Compression effects on time-series data

The cumulative effects of compression on the database were investigated from the viewpoint of the velocity field and concentration of the simulated passive scalar.

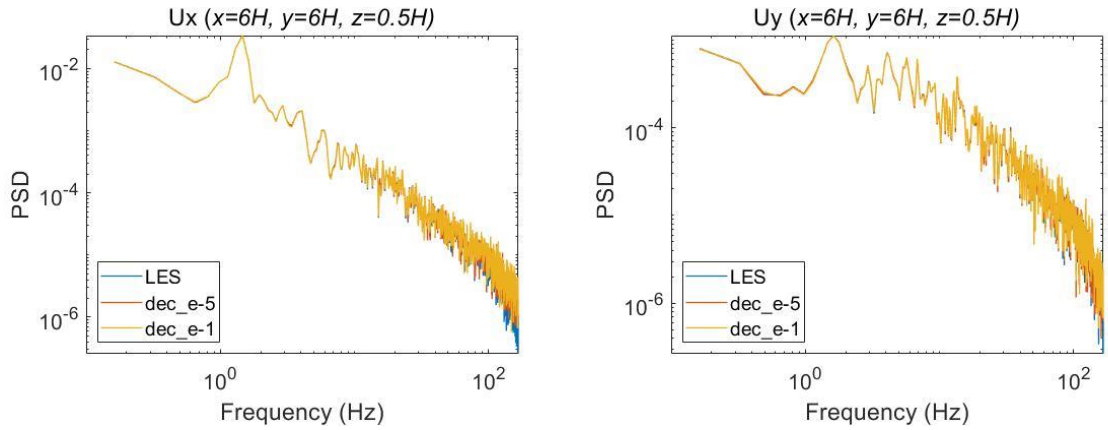
First, the power spectrum densities (PSDs) of velocities at two points located at the canyon region behind the buildings ($x = 6H$, $y = 7H$, $z = 0.5H$) and the open street region between the

rows of blocks ($x = 6H$, $y = 6H$, $z = 0.5H$) were compared for the original LES, compression database with $\varepsilon = 10^{-5}$, and compression database with $\varepsilon = 10^{-1}$. The results in **Fig. 4.14** show that the PSD curves of the three cases are almost the same when the frequency is relatively small, indicating that large-scale vortex structures are well preserved in the compression. For the high-frequency area, the larger ε, the larger the difference between the compressed and original data. This phenomenon can be explained by the principles of the WCM. WD decomposes the flow field into an approximation part with low frequency and a detailed part with high frequency. Because the truncation error in the quantization mainly deals with the detailed part whose wavelet coefficients are much smaller than those of the approximation part, after the compression, the high-frequency structures are highly coupled with truncation errors, which are manifested as white noise in the PSD. For the same reason, the compression data with $\varepsilon = 10^{-5}$ coincide with the original data longer and more closely than the data with $\varepsilon = 10^{-1}$ because the truncation error is set smaller. Additionally, for the point in the open street region, these three curves are closer than that in the wake region. The reason is that the velocity value is higher in this area, so the effects of compression error are weaker.

The probability density functions of the velocity at the same two points are summarized in **Fig. 4.15**. At both points, the velocity distribution of case $\varepsilon = 10^{-5}$ coincides well with the original data. As for $\varepsilon = 10^{-1}$, the distribution deviates at the place where the velocity is near 0 for the point in the canyon region. These small velocities are mainly contributed by a high-frequency small vortex, which suffers from the damage from the compression process to some degree. It is reasonable to conclude that the compression barely disturbs the occurrence probability distribution of the velocity field.



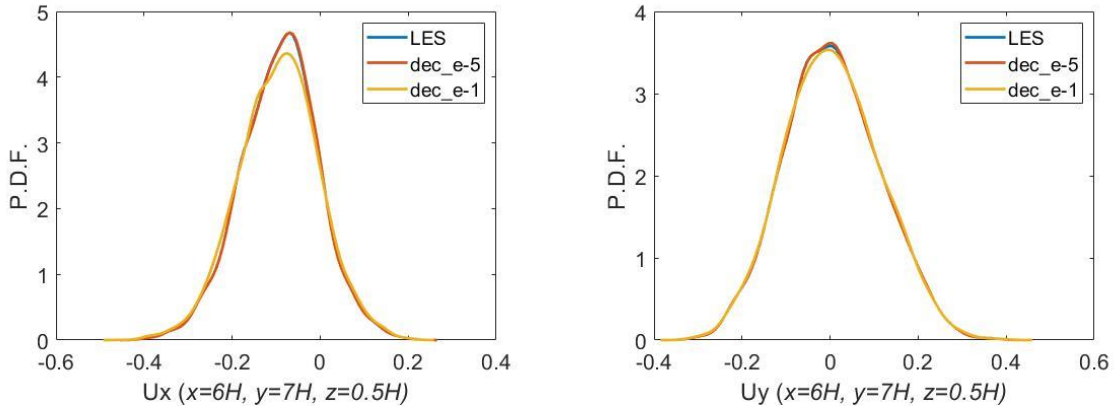
(a) Point at canyon region ($x = 6H$, $y = 7H$, $z = 0.5H$)



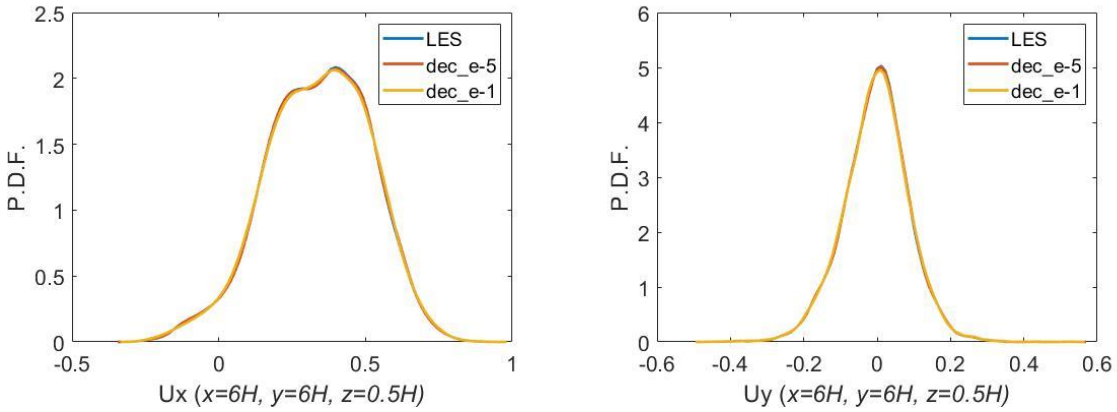
(b) Point at open street region ($x = 6H$, $y = 6H$, $z = 0.5H$)

Figure 4.14. Comparison of the PSD of velocity (streamwise and spanwise) at two points in three cases: original LES, compression database with $\varepsilon = 10^{-5}$ (dec_e-5), and compression database with $\varepsilon = 10^{-1}$ (dec_e-1)

1



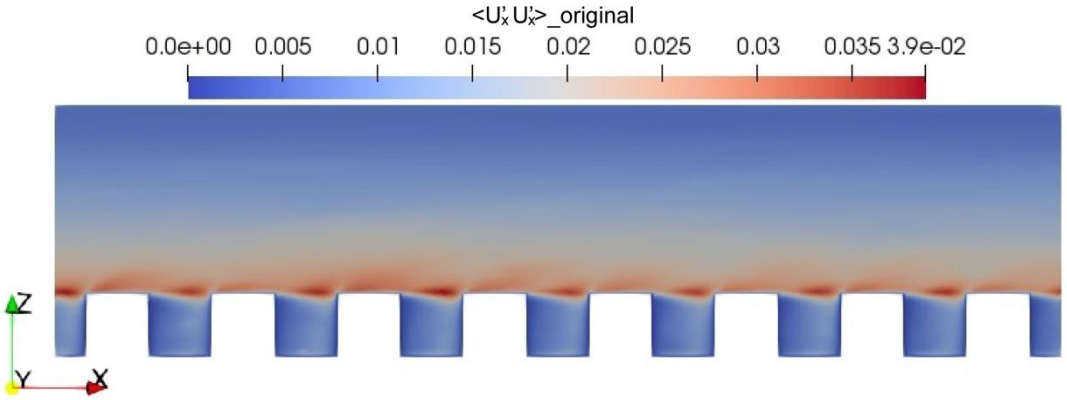
(a) Point at canyon region ($x = 6H$, $y = 7H$, $z = 0.5H$)



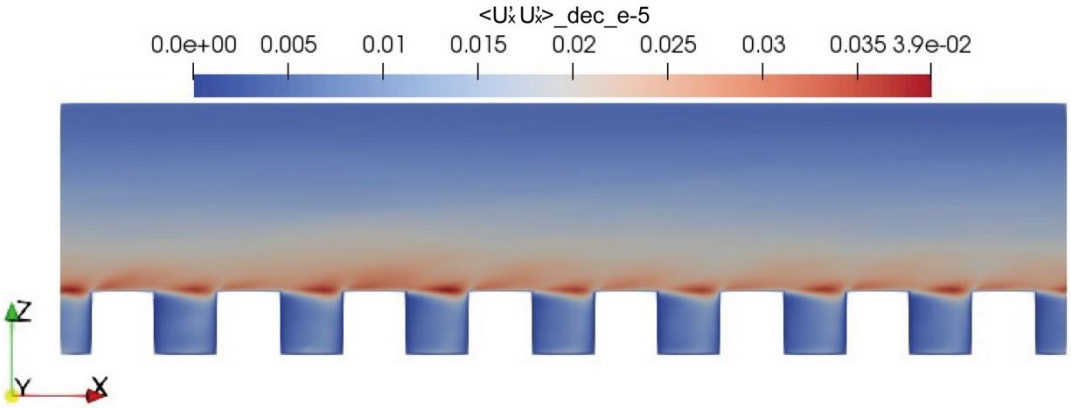
(b) Point at open street region ($x = 6H$, $y = 6H$, $z = 0.5H$)

Figure 4.15. Comparison of probability density functions of velocity at two points in three cases: original LES, compression database with $\varepsilon = 10^{-5}$ (dec_e-5), and compression database with $\varepsilon = 10^{-1}$ (dec_e-1)

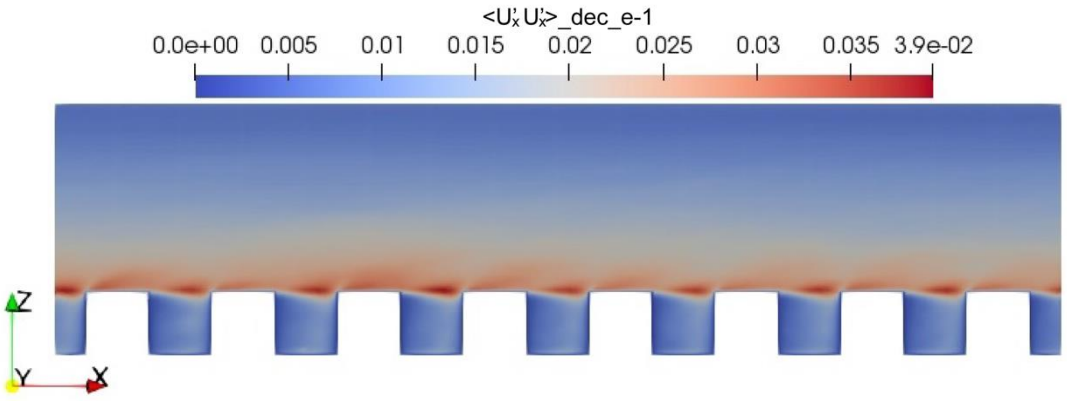
The effects of compression on the higher order statistics (variance and covariance of velocity fluctuations) are also checked. The distributions of variance $< U_x' U_x' >$, $< U_y' U_y' >$, $< U_z' U_z' >$, $< U_x' U_z' >$, $< U_x' U_y' >$ and $< U_y' U_z' >$in the middle vertical plane are presented in **Fig. 4.16 ~ 4.21**. Here, x, y, z denote the three components of the velocity field, $'$ means the fluctuations, and $<\bullet>$ is the time-averaged operator. It is shown that the 100 times compression preserves the second order statistics as well as the instantaneous flow field. Both the magnitude and shapes are almost the same as the original LES. The dominant co-variance structures caused by the separation and wake of the buildings are completely maintained.



(a) Original LES



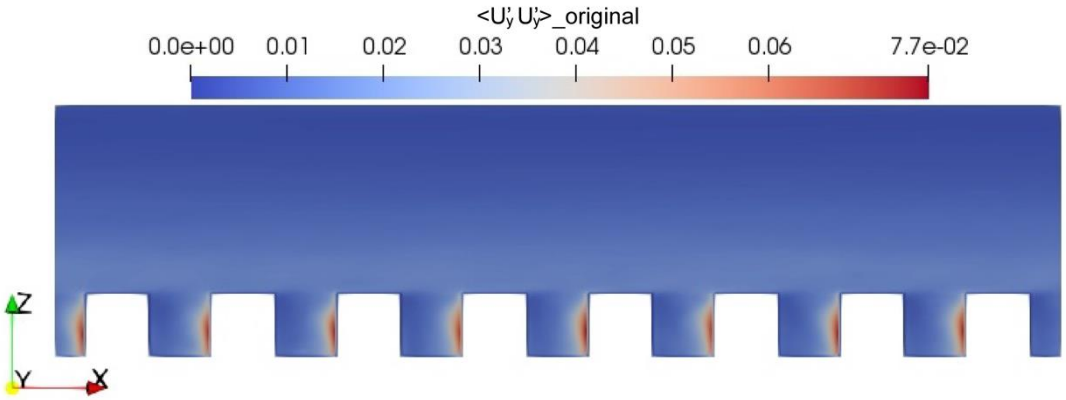
(b) Compression database with $\varepsilon = 10^{-5}$



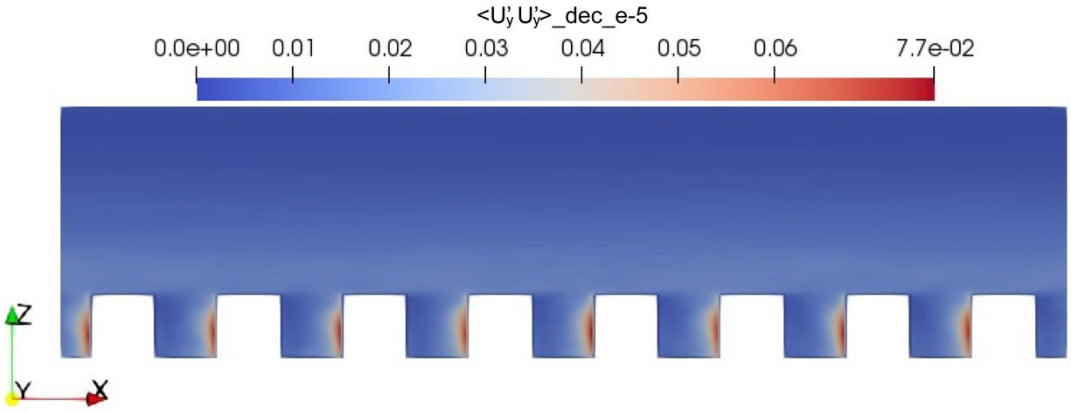
(c) Compression database with $\varepsilon = 10^{-1}$

Figure 4.16. The distribution of $< U_x' U_x' >$ of different cases in a vertical plane ($y = 7H$)

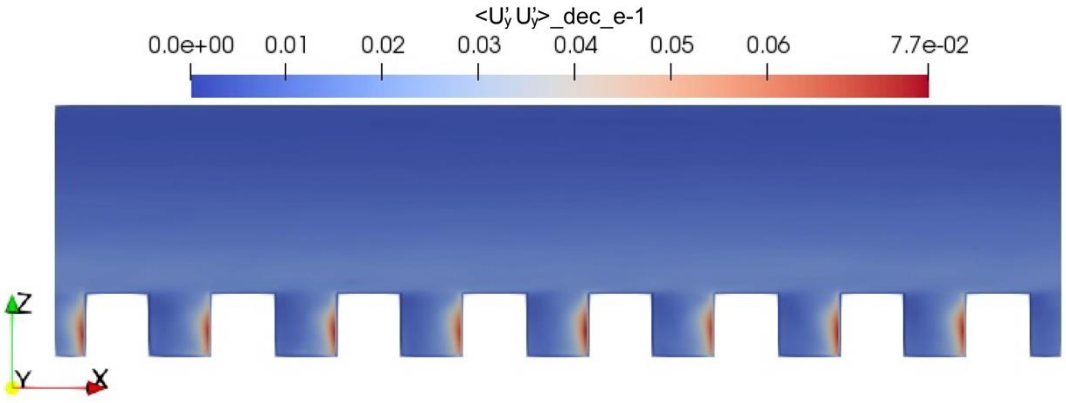
1



(a) Original LES



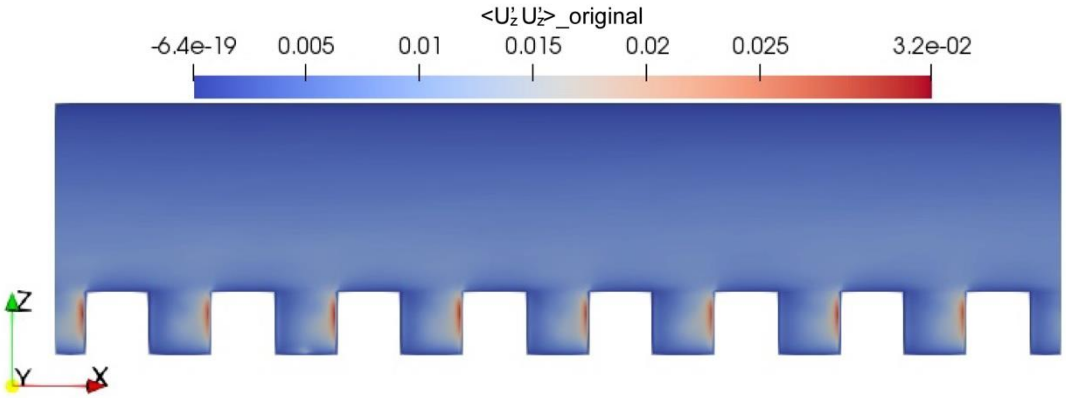
(b) Compression database with $\varepsilon = 10^{-5}$



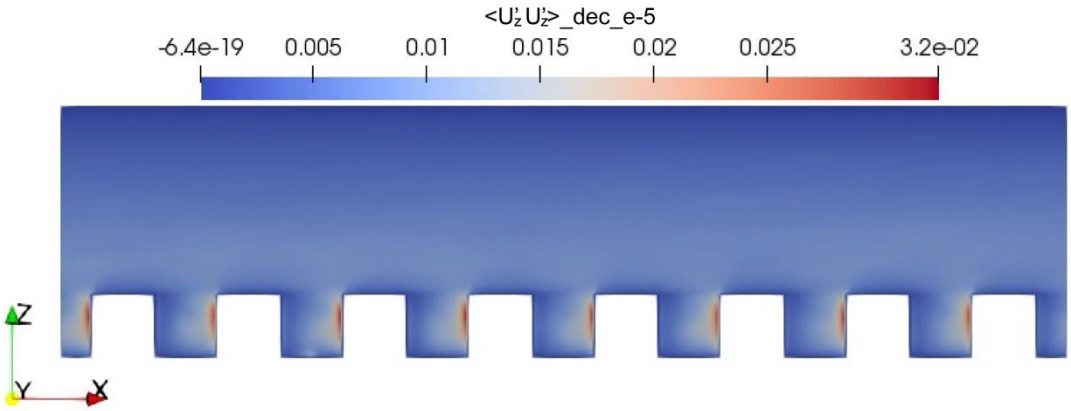
(c) Compression database with $\varepsilon = 10^{-1}$

Figure 4.17. The distribution of $< U_y' U_y' >$ of different cases in a vertical plane ($y = 7H$)

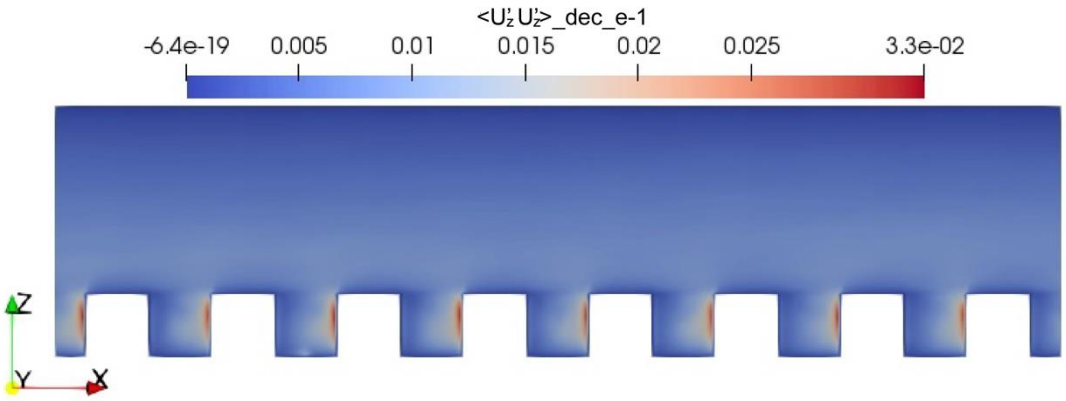
2



(a) Original LES



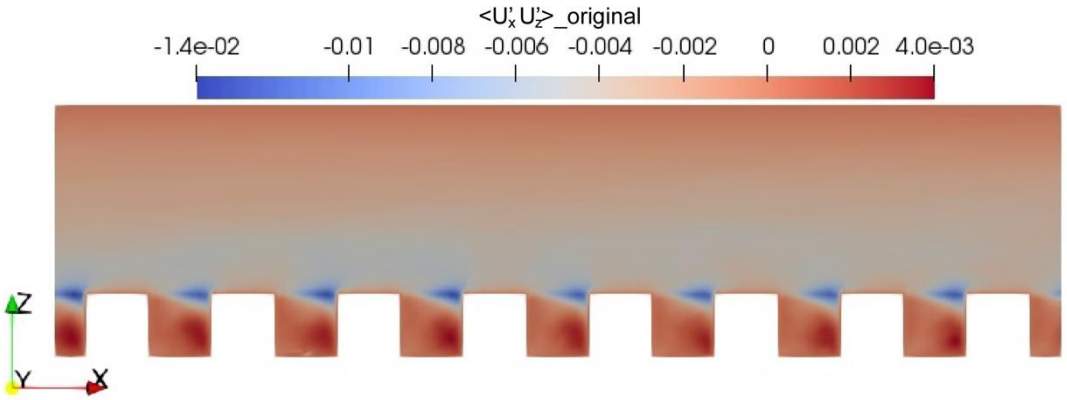
(b) Compression database with $\varepsilon = 10^{-5}$



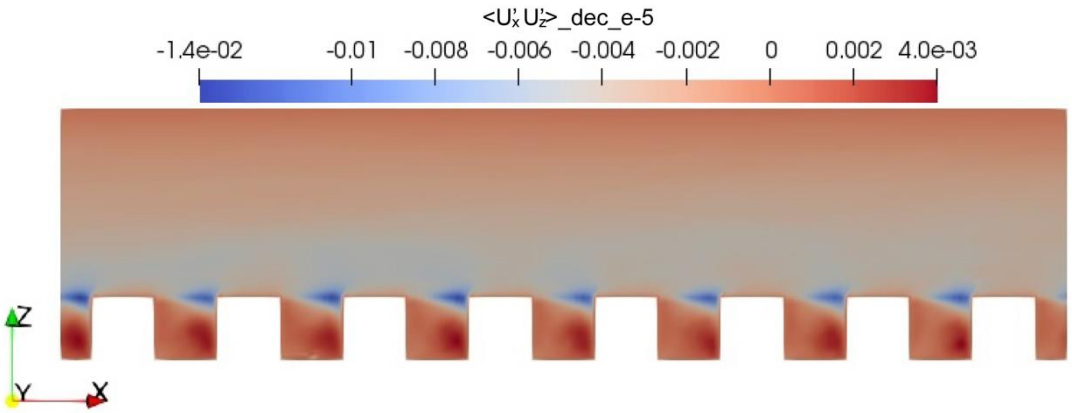
(c) Compression database with $\varepsilon = 10^{-1}$

Figure 4.18. The distribution of $< U_z' U_z' >$ of different cases in a vertical plane ($y = 7H$)

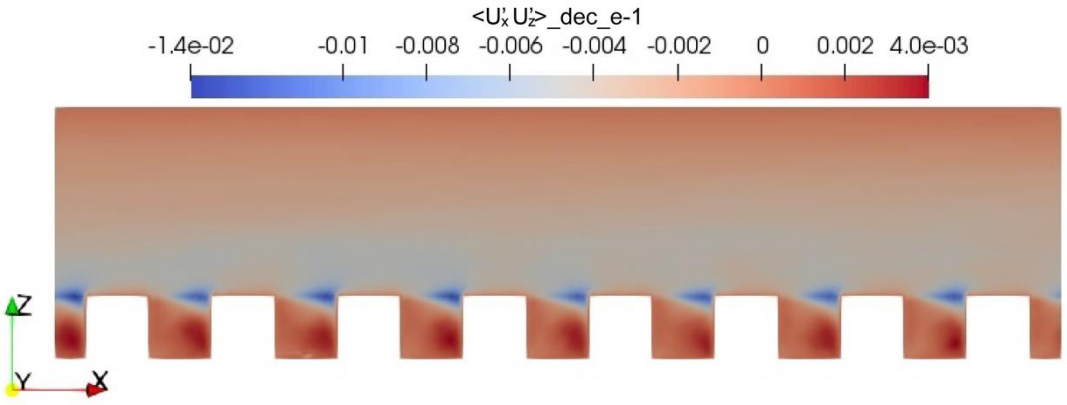
1



(a) Original LES



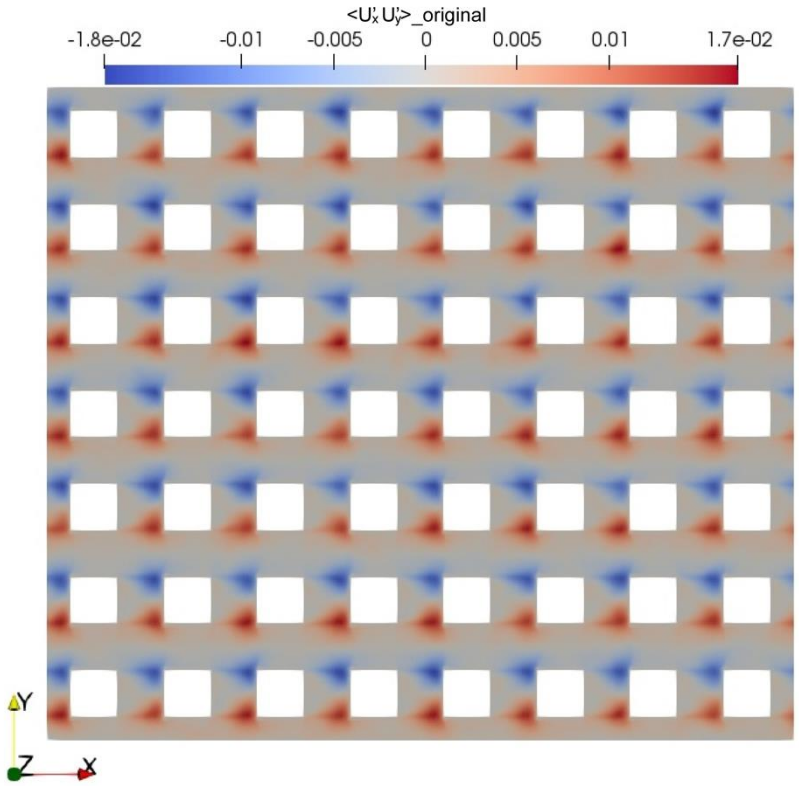
(b) Compression database with $\varepsilon = 10^{-5}$



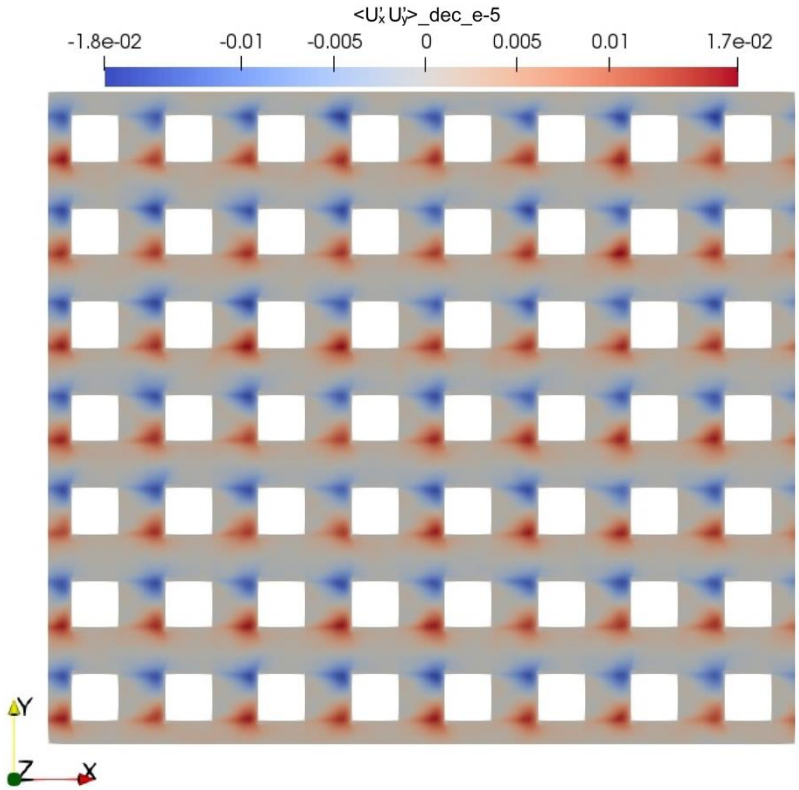
(c) Compression database with $\varepsilon = 10^{-1}$

Figure 4.19. The distribution of $< U_x' U_z' >$ of different cases in a vertical plane ($y = 7H$)

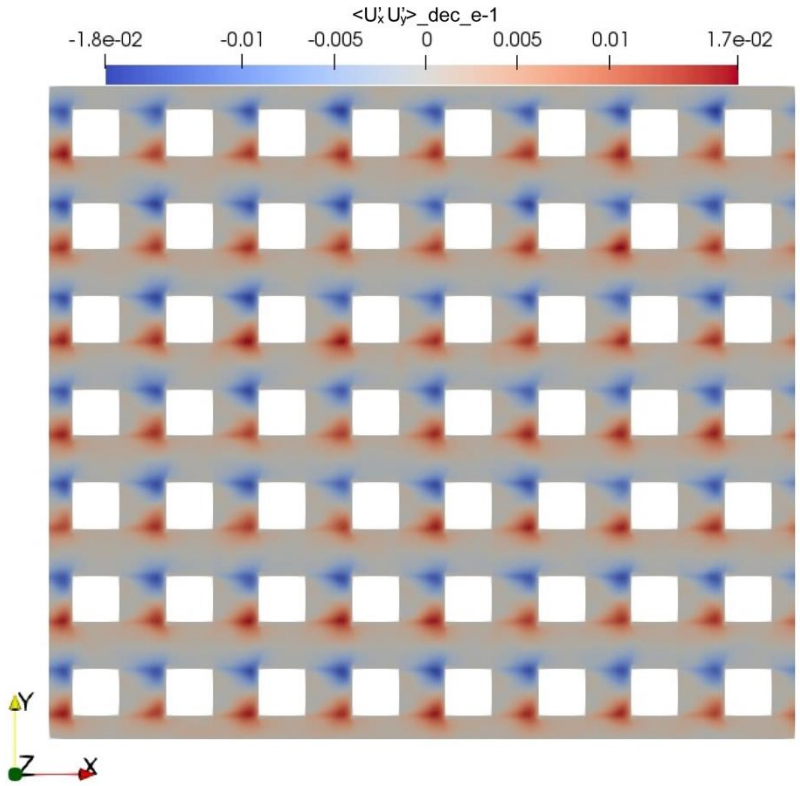
2



(a) Original LES



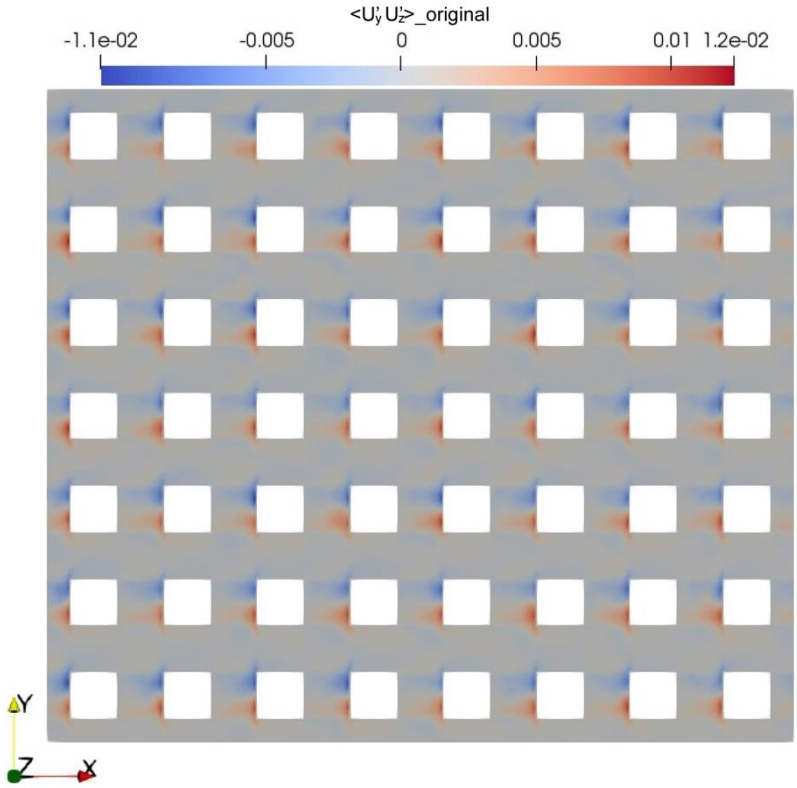
(b) Compression database with $\varepsilon = 10^{-5}$



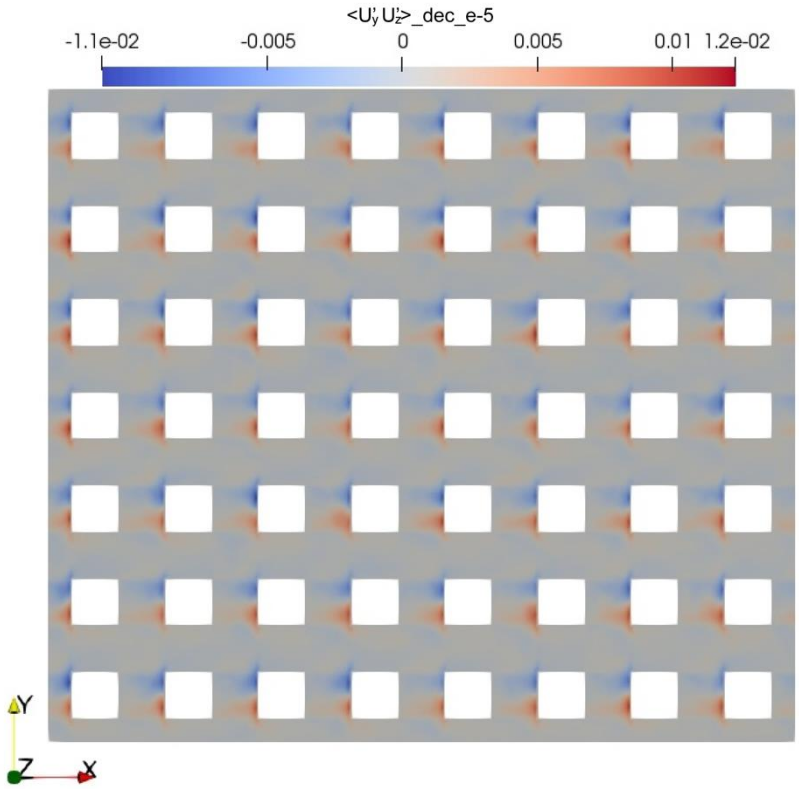
(c) Compression database with $\varepsilon = 10^{-1}$

Figure 4.20. The distribution of $< U_x' U_y' >$ of different cases in a vertical plane ($y = 7H$)

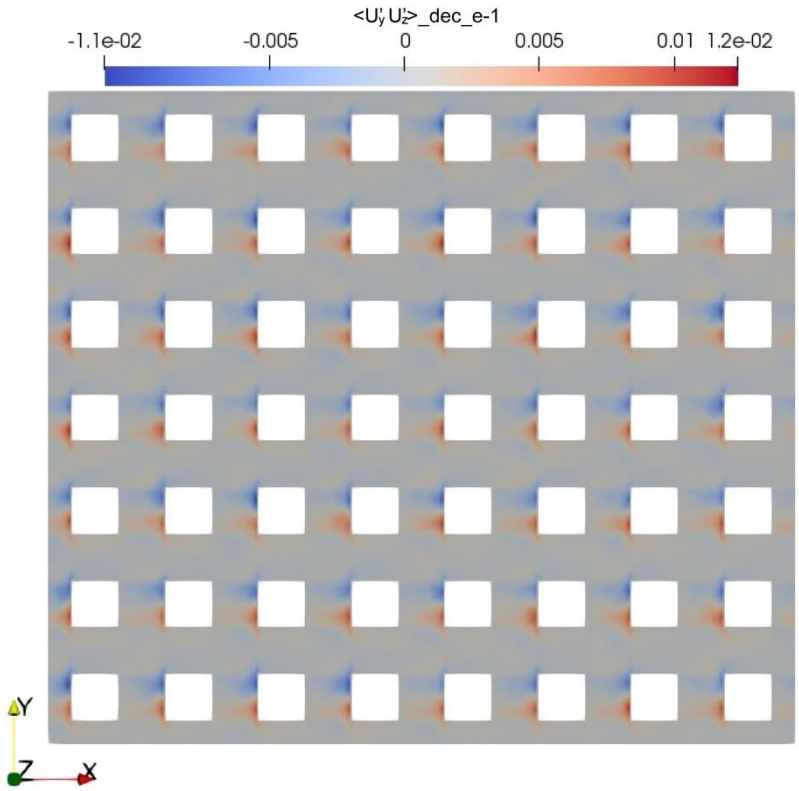
1



(a) Original LES



(b) Compression database with $\varepsilon = 10^{-5}$



(c) Compression database with $\varepsilon = 10^{-1}$

Figure 4.21. The distribution of $< U'_y U'_z >$ of different cases in a vertical plane ($y = 7H$)

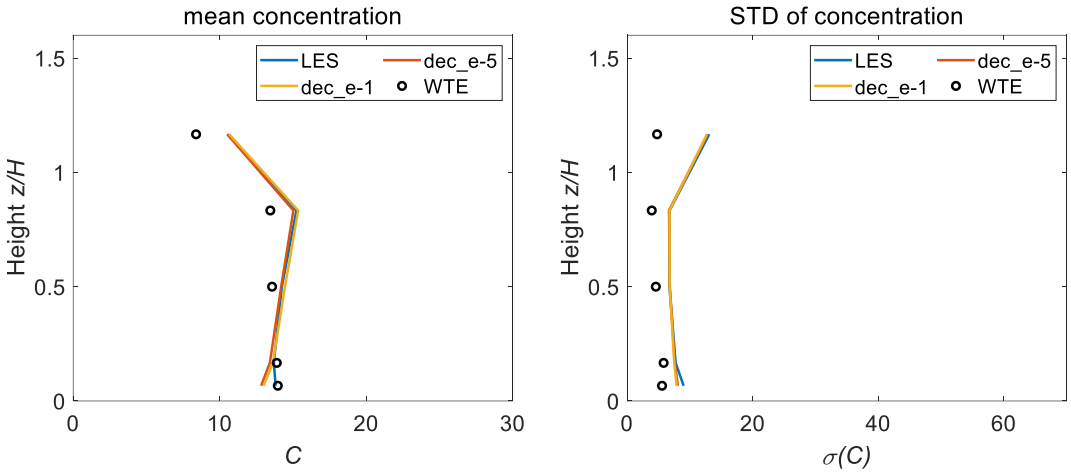
1

2 **4.4.5 Dispersion simulation with compression database**

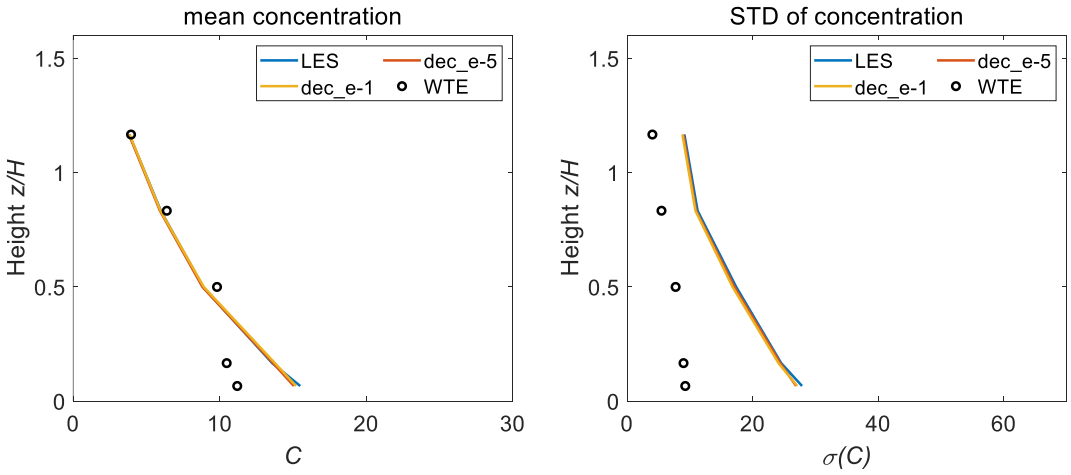
3 Then, the re-simulation of the passive scalar dispersion was evaluated with the compressed
4 database. Comparisons between the wind tunnel experiment, original LES, and compressed
5 databases at four representative locations — the canyon region near the source ($x = 6H$, $y =$
6 $7H$), street region near the source ($x = 6H$, $y = 8H$), canyon region away from the source ($x =$
7 $10H$, $y = 7H$), and street region away from the source ($x = 10H$, $y = 8H$) — can be observed in
8 **Fig. 4.22**. The original LES basically predicted the concentration distribution in accordance
9 with the experiment. At the location in the open street region near the source, LES slightly
10 overestimated the turbulent scalar flux. Because a method to simulate accurately the scalar
11 dispersion in the complicated building blocks is still being researched (Tominaga and
12 Stathopoulos, 2012) and was not the concern of this study, the overestimation was not analyzed
13 in detail.

14 Meantime, the concentration of the original LES was successfully reproduced by the

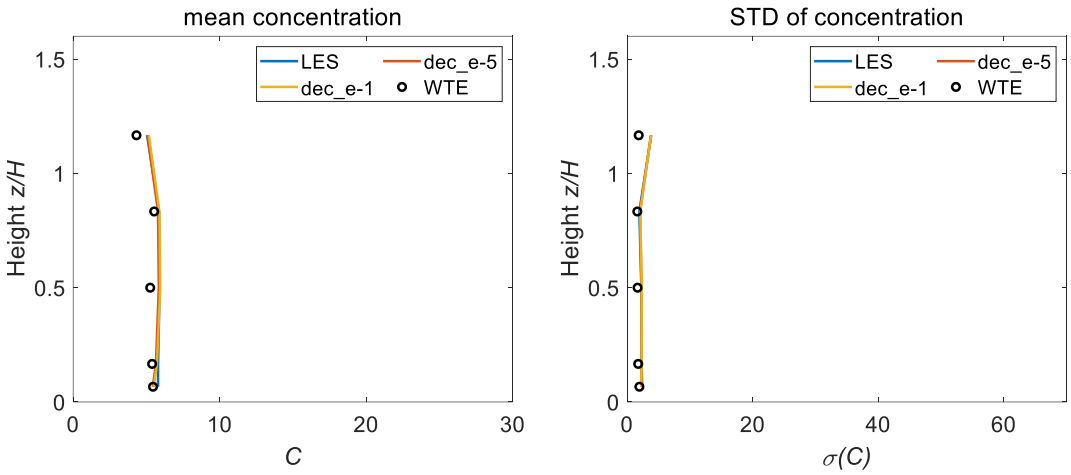
compressed data, among which the smallest data volume was only approximately 1% of the original one (compression database with $\varepsilon = 10^{-1}$). At the point in the canyon region near the source, the concentration of the compressed database had a small deviation near the ground. The velocity at this height was somewhat small, indicating that it is vulnerable to compression error. Given that the concentration flux is rather strong near the source, a slightly different velocity can result in a different concentration distribution and cause this kind of deviation. At the places that were relatively far from the source, the effects of compression became almost unnoticeable. Generally, not only the mean concentrations but also the standard deviation of all the simulations of the compressed database were close to that of the original LES results.



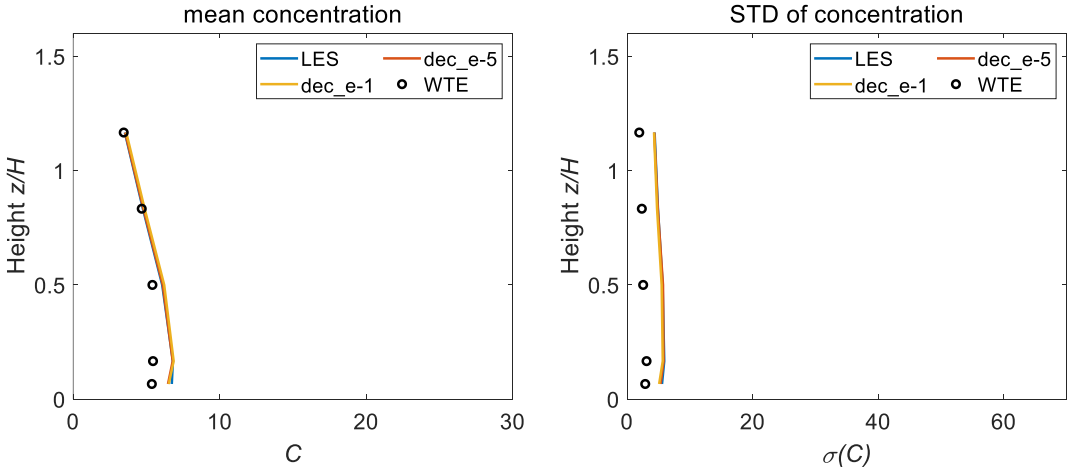
(a) $x = 6H$ $y = 7H$



(b) $x = 6H$ $y = 8H$



(c) $x = 10H$ $y = 7H$

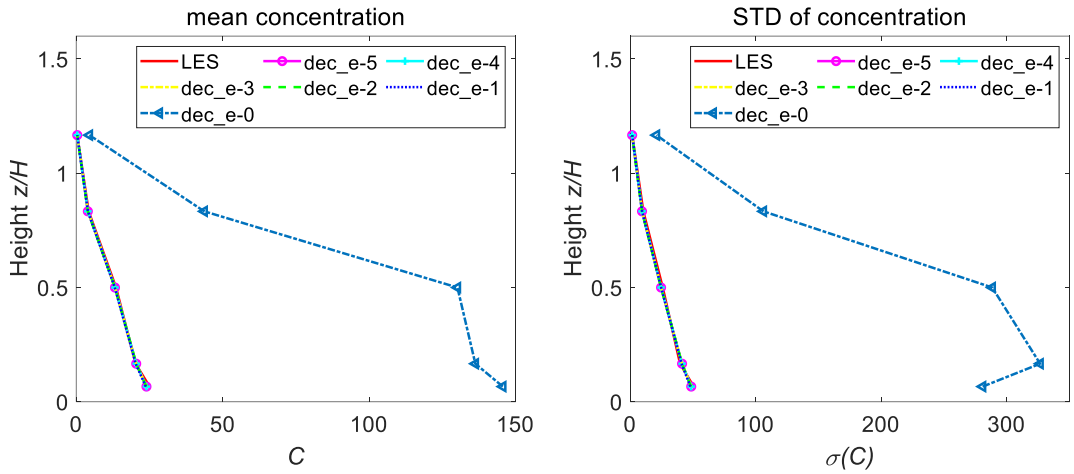


(d) $x = 10H$ $y = 8H$

Figure 4.22. Comparison of the concentration at different locations in four cases: wind tunnel experiment, original LES, compression database with $\varepsilon = 10^{-5}$ (dec_e-5), and compression database with $\varepsilon = 10^{-1}$ (dec_e-1) — left: time mean concentration; right: standard deviation of concentration

When the compressed database with $\varepsilon = 10^{0}$ was used, the statistical result of concentration totally collapsed within only 1s (approximately $5T$) simulation time, while the other cases remained stable (**Fig. 4.23**). As a consequence, the concentration simulation was insensitive to the compression error under the threshold. When ε was loosened to 1, an enormous error appeared. Once again, the reason was that the truncation error in the compression was so large that the main information in the wavelet coefficients was removed;

therefore, the flow structures were damaged, and the reconstructed flux could no longer transport the passive scalars correctly.



$x = 4.5H \quad y = 6H$

Figure 4.23. Comparison of the concentration sampled in a short time in seven cases: original LES and all the compression cases — left: time mean concentration; right: standard deviation of the concentration

To sum up, the dispersion of passive scalars can be obtained later, as in the original LES simulation with a compressed database with $\varepsilon = 10^{-1}$, whose volume was approximately 1% of the original one according to **Fig. 4.7**.

4.5 Conclusions and discussions

4.5.1 Conclusions

The feasibility of constructing a small urban flow database by compressing the original CFD data with the WCM was investigated. The original data were flow fields in a block-arrayed urban area simulated by the LES model. Several compressed databases were constructed with different error controls ε: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, and 10^{0}.

The compression method was based on WD, quantization, and entropy encoding. The original data were first transformed by WD to the approximate part with a small volume and detailed parts with a large volume. The quantization process decided the proportion of detailed parts that should be retained according to the user-specified error control ε and caused the

truncation error ε_F. Finally, entropy encoding was implemented to compress the data further.

The compression ability of the WCM was studied. The empirical formula for error control guarantees that the prescribed error is exactly realized in the compression. The compression ratio increases with ε exponentially until it reaches the compression limitation, approximately 1000 times in this study, where $\varepsilon = 10^0$ is so large that most of the wavelet coefficients are deeply stained. It was found that, although the error control that can reach this compression limitation is found according to each case, it can be calculated from the distribution of the wavelet coefficient in advance.

The influence of compression on the quality of the database was checked from two viewpoints: single snapshot and time-series data. On one hand, for a single snapshot used in postprocessing and visualization, the velocity and vorticity fields are almost the same as long as ε is less than the threshold of 10^0, which means that the compression database with approximately 1% of the volume of the original data is sufficient for visualization. When compression reaches the threshold, a considerable number of unphysical large values in velocity and small vorticities appear, and the flow structures are highly contaminated. On the other hand, the cumulative effects were inspected by comparing the velocity field and different dispersion results re-simulated with the compressed database to the original LES results. According to the PSD velocity results, the large-scale vortex structures with low frequencies are well preserved in the compression. The compression errors function as white noise superimposed on the high-frequency structures. Moreover, the dispersion re-simulation is insensitive to the compression error below the compression limitation. The dispersion process can be reproduced closely, even by rough compressed data with approximately 1% of the original volume. A considerable discrepancy appears near the compression limitation where the flow field can no longer be reconstructed properly.

Overall, the results in this chapter show that WCM is a powerful tool to construct a portable flow database. The possible applications of a database, such as visualization of a single snapshot and re-simulation with time-series data, can be appropriately realized, even under approximately 100-times compression. Therefore, in the next chapter, the unsteady adjoint equation will be simulated by the compression database constructed here, and the performance in source identifications will be checked.

4.5.2 Method limitation and future research

It has to be admitted that this chapter is one of the early attempts of a compressed turbulence database focusing on a complicated urban environmental flow, more situations in other areas should be examined before the general application of WCM. The current method still has some limitations which can be improved in future research. Since the method consists of three steps: WD, quantization, and entropy encoding, a brief discussion about method limitation and possible future research plans is conducted in these three aspects.

The first one is WD. In the current method, all the wavelet coefficients are passed exactly to quantization, which means it is lossless and unrelated to compression error control. The improvement strategy should focus on the calculation speed and coefficients production. The core part of WD is the wavelet kernel function. The kernel function that requires less calculation resource and yields fewer coefficients is favorable to the compression process. However, it seems that these two properties are contradictory to each other. Although the current kernel CDF9/7 performs well in the coefficients production, it is one of the most complicated kernels because of its irrational decomposition coefficients. It is worthwhile to testify the performance of other improved kernels based on CDF9/7 in the future. Meanwhile, even though CDF9/7 is highly recommended in image compression, it cannot be simply asserted to be the best one for turbulence compression. After all, image compression focuses on the visual effects to humankind, in which the low-frequency domain is the most important part, while turbulence focuses on the physical information, which is contained in different scales structures with a large range of frequency. Deeper research is necessary to decide the most appropriate kernel for turbulence decomposition.

The only step concerning error control is quantization. The main limitation of the method is the way to control compression error. According to the definition of the error metric $\varepsilon_r = \frac{\max|f-\check{f}|}{\max|f|}$, the error control is closely related to the maximum value of the flow field data and maximum compression error rather than the spatial mean value or the range of all values. However, the place where the maximum compression error happens is not promised to be the place where the flow field takes the maximum value. In other words, the situation of all places in the field cannot be independently determined by the current error metric so that someplace may suffer unexpected errors resulting from the maximum field data. For example, if the inflow

speed is increased in the case of this paper, the maximum velocity in the high place will increase correspondingly. It is possible that the compression error of the low velocity area near walls will be increased even if the compression algorithm is set with the same ε and yields a close compression ratio. Therefore, it has to be very cautious to apply this method to compress the flow field with a wide range of values. In the common urban flow simulation, the range of physical values is mild and high Reynolds number flow is the main concern, author is optimistic about the application of the current method. However, special attention needs to be paid to when both the low and high Reynolds number turbulence matters.

Entropy encoding is a lossless compression algorithm that is also included in other compression methods like JPEG2000 as a standard process. It is reasonable to say that the effectiveness of this step would not be changed by different application scenarios of turbulence compression.

1 Symbols

C : the concentration distribution

C_r : standard (reference) concentration

C_s : Smagorinsky constant for LES model (=0.12)

C_{gas} : the emission strength of the source

D_e : effective diffusion ($D_m + D_{sgs}$)

D_m : molecular diffusivity

D_{sgs} : sub-grid scaled turbulent diffusivity

$\widetilde{D_{sgs}}$: decompressed sub-grid scaled turbulent diffusivity

$f_{i,j,k}$: the original data of a three-dimensional scalar field in the Cartesian indexing mesh grid

$\check{f}$: the decompressed flow field data

$F_{i,j,k}$: the wavelet coefficients obtained from the wavelet decomposition of raw data $f_{i,j,k}$

$\check{F}$: the wavelet coefficients reconstructed from quantization results Q

H : the edgy length of blocks in the simulation (=60 mm)

h_r : reference height

N : the number of 1-byte memory spaces

q : the gas flow rate at the source

Q : the results produced by quantization operation

r : the compression ratio

S : source term in the transport equation

Sc_{sgs} : sub-grid scaled turbulent Schmidt number

T : standard time-scale of LES simulation

u_r : reference speed

u_* : friction speed

$U_x,\ U_y,\ U_z$: streamwise, spanwise, and vertical velocity

$\boldsymbol{U}$: velocity field

$\breve{\boldsymbol{U}}$: decompressed velocity field

V_o : storage volume of raw data

V_c : storage volume of compressed data

ε_F : truncation errors between F and $\breve{F}$ caused by quantization operation

ε : user-specified allowable error for the WCM

ε_r : real compression error metric between f and $\breve{f}$

η : empirical coefficient for error control of WCM (=1.75)

ν_{sgs} : eddy viscosity coefficient at the sub-grid scale

$\sigma(\cdot)$: standard deviation

$(\cdot)'$: fluctuation value with time

$<\bullet>$: time-averaged operator

1

2

Chapter 5
Source term estimation with unsteady adjoint equations modeled by large-eddy simulation and compression database

Abstract

This chapter tries to develop a new source term estimation (STE) method in which the unsteady simulation of adjoint equations is embedded via large eddy simulation (LES) into a Bayesian inference framework. The LES of adjoint equations is based on the compression database constructed in **Chapter 4**. The performance of the proposed model is validated using a point source dispersion case in a regular, block-arrayed, urban model wind tunnel experiment. To clarify the effects of different computational fluid dynamics models on the simulation of adjoint equations and the accuracy of STE, an existing (Xue et al., 2018b) Reynolds averaged Navier-Stokes model using a time-averaged LES flow field is selected for comparison.

5.1 Introduction

An adjoint equation is a partial differential equation that is similar in form to a dispersion equation and can be simulated by computational fluid dynamics (CFD) models. Among them, two popular methods are the Reynolds averaged Navier-Stokes (RANS) model and large eddy simulation (LES). The main difference between these models is the way in which turbulent diffusion is simulated. The RANS model is based on an ensemble-averaged governing equation and approximates fluctuations in the turbulent flux using a mean field according to the gradient diffusion hypothesis (GDH), which assumes that turbulent diffusion is proportional to the gradient of the mean concentration field. Meanwhile, an LES explicitly resolves most turbulent effects using a fine mesh. The RANS model has been frequently used to simulate adjoint equations (Efthimiou et al., 2018b; Keats et al., 2007b; Kumar et al., 2015b) and is capable of providing more accurate dispersion predictions than a Gaussian dispersion model, with a mild increase in computational resources. However, the prediction accuracy of RANS models for the time-averaged flow fields around buildings has been shown to be insufficient when compared with LES (Tominaga et al., 2008a), which undermines the accuracy of their adjoint equation simulations and source term estimation (STE). Xue et al. (2018b) noticed this defect and suggested that the coupling of a RANS-like simulation of adjoint equations with the time-averaged flow of LES can improve the accuracy of STEs. In these RANS and RANS-like simulations, turbulent diffusion is always modeled according to the GDH, regardless of whether the mean flow field was produced by a RANS model or LES. However, the validity of the GDH on dispersion remains unclear, especially across complex urban terrains.

It has been confirmed that the dispersion prediction accuracy of RANS models operating according to the GDH is limited when compared with that of explicit models with LES (Tominaga and Stathopoulos, 2012, 2011b). Until now, the LES of adjoint equations has been regarded as impractical because the embedded adjoint equation represents an inverse dispersion process, such that the time-series flow field data of the entire domain must be produced by forward simulation and stored in advance to realize the unsteady simulation. One important challenge is that the volume of data acquired by LES seems too large for practical application. Nevertheless, this problem can be solved by the application of the compression method. It has been shown in **Chapter 4** that the wavelet-based compression method can now compress data to ~100 times their original size, while simultaneously conserving their accuracy.

Another limitation is the heavy calculation burden brought by LES of highly-complicated real urban areas and data compression. Considering that the source-receptor relationship can be constructed as databases long before the emergent injection of hazardous materials into the atmosphere, the requirement for more calculations caused by this method is not a problem. Indeed, it enables the real-world application of the LES of adjoint equations. Therefore, this chapter evaluates the applicability of STE with LES of the adjoint equation in an urban-like experiment.

5.2 Simulations for adjoint equation

Before the Bayesian inference, it is necessary to clarify the difference between RANS and LES of adjoint equation simulation.

5.2.1 Reynolds averaged Navier-Stokes simulation of the adjoint equation

The RANS simulation for the adjoint equation has been introduced in **Chapter 2**. It can be noticed that Eq. (2.24), RANS form of the adjoint equation, only explicitly resolves the mean flow advection, so the effects of turbulent diffusion, $\overline{\boldsymbol{u}'C^{*\prime}}$ (where $'$ denotes the temporal fluctuations of variables) must be modeled by the mean field. One simple way of achieving this is through the GDH (Combest et al., 2011):

$$\overline{\boldsymbol{u}'C^{*\prime}} = -D_t \frac{\partial(\overline{C^*})}{\partial \boldsymbol{x}} = -\frac{\nu_t}{Sc_t}\frac{\partial(\overline{C^*})}{\partial \boldsymbol{x}} \tag{5.1}$$

where D_t is the turbulent diffusivity, ν_t is the eddy viscosity, which can be estimated based on the mean velocity and Reynolds stresses, and Sc_t is the turbulent Schmidt number. However, there are at least three problems in this GDH approximation. First, it assumes that the turbulent scalar flux is aligned with the mean scalar gradient, which is invalid under anisotropic turbulence. Secondly, ν_t is estimated by the mean velocity and Reynolds stresses predicted by the RANS model, which has been shown to inaccurately predict reality (Tominaga et al., 2008a). Finally, Sc_t is usually set as a global constant in the domain for simplicity, but it actually spans a large range of spatial differences (Combest et al., 2011; Tominaga and Stathopoulos, 2012). Thus, there remains no satisfactory way to model the spatial distribution of Sc_t.

Following Tominaga and Stathopoulos (2012), who clarified that accurate predictions of mean flow fields can improve the estimation of ν_t and further improve RANS predictions of dispersion fields, Xue et al. (2018) conducted LES-based forward simulations to obtain a time-averaged

velocity field and used it in a RANS-like (hereafter, RANS) simulation for adjoint equations. In this case, the eddy viscosity was calculated as:

$$\nu_t = -\frac{\sum_{i,j} \overline{u_i' u_j'} S_{ij}}{\sum_{i,j} 2{S_{ij}}^2} \tag{5.2}$$

where S_{ij} is the main strain rate:

$$S_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u_i}}{\partial x_j} + \frac{\partial \overline{u_j}}{\partial x_i}\right) \tag{5.3}$$

The mean velocity, Reynolds stresses, and strain rate were obtained by the forward LES. The accuracy of STE was confirmed to have improved because LES can predict the mean flow field in urban areas more accurately than RANS models. Despite this, the RANS simulation of adjoint equations is still imperfect due to the existence of the other two problems related to the GDH. In real applications, this simplification may cause significant errors in the modeling concentration, $\boldsymbol{R}$, especially in complex urban areas with many buildings.

5.2.2 Large eddy simulation of the adjoint equation

In contrast, LES appears to be a promising choice since it can explicitly resolve most of the turbulent diffusion with the following adjoint equation.

$$-\frac{\partial(\widetilde{C^*})}{\partial t} - \widetilde{\boldsymbol{u}}\frac{\partial(\widetilde{C^*})}{\partial \boldsymbol{x}} - \frac{\partial}{\partial \boldsymbol{x}}\left([D_{sgs} + D_m]\frac{\partial(\widetilde{C^*})}{\partial \boldsymbol{x}}\right) = \delta(\boldsymbol{x} - \boldsymbol{x}_m) \tag{5.4}$$

Here $\widetilde{C^*}$ represents the grid-scale value of the variables. The sub-grid scale (SGS) turbulent diffusion is modeled by:

$$\widetilde{\boldsymbol{u}'C^{*\prime}} = -D_{sgs}\frac{\partial(\widetilde{C^*})}{\partial \boldsymbol{x}} = -\frac{\nu_{sgs}}{Sc_{sgs}}\frac{\partial(\widetilde{C^*})}{\partial \boldsymbol{x}} \tag{5.5}$$

where ν_{sgs} is the SGS eddy viscosity and Sc_{sgs} is the SGS turbulent Schmidt number. When the grid is fine enough, most of the turbulent diffusion is directly calculated and the modeled SGS is very small. Compared with RANS models, LES can more accurately simulate adjoint equations.

The main limitation to the practical application of LES is that since the adjoint tracer dispersion field is a reverse simulation with $-\widetilde{\boldsymbol{u}}(\boldsymbol{x}, t)$, simulating the unsteady dispersion field requires saving all of the $\widetilde{\boldsymbol{u}}(\boldsymbol{x}, t)$ and D_{sgs} data for the entire domain and across all timesteps in advance.

Furthermore, LES always requires a fine mesh and short timesteps to ensure the accuracy and convergence of simulations, meaning that the data saved for the LES of adjoint equations require such massive storage volumes that it renders the method impractical. As a proposed solution, the compression database in **Chapter 4** is used for an LES of adjoint equations.

5.3 Case description

As for the study case, the block-arrayed urban model in **Chapter 4** is applied to represent the geometry of a real urban neighborhood. This model has been frequently used in previous studies on airflow (Cheng and Castro, 2002; Uehara et al., 2000; Xie and Castro, 2006) and the dispersion characteristics of atmospheric pollutants (Branford et al., 2011b; Coceal et al., 2014b; Tominaga and Stathopoulos, 2012) in urban areas. It includes complicated unsteady turbulent structures that challenge the robustness of STE methods. Moreover, there is an open WTE database with the same model that can be used as the measurements of concentration $\boldsymbol{D}$.

A forward simulation with an LES model was utilized to predict the unsteady flow field, which was compressed and stored in each time step. The source–receptor relationship was then obtained by the unsteady simulation of adjoint equations using compressed flow data. To evaluate the improvement made to the estimation accuracy of the proposed method, the RANS simulation of adjoint equations was also performed with the time-averaged flow field of the LES. Bayesian inference of the source was conducted twice based on the RANS and LES models.

5.3.1 Wind tunnel experiment

In the wind tunnel experiment, a continuously releasing point source (a hole with a diameter of 2 mm) was placed at the ground with coordinates of ($4H$, $7H$, 0). Pure ethylene gas (C_2H_4) was released at a flow rate of Q= 0.216 L/min from the source. Because the gas was released in the wake region of one block and transported downstream via complex turbulence, the setting was identified as a difficult scenario for the STE.

Velocities and concentrations were measured at different heights (H/15, H/6, H/2, 5H/6, 7H/6) following the same symmetrical horizontal sensor configuration shown in **Fig. 5.1**. The wind velocity was measured with split-film probes and the concentration was measured with a rapid-

response flame ionization detector. The measurements were sampled every 120 s in the frequency of 1000 Hz. Here, considering the symmetrical characteristics of the sensor configuration and the calculation burden of simulating adjoint equations, the time-averaged measurements of only 16 sensors in the horizontal plane with a height of 0.5H were used to construct the vector $\boldsymbol{D}$ (solid cycles in **Fig. 5.1**).

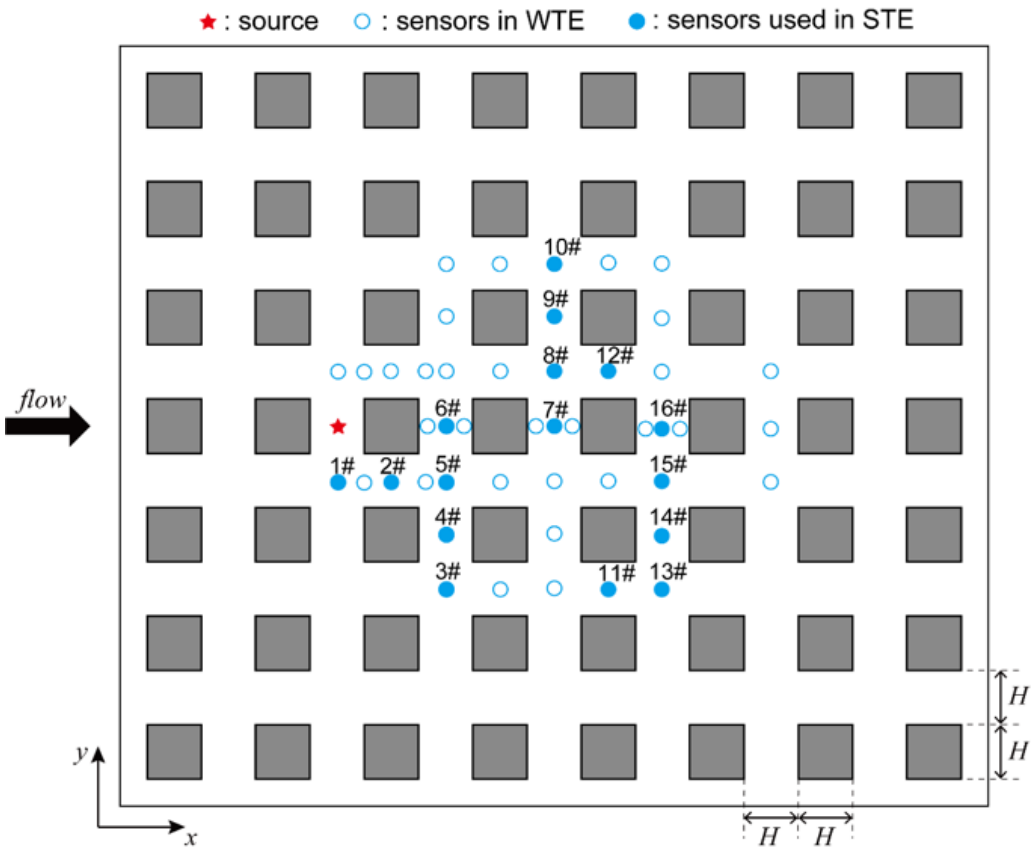


Figure 5.1. Schematic of sensor configuration (horizontal plane with $z = H/2$).

5.3.2 Simulation settings

In order to evaluate the improvements on STE accuracy brought by LES of the adjoint equation, two inferences based on RANS and LES of the adjoint equation are conducted and compared. One of the databases in **Chapter 4** for which the data volume was compressed 10 times to ensure the error $\frac{\max|f-\tilde{f}|}{\max|f|} < 10^{-5}$ is selected. The original volume of 240T data was ~24 TB. The compression reduced this size significantly to ~2.4 TB, which made it much easier to handle, even with a portable

hard disk.

The RANS simulation of adjoint equations followed Eqs. (2.24, 5.1-5.3) with a global turbulent Schmidt number $Sc_t = 0.7$. All time-averaged values were obtained from the $240T$ original LES. The simulated ensemble-averaged concentration of the adjoint tracer was regarded as the modeling concentration $\boldsymbol{R}$. The unsteady adjoint concentration distribution was predicted by LES according to Eq. (5.4-5.5), where $\widetilde{\boldsymbol{u}}$ and D_{sgs} were the instantaneous values of the compressed LES flow field at each time step and Sc_{sgs} was set as 0.7 for the entire domain. Of the total $240T$ of compressed flow field data, the first $120T$ were used to initialize the simulation of the adjoint equation. The time-averaged C^* distribution of the second $120T$ was then sampled as the modeling concentration $\boldsymbol{R}$.

5.3.3 Bayesian inference settings

Bayesian inference was implemented twice based on the different modeling concentrations, $\boldsymbol{R}$, obtained from the RANS and LES models of adjoint equations. The measurements, $\boldsymbol{D}$, were collected in the wind tunnel experiment. The variance of errors is set as $\sigma_{d,i}^2 + \sigma_{m,i}^2 = r * D_i$, where $r = 0.2$. Other settings are the same as inference introduced in **Chapter 2**.

5.4 Results and discussions

Here, all the results were nondimensionalized. The coordinate and length values were nondimensionalized by H. The velocity results were nondimensionalized by U_r. The concentration results were nondimensionalized by the standard concentration $C_r = C_{\text{gas}} Q/(U_r h_r^2)$. Here, C_{gas} is the concentration of gas in the injected flow from the source in the wind tunnel experiment, Q is the injected flow rate, and h_r is the reference height, which was set as $3.33H$.

5.4.1 Flow fields of forward simulations

In the LES flow fields of the forward simulations, the velocity results were nondimensionalized by the reference velocity, U_r. The mean velocity fields in several planes are shown in **Fig. 5.2**. The horizontal velocity distribution shows the symmetry caused by the regular block-arrayed

configuration. The most dominant velocity is the streamwise velocity in open street areas and the separation between the windward sides of buildings and circular vortices in wake regions can be distinguished through the spanwise velocity distribution. Apart from this, the spanwise velocity in most areas is nearly zero. In the vertical plane of open street regions ($y = 6H$, **Fig. 5.2(c, d)**), the mean velocity is layered such that the streamwise velocity at the same height does not change with the x-coordinate and the vertical velocity is nearly zero. There is essentially no momentum exchange in the vertical direction. This averaged flow field was used in the RANS simulation of adjoint equations for all sensors as the conventional method.

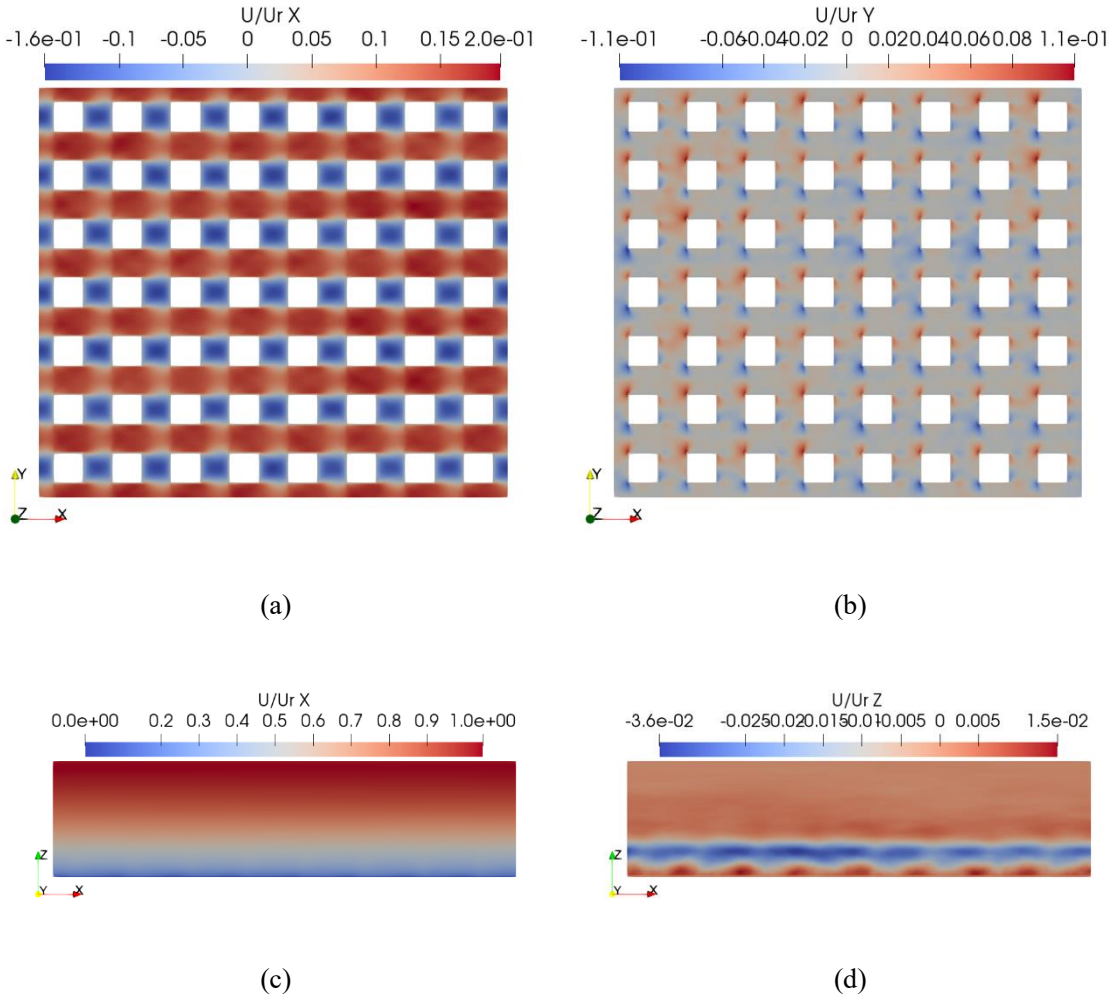


Figure 5.2. Mean flow fields produced by the forward LES. (a) streamwise velocity in a horizontal plane; (b) spanwise velocity in a horizontal plane; (c) streamwise velocity in a vertical plane; (d) vertical velocity in a vertical plane. Horizontal: $z = H/20$; vertical: $y = 6H$.

The validation profiles of the streamwise velocity can be checked in **Fig. 4.4 & 4.5**. Since the compression error was very limited in this study, it is reasonable to assume that the velocity field in

the LES of adjoint equations does not contribute greatly to the overall modeling error.

5.4.2 Comparison of adjoint concentration

The adjoint concentration fields simulated by RANS and LES models are compared. **Fig. 5.3** shows the mean distribution of the adjoint tracer released from sensor No. 1, which was located in the open street region between buildings ($x = 4H$, $y = 6H$, $z = 0.5H$). The horizontal plane is shown as the lowest plane in the domain, which contains the true source and can demonstrate its adjoint concentration. The vertical plane contains the sensor.

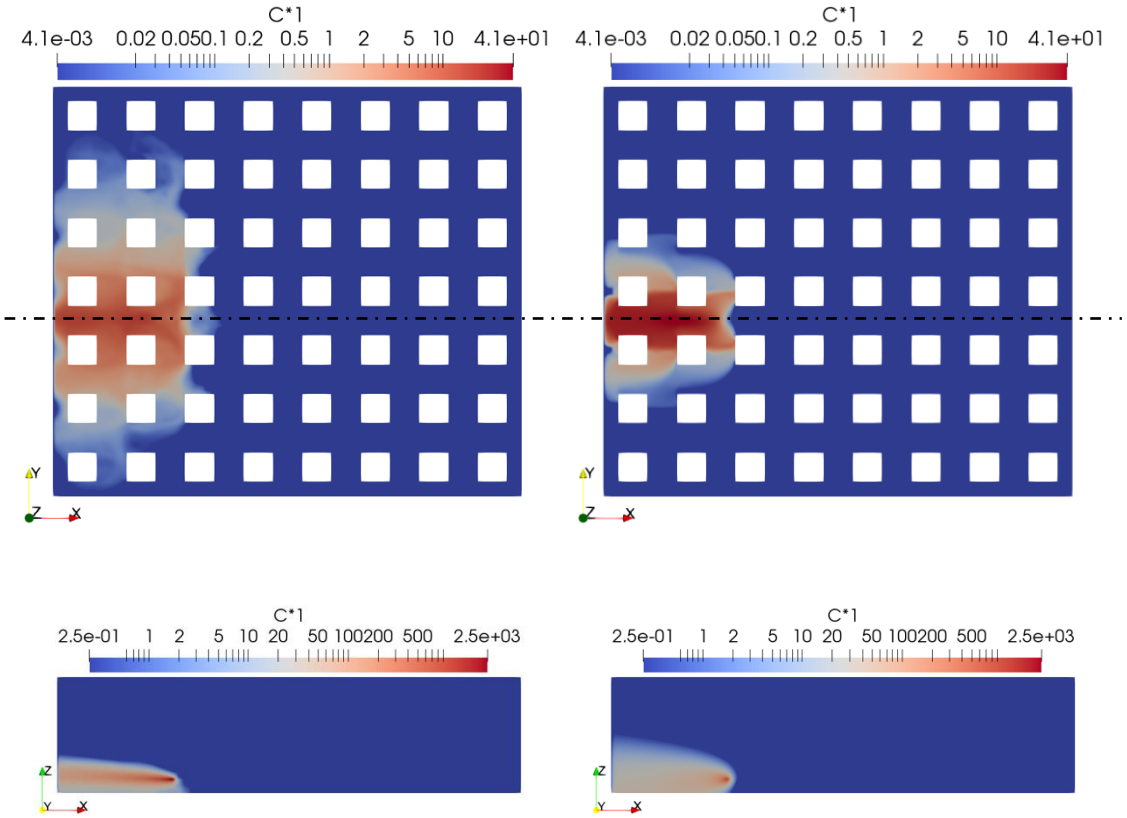


Figure 5.3. Mean adjoint concentration field of sensor No. 1 located in the open street region simulated by different methods: (left) LES and (right) RANS. (Top) Distribution in the horizontal plane ($z = H/20$); (bottom) distribution in the vertical plane ($y = 6H$).

In the RANS simulation of adjoint equations, it was confirmed that the dispersion area in the horizontal plane was limited within $5H$ in width, which is less than that of the LES results. Most of the adjoint tracers were clustered along the open street where the sensor is located. The GDH in the RANS model makes the turbulent diffusion proportional to the gradient of the mean concentration,

which is only true for small turbulent structures that can be simplified as isotropic. However, according to the mean flow shown in **Fig. 5.2**, the advection of the tracers was dominated by the streamwise velocity and separation flow, which constructed large-scale flow structures stretched in the streamwise direction. The strong anisotropy of the flow fields invalidated the efficiency of the GDH, making the turbulent diffusion also follow this large structure and causing simulation errors in the adjoint equations.

As most tracers are transported along the streamwise direction, dispersion in the spanwise direction is very weak. Even though the separation flow caused by the front walls of blocks can suck tracers into the wake region in the reverse flow, its strength is limited when compared with the streamwise velocity. The asymmetry of wake flow structures prevents the dispersion from expanding into the other half of the wake. Moreover, the mean velocity field is layered in the vertical direction, so the transport of the adjoint tracer in this direction can only rely on the GDH and molecular diffusion. Hence, in the vertical plane, the concentration distribution of the RANS model demonstrated a higher level of isotropy than the LES, where the distribution was stretched more in the streamwise direction and squeezed downward. Consequently, fewer tracers were moved to the bottom boundary in the RANS simulation. When the source was far away from a sensor in the spanwise direction or they were at different heights, it would not be notably impacted from the tracer and the adjoint relationship would be incorrectly constructed. In contrast, the situation changed in the LES of adjoint equations. Because most of the turbulent diffusion was explicitly resolved, the dispersion of adjoint tracers was wide throughout the domain in both the horizontal and vertical directions. Instantaneous turbulent flows carried the tracers emitted from the sensor across several wake regions.

Fig. 5.4 shows the mean distribution of the adjoint tracer released from sensor No. 4 in the wake region behind a building ($x = 6H$, $y = 5H$, $z = 0.5H$). In the RANS model, compared with the sensor in the open street area, the dispersion area widened and the concentration in the lowest horizontal plane increased. The time-averaged flow structures in the wake region were more effective for dispersion than the layered flow in the open street region. Nevertheless, the dispersion was not able to expand across two wake regions in the spanwise direction; instead, it clustered around the blocks in the same row as the sensor. Thus, the prevention of dispersion caused by asymmetric wake flows remained. As before, in the LES, the adjoint tracers were transported over a larger area by the explicit turbulent diffusion.

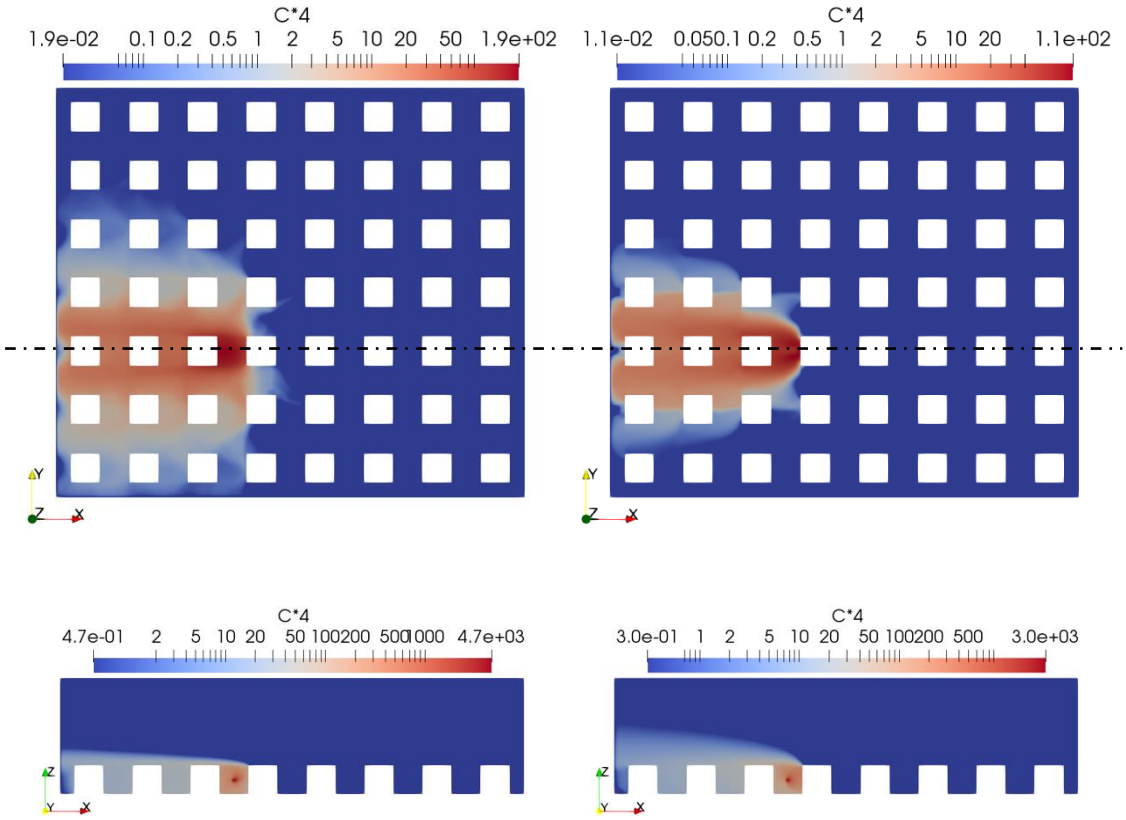


Figure 5.4. Mean adjoint concentration field of sensor No. 4 located behind a building simulated using different methods: (left) LES and (right) RANS. (Top) Distribution in the horizontal plane ($z = H/20$); (bottom) distribution in the vertical plane ($y = 5H$).

The measurements, $\boldsymbol{D}$, from the wind tunnel experiment are compared with the modeling concentration, $\boldsymbol{R}$, of the RANS and LES models in **Fig. 5.5**. According to the adjoint relationship shown in Eq. (2.23), the $\boldsymbol{R}$ of each sensor in **Fig. 5.5** is the nondimensionalized concentration of adjoint tracers emitted by that sensor at the true source's location. We confirmed that the LES modeling concentration was closer to the experimental data than that in the RANS model because most of the adjoint tracers were transported along the streamwise direction and the spanwise diffusion was insufficient in the latter model. For the sensors placed at large distances from the true source in the y-direction, the RANS model underestimated the concentration measurements. Meanwhile, for sensors placed in the same row as the true source, concentrations were overestimated due to the over-concentration of the adjoint tracers. Therefore, regardless of whether the sensor was in an open street or wake region, simulating turbulent diffusion is critical to dispersion.

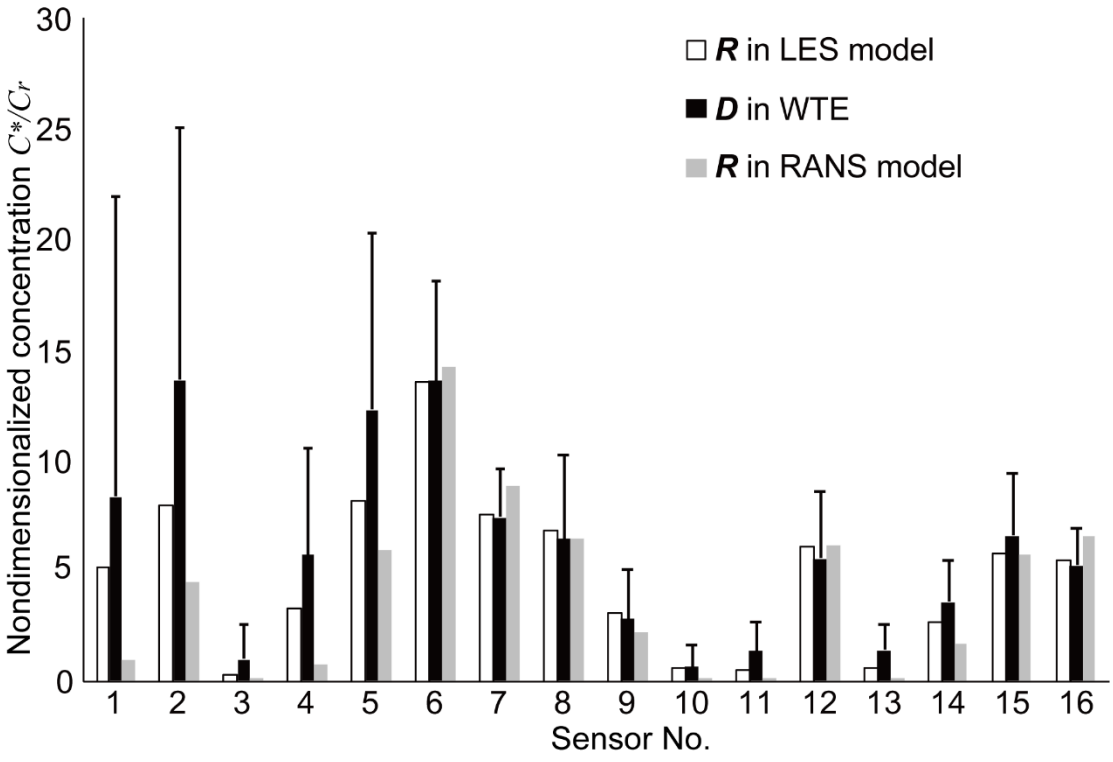


Figure 5.5. Comparison between the measurements, $\boldsymbol{D}$, in the wind tunnel experiment (WTE) and modeling concentration, $\boldsymbol{R}$, of the RANS and LES models. Error bars denote the standard deviation of $\boldsymbol{D}$ in the WTE.

The numerical model employed to simulate adjoint equations requires special attention for complex urban areas, as in the case study. In general, the LES performed better than the RANS model for most sensors even though there were also discrepancies when compared to the experimental data. One possible reason for this result could be due to the fact that the RANS Sc_t was still set as a global constant and ν_t was modeled under the assumption of isotropic turbulence in the current case. Thus, deeper research is still needed to advance the modeling of Sc_t and ν_t.

5.4.3 Source term estimation results

We compared the estimation results of Bayesian inference using the modeling concentrations, $\boldsymbol{R}$, of the RANS and LES models of adjoint equations. **Fig. 5.6** presents the marginal probability distribution of each source parameter. In the results of the proposed method based on the LES of adjoint equations, the estimation of each parameter was accurate. The peak values of the *x*- and *y*-coordinates were close to the true ones and the deviations were smaller than 0.5H. The strength of the source was also well-defined since the peak value was almost the same as the true one, and the

probability density function (PDF) was concentrated around the true value, similar to a normal distribution with a small standard deviation.

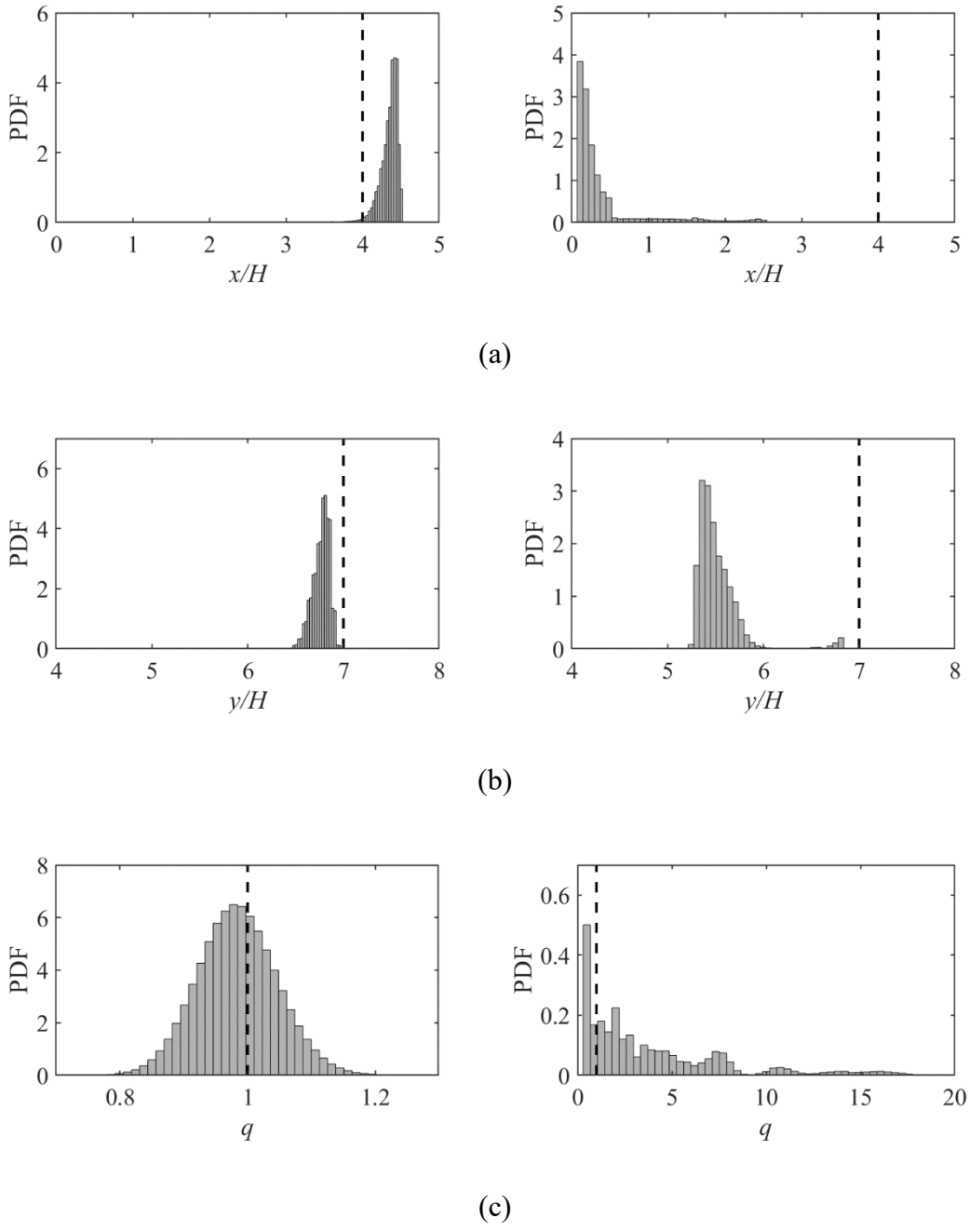


Figure 5.6. Marginal probability distributions of single-source parameters (a: *x*-coordinate; b: *y*-coordinate; c: strength) estimated by Bayesian inference using different adjoint concentration fields: (left) LES and (right) RANS. The dashed line represents the true value.

The estimation performance with RANS modeling of the adjoint equations was worse than that of the LES. The PDFs of the *x*- and *y*-coordinates greatly deviated from true values. The zero

probability of estimating the true value indicates that the Markov chain did not explore this area or did not credit it after exploration; both situations mean that the probability of the true value was too low because the difference was too great between the modeling concentration and measured value. The estimated PDF of the strength revealed even more limitations of the conventional method. In spite of the fact that the peak value was close to the true strength, the distribution was so wide that it rendered the estimate invalid. There was also a second peak in the distribution far from the true value. In a real-world application, with more complicated measurement noise, it is possible that the width of the distribution would increase and the wrong peak would be identified as the dominant one, which may hinder subsequent risk management.

The main reason for the estimation failure of the RANS model was attributed to the inaccurate prediction of adjoint concentration fields, where the turbulent diffusion was modeled according to the GDH. The difference between the measurements and modeling concentrations made it difficult to identify the true parameters via Bayesian inference. It is worth noting that, although the RANS modeling concentration errors may seem tolerable when compared with the variance within the measured values, the resulting estimation error is considerably larger. The reason for this is that in sampling the posterior distribution, the modeling concentration is coupled with the inference input coefficient, $\sigma_{d,i}^2 + \sigma_{m,i}^2$. As the optimal selection of this variance remains unknown, this amplification could be inefficient in practical applications. For modeling concentrations that are not accurate enough, the inference is vulnerable to the selection of the input coefficient. When confronting complex measurements in reality, establishing an accurate modeling concentration is key to effective STE before systematic research into the selection of the input coefficient is completed.

The improvements made to the accuracy of the STE yielded by the proposed model is better demonstrated by the joint probability distribution, $p((x, y)|D, I)$, of the two methods (**Fig. 5.7**). We confirmed that the estimated source locations based on the LES model were very close to that of the true source. The inference narrowed the probability into the same wake region and the distance between the estimated location and the true one was smaller than $0.5H$. In contrast, the estimation based on the RANS model was inaccurate. The joint probability clustered around a small area that was approximately two columns away from the true location, which suggests the limited credibility of the conventional RANS-based method.

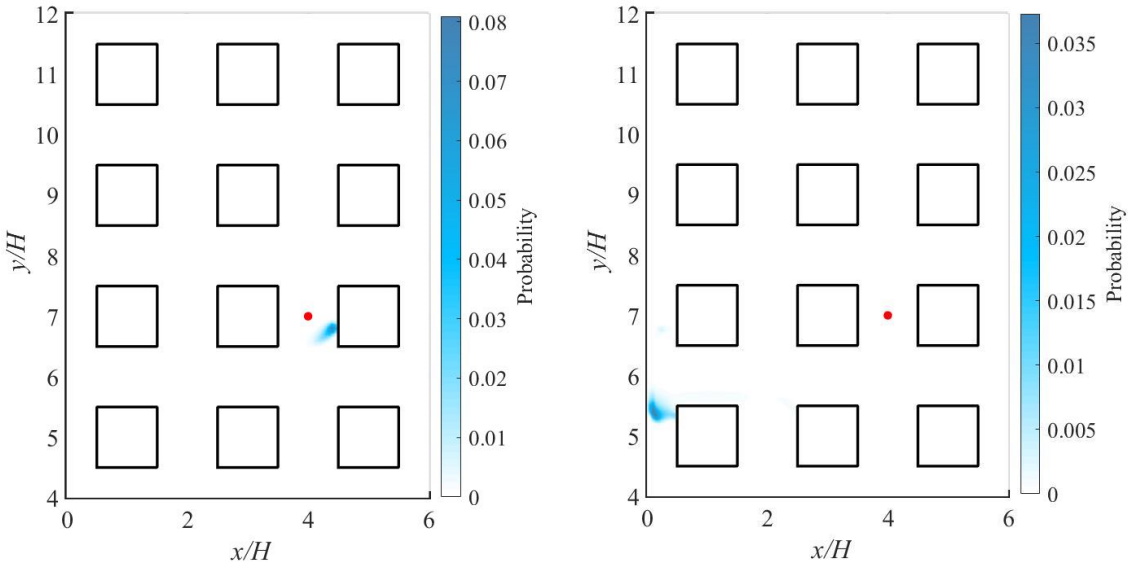


Figure 5.7. Joint probability distribution of source locations estimated by Bayesian inference using different adjoint concentration fields: (left) LES and (right) RANS. The red point represents the true source.

To quantify the improvement to STE achieved by the LES modeling of adjoint equations, we selected the 50th percentile values of the PDFs shown in **Fig. 5.6** as the estimated results of the two methods and defined the following two indices.

The location error E_d is defined as the expectation of the Euclidian distance between the estimated location and the true location, which was calculated using the following equation:

$$E_d = \int\int p(x, y|\boldsymbol{D}, \boldsymbol{I})\sqrt{(x - x_s)^2 + (y - y_s)^2}dxdy \tag{5.6}$$

where (x_s, y_s) is the true location of the source.

The strength error E_q is the expectation of the normalized difference between the estimated strength and true strength, which is expressed as follows:

$$E_q = \int p(q|\boldsymbol{D}, \boldsymbol{I})\frac{|q - q_s|}{q_s}\,dq \tag{5.7}$$

where q_s is the true strength of the source.

These indices are summarized in **Table 5.1**. We found that the LES model reduced the errors of location and strength estimates by 89% and 99%, respectively, when compared with the RANS model.

Table 5.1. Summarized estimation results of the RANS and LES models.

Method	Location		Strength	
True value	(x_s, y_s)	E_d	q_s	E_q
	(4H, 7H)		1	
Steady adjoint equation (RANS)	(0.21H, 5.46H)	4.09H	2.5	1.5
Unsteady adjoint equation (LES)	(4.38H, 6.78H)	0.44H	0.98	-0.02

5.5 Conclusions

In this chapter, to improve the accuracy of STE in complex urban applications, a new method was developed based on Bayesian inference coupled with unsteady adjoint equation modeling via LES. The performance of the proposed method was evaluated through a regular, block-arrayed urban model with the continuous dispersion of a point source in the wind tunnel experiment. The LES approach was applied to predict the flow fields of the urban model and the full spatial distribution and time-dependent dynamics were saved for the simulation of adjoint equations. To relieve the pressure imposed by storing such a large amount of data, the wavelet-based compression method was employed to compress the original data to 10% of its original volume. The source-receptor relationship obtained from the LES model of the adjoint equations and the measurements from 16 sensors in the wind tunnel experiment were used as inputs for Bayesian inference to calculate the posterior distribution of the source term. To clarify the improvements made to the accuracy of STE using the proposed method, an existing RANS-based method of simulating adjoint equations with a time-averaged LES flow field was conducted for comparison.

The results showed that the modeling of the adjoint equation was significantly improved by the new LES model. The proposed method reflected the crucial effects of an explicit resolution of turbulent diffusion with time-series data. The modeling concentration, $\boldsymbol{R}$, of the LES model was closer to the measurements, $\boldsymbol{D}$, than that of the RANS model. Since turbulent diffusion was modeled by the GDH in the RANS method, the adjoint concentration predicted with the mean flow field did not reflect the true source-receptor relationship well. The different degrees of accuracy of the modeling concentrations between the two methods further resulted in different estimated

posterior probabilities of the source parameters. For the RANS model, neither the coordinates nor the strength of the point source was successfully identified. In contrast, the LES model provided much more accurate estimates, which were robust to variations in noise, and the estimated location was in the same wake region of the true source. The errors in estimates of location and strength were reduced by 89% and 99%, respectively, using the LES model of adjoint equations when compared with the RANS approach.

It must be noted that the regular, block-arrayed model employed in this chapter is an idealized scenario of real urban geometry. Even in this simple case, the difference between the adjoint tracer concentration predicted by the RANS and LES models was large enough to significantly affect the STE results. The critical factor was the simulation of dispersion in the adjoint equations. Even though the RANS model imposed lower computational costs, it was demonstrated that the GDH could not satisfy the accuracy requirements in strongly anisotropic flow fields. In real urban areas, the situation is much more complicated because of the diverse geometries and configurations of buildings and equipment. Hence, to effectively perform STE, it is beneficial to spend more time explicitly resolving turbulent diffusion for the source–receptor relationship via LES until more accurate numerical models for turbulent diffusion are proposed.

Even though unsteady LES of adjoint equation was realized in the current research, the time-averaged adjoint relationship was used in the inference instead of time-series data. The reason is that only time-averaged measured concentration data is available in the case study. Besides, advanced measurement equipment that can record concentration in a time-series form is still unpopular due to high cost. Existing research in the literature is still mainly based on mean concentration measurement. Hence, for the convenience of comparison, this research also used time-averaged simulated results. Although it seems to be a waste of calculation resources to only use mean value of unsteady simulation, it was confirmed that mean results of the unsteady simulation outperform that of the steady simulation.

However, it must be admitted that time-series data contains much more useful information to STE than time-averaged data. The performance of STE will be improved if the time-series data can be effectively utilized, which is also helpful to estimations considering multiple source or response time. Therefore, it is necessary to develop appropriate ways to fully use unsteady adjoint relationships.

1 Symbols

C : the concentration field

C^* : the adjoint concentration field

C_r : standard (reference) concentration

C_{gas} : the emission strength of the source

$\boldsymbol{D}$: the measurements vector

D_i : the measurement of the sensor with index i

D_m : molecular diffusivity

D_{sgs} : sub-grid scaled turbulent diffusivity

D_t : turbulent diffusivity

E_d : the expectation of the Euclidian distance between the estimated location and the true location

E_q : the expectation of the normalized difference between the estimated strength and true strength

f : the original data of a three-dimensional scalar field in the Cartesian indexing mesh grid

$\check{f}$: the decompressed flow field data

H : the edgy length of blocks in the simulation (=60 mm)

h_r : reference height

I : the background information for Bayesian inference

$p(A|B)$: conditional probability of event A occurring given that B is true

Q : the gas flow rate at the source

q : strength samplings produced in MCMC for the point source

q_s : the true strength of the point source

r : the ratio between the specified error covariance in the Bayesian inference and the measurements

S_{ij} : mean strain rate of i and j directions

Sc_{sgs} : sub-grid scaled turbulent Schmidt number

Sc_t : turbulent Schmidt number

T : standard time-scale of LES simulation

U_r : reference speed

$\boldsymbol{u}$: velocity field

$\boldsymbol{x}_m$: coordinates of sensors

(x, y) : location samplings produced in MCMC for the point source

(x_s, y_s) : the true location of the point source

$\delta(\cdot)$: Dirac delta function

ν_{sgs} : eddy viscosity coefficient at the sub-grid scale

ν_t : eddy viscosity

$\sigma_{d,i}^2$: the variance of error in the measurement of the sensor with index i

$\sigma_{m,i}^2$: the variance of error in the modeling concentration for the sensor with index i

$\sigma(\cdot)$: standard deviation

$(\cdot)'$: fluctuation value with time

$\bar{}$: Reynolds average operator

$\tilde{}$: filtering operator in LES

1

2

Chapter 6
Sensor configuration optimization for source term estimation based on the entropy of adjoint equation

Abstract

Until now, few studies have developed sensor configuration optimization methods aiming to ensure good source term estimation (STE) performance in the monitoring area, such that the posterior probability could aggregate around the truths while addressing most sources. This chapter proposes a method by designing an objective function and applying a simulated annealing algorithm. The objective function is set as the information joint entropy of the adjoint concentration. The performance of the proposed method was assessed by Bayesian inference STE for 25 unknown sources based on the obtained optimal configuration in a regular block-arrayed building group model. The STE results were compared with those of uniform and random configurations.

6.1 Introduction

Source term estimation (STE) has three critical factors: measurements, source-receptor relationship, and estimating algorithm. In the previous research, most efforts have been devoted to the source-receptor relationship modeling and estimating algorithms. Many methods have been proposed to accurately simulate the predicted concentration and accelerate the estimation algorithm, while measurements have been neglected for a long time, and the related research is sparse. However, measurements are the foundation of STE. Theoretically, all STE methods would be useless if sensors could not efficiently measure the concentration information (Keats et al., 2010). In **Chapter 3**, it has been confirmed that some configurations may not provide sufficient measurement information for the STE of the line source, and the resultant estimations are inaccurate. In addition, the uniform and random sensor configurations used in wind tunnel experiments or field tests in previous research are impractical because of the irregular building distribution and the limitation of deployment cost in real applications. Random configuration may also cause considerable errors in the measurements of certain sources, which may be dominant over the truth.

Most existing research on sensor configuration optimization (SCO) design is still insufficient because of the special requirements of STE applications. Several studies used the difference between measurements and simulated concentrations to evaluate the sensor configuration. For example, a sensor network has been designed to monitor nuclear power stations in France (Abida et al., 2008; Saunier et al., 2009). Kouichi et al., (2019) proposed an optimization method to reduce the existing sensors for an urban model area. However, the simulated concentration in these methods needs prior knowledge of sources in advance, which is unavailable in STE applications. Other published SCO methods aim to monitor specific atmospheric pollutants such as ozone, rather than a specific source (Araki et al., 2015; Fuentes et al., 2007; Wu and Bocquet, 2011). Sources emitting ozone are too numerous to effectively identify, so their target is to estimate the global distribution using several point measurements. Keats et al. (2010) proposed a design method related to STE, but the research problem is to place an additional sensor as a supplementary for a fixed configuration after the dispersion emergency. The method cannot be used in the preparatory sensor configuration design before the emergency. Until now, one of the most applicable SCO methods for STE was proposed by Ngae et al. (2019), in which the sensor configuration is evaluated based on the entropic criterion of measurement information provided by sensors (Issartel, 2005b) without the prior knowledge of sources. However, this method has a deep relationship with the renormalization inversion theory, which belongs to the deterministic STE category. It is still necessary to develop other SCO methods

from the view of stochastic STE theory.

Therefore, in this chapter, an SCO method for the stochastic STE is proposed. This method has the same basic idea as Ngae et al. (2019) that evaluates the quality of sensor configuration based on the information entropy contained in sensors' measurements. Despite that, these two methods use totally different ways to calculate the information entropy. Ngae et al. (2019) obtained the entropy by the determinant of the weighted Gram matrix (Issartel, 2005b), which is an important element transformed from adjoint concentration field in the renormalization inversion theory, while our method calculated the entropy with the spatial distribution of adjoint concentration, which is expected to be more intuitive and easy to understand. Besides, because entropy is an important concept in stochastic theory, its physical meaning is closely related to stochastic STE and is interpreted in this chapter.

The application scenario is assumed to be that a point source with constant release strength may appear anywhere in the target domain with a statistically steady flow field. To monitor the target area, the proposed method finds a user-specified number of sensors among candidates to construct an optimal configuration. Regarding stochastic STE, the Bayesian inference method in **Chapter 2** is applied. In an accurate Bayesian STE, the probability mass in the posterior probability distribution should concentrate around the true value with a narrow distribution width. The optimal configuration and resultant estimations are expected to be superior to others such as uniform, random, and experience-based configurations.

6.2 Entropy-based configuration optimization

In this section, the proposed SCO method is introduced. It consists of two parts: the objective function to evaluate the rank of any configuration and the simulated annealing (SA) algorithm to efficiently identify the optimal configuration.

6.2.1 Objective function

The objective function is used to evaluate the measurement ability and steadiness of each sensor configuration. It is proposed that all configurations can be ranked by the joint entropy of the spatial probability distribution of adjoint concentration fields of their sensors. The optimal

configuration should have the largest entropy. The author first proves this mathematically and then explains its physical meaning.

a. Mathematical explanation

Here is the definition of the information entropy of a probability distribution, $p(x)$ (Cover and Thomas, 2006).

$$H(x) = -\int p(x)\, log\, p(x)\, d\boldsymbol{x} \tag{6.1}$$

In the statistics, entropy represents the information or uncertainties of $p(x)$. A distribution with a narrow spread and a sharp peak possesses small entropy, meaning that the uncertainty of the objective or information contained in the distribution is small. Similarly, a distribution with a wide spread and slight slope has a large entropy. Naturally, a uniform distribution has the largest entropy, information, and uncertainty (Fuentes et al., 2007).

For any unknown source $\boldsymbol{s}$, if we only have background information I without any sensors, the prior probability distribution of the unknown source parameter is $p(\mathbf{s}|I)$. When we have n sensors to monitor an area with the measurement vector $\boldsymbol{D} = (D_1, D_2, D_3, \dots, D_n)$, after the Bayesian inference, the posterior probability distribution of the unknown source parameter becomes $p(\mathbf{s}|\boldsymbol{D}, I)$. The information about unknown sources provided by these sensors can be quantified by mutual information I_m, which is defined as

$$I_m(\boldsymbol{s}; \boldsymbol{D}|I) = H(\boldsymbol{s}|I) - H(\boldsymbol{s}|\boldsymbol{D}, I) \tag{6.2}$$

where $H(\mathbf{s}|I) = -\int p(\mathbf{s}|I) \log p(\mathbf{s}|I)\, d\boldsymbol{s}$ and $H(\mathbf{s}|\boldsymbol{D}, I) = -\int p(\mathbf{s}|\boldsymbol{D}, I) \log p(\mathbf{s}|\boldsymbol{D}, I)\, d\boldsymbol{s}$ are the entropies of the corresponding probability distributions. The optimum sensor configuration should provide the largest information $I_m(\boldsymbol{s}; \boldsymbol{D}|I)$, which means it should minimize $H(\boldsymbol{s}|\boldsymbol{D}, I)$ because $H(\boldsymbol{s}|I)$ is constant before and after measurement. According to the properties of the entropy definition, $H(\boldsymbol{s}|\boldsymbol{D}, I)$ becomes smaller when $p(\mathbf{s}|\boldsymbol{D}, I)$ has a narrower distribution. Therefore, the minimum $H(\boldsymbol{s}|\boldsymbol{D}, I)$ denotes the posterior probability density function (PDF) of unknown source $p(\mathbf{s}|\boldsymbol{D}, I)$ concentrates around a certain value, which is expected to be close to the true value when the Bayesian inference is sufficiently accurate.

To find $\boldsymbol{D}$, which minimizes $H(\boldsymbol{s}|\boldsymbol{D}, I)$, let's start with a simple scenario. Imagine the situation that we determine (randomly or by experience) first as $n-1$ sensors, fix them, and find the

optimal position for the next sensor e. The measurement vector can be divided into $\boldsymbol{D} = (\boldsymbol{D}_{n-1}, D_n)$. The probability that sensor e can provide the measurement D_n is denoted as

$$\begin{aligned} p(D_n|\boldsymbol{D}_{n-1}, I) &= \int p(D_n|\boldsymbol{s}, \boldsymbol{D}_{n-1}, I) p(\boldsymbol{s}|\boldsymbol{D}_{n-1}, I) d\boldsymbol{s} \\ &= \int p(D_n|\boldsymbol{s}, I) p(\boldsymbol{s}|\boldsymbol{D}_{n-1}, I) d\boldsymbol{s} \end{aligned} \tag{6.3}$$

When we assume that the measurement of each sensor is only determined by source $\boldsymbol{s}$ and would not be affected by other sensors, $\boldsymbol{D}_{n-1}$ can be removed from $p(D_n|\boldsymbol{s}, \boldsymbol{D}_{n-1}, I)$ on the right-hand side (RHS). This probability contains the real measurement error and the posterior guess of the source using the known data $\boldsymbol{D}_{n-1}$. The optimal selection of e should minimize the entropy $H(\boldsymbol{s}|\boldsymbol{D}, I)$ as follows:

$$H(\boldsymbol{s}|\boldsymbol{D}, I) = H(\boldsymbol{s}|\boldsymbol{D}_{n-1}, D_n, I) = -\int p(\mathbf{s}|\boldsymbol{D}_{n-1}, D_n, I) \log p(\mathbf{s}|\boldsymbol{D}_{n-1}, D_n, I)\, d\boldsymbol{s} \tag{6.4}$$

Because D_n also has uncertainty, as shown in Eq. (6.3), the expectation of $H(\boldsymbol{s}|\boldsymbol{D}, I)$ can be evaluated by

$$E[H(\boldsymbol{s}|\boldsymbol{D}_{n-1}, D_n, I)] = \int p(D_n|\boldsymbol{D}_{n-1}, I) H(\boldsymbol{s}|\boldsymbol{D}_{n-1}, D_n, I) dD_n \tag{6.5}$$

The joint entropy has a chain rule (Cover and Thomas, 2006):

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z) = H(X|Z) + \int p(x) H(Y|x, Z) dx \tag{6.6}$$

If we let $X = \boldsymbol{s}$, $Y = D_n$, $Z = (\boldsymbol{D}_{n-1}, I)$ and notice that neglecting Z in the equation for a moment for simplicity would not change the final result because it appears in each term, we have:

$$H(\boldsymbol{s}, D_n) = H(\boldsymbol{s}) + \int p(\boldsymbol{s}) H(D_n|\boldsymbol{s}) d\boldsymbol{s} \tag{6.7}$$

By switching the position of $\boldsymbol{s}$ and D_n, we can further obtain:

$$H(\boldsymbol{s}, D_n) = H(D_n) + \int p(D_n) H(\boldsymbol{s}|D_n) dD_n \tag{6.8}$$

The second term on the RHS of Eq. (6.8) is the same as the RHS in Eq. (6.5), the expectation of $H(\boldsymbol{s}|\boldsymbol{D}, I)$, and RHS of Eq. (6.8) also equals the RHS of Eq. (6.7). Then, we return $Z = (\boldsymbol{D}_{n-1}, I)$ to obtain

$$\begin{aligned} E[H(\boldsymbol{s}|\boldsymbol{D}_{n-1}, D_n, I)] &= \int p(D_n|\boldsymbol{D}_{n-1}, I) H(\boldsymbol{s}|\boldsymbol{D}_{n-1}, D_n, I) dD_n \\ &= H(\boldsymbol{s}|\boldsymbol{D}_{n-1}, I) + \int p(\boldsymbol{s}|\boldsymbol{D}_{n-1}, I) H(D_n|\boldsymbol{s}, \boldsymbol{D}_{n-1}, I) d\boldsymbol{s} \\ &- H(D_n|\boldsymbol{D}_{n-1}, I) \end{aligned} \tag{6.9}$$

In the RHS of Eq. (6.9), the first term is the entropy of the posterior PDF of the STE by $\boldsymbol{D}_{n-1}$. Because these sensors are fixed, the value is constant and independent of D_n. The second term describes the influence of measurement noise on the data. $H(D_n|\boldsymbol{s}, \boldsymbol{D}_{n-1}, I)$ means the uncertainty contained in the measurement signals when a source $\boldsymbol{s}$ and a No. *n* sensor are selected. This uncertainty is usually affected only by the measurement noise since both source and sensor are fixed. If we assume that the noise is independent of the measurement location, the second term is also constant (Loredo, 2004). Therefore, to minimize the left-hand side of Eq. (6.9), the third term must be maximized.

Before further derivation, we first check the meaning of the third term $H(D_n|\boldsymbol{D}_{n-1}, I)$, which is the entropy of probability $p(D_n|\boldsymbol{D}_{n-1}, I)$. According to Eq. (6.3), $p(D_n|\boldsymbol{D}_{n-1}, I) = \int p(D_n|\boldsymbol{s}, I) p(\boldsymbol{s}|\boldsymbol{D}_{n-1}, I) d\boldsymbol{s}$. $p(\boldsymbol{s}|\boldsymbol{D}_{n-1}, I)$ indicates that measurements $\boldsymbol{D}_{n-1}$ constrain the estimation of an unknown source in a limited parameter space $\mathbb{V}$ with certainty. When the unknown source is located at different places in $\mathbb{V}$, it may cause different concentration measurements D_n at sensor *e*, which is the meaning of the first term $p(D_n|\boldsymbol{s}, I)$.

Assuming that the adjoint equation is accurately simulated, according to the adjoint relationship between the source and sensor, as shown in Eq. (2.24), we obtain:

$$D_n = \overline{C_{\boldsymbol{s}}(e)} = q_{\boldsymbol{s}} \overline{C_e^*(\boldsymbol{s})} \tag{6.10}$$

Therefore, $p(D_n|\boldsymbol{s}, I) = p\left(\overline{C_{\boldsymbol{s}}(e)}\middle|\boldsymbol{s}, I\right) = p\left(q_{\boldsymbol{s}} \overline{C_e^*(\boldsymbol{s})}\middle|\boldsymbol{s}, I\right)$.

To maximize $H(D_n|\boldsymbol{D}_{n-1}, I)$, $p(D_n|\boldsymbol{D}_{n-1}, I)$ should have as wide a distribution as possible; consequently, $p(D_n|\boldsymbol{s}, I)$ or $p\left(q_{\boldsymbol{s}} \overline{C_e^*(\boldsymbol{s})}\middle|\boldsymbol{s}, I\right)$ should be wide, since $p(\boldsymbol{s}|\boldsymbol{D}_{n-1}, I)$ has been fixed in Eq. (6.3). In the Bayesian inference for STE, $q_{\boldsymbol{s}}$ is sampled by the MCMC with a uniform prior distribution. Meanwhile, in the process of location optimization for sensor *e*, changing the location of *e* would only affect $\overline{C_e^*(\boldsymbol{s})}$, the spatial distribution of adjoint concentration corresponding to $\mathbb{V}$. The wider this spatial distribution is, the wider $p\left(q_{\boldsymbol{s}} \overline{C_e^*(\boldsymbol{s})}\middle|\boldsymbol{s}, I\right)$ becomes. In other words, the uncertainty that the unknown source causes different measurements in sensor *e* can be transferred to the uncertainty of the spatial adjoint concentration resulting from *e*.

Returning to the optimal process, we use the chain rule of joint entropy (Eq. (6.6)) once more to obtain

$$H(\boldsymbol{D}|I) = H(\boldsymbol{D}_{n-1}, D_n|I) = H(\boldsymbol{D}_{n-1}|I) + H(D_n|\boldsymbol{D}_{n-1}, I) \tag{6.11}$$

Because the first term on the RHS of Eq. (6.11) is fixed and our target is to maximize the second term in the RHS, $H(\boldsymbol{D}|I)$ must be maximized. It is the joint entropy of probability of all measurements at fixed sensors (if e has been decided) caused by different sources $\boldsymbol{s}$. According to the discussion on the adjoint relationship above, this joint probability can be transferred to the joint spatial probability of adjoint concentrations at different places $p(\boldsymbol{R}|I)$ resulting from each sensor. Therefore, we find a sensor location e that makes the spatial distribution $p(\boldsymbol{R}|I)$ as wide as possible, and the corresponding entropy $H(\boldsymbol{R}|I)$ or $H(\boldsymbol{D}|I)$ is maximized.

In the above content, we proved the theorem that when $n-1$ sensors are fixed, the best location of the remaining sensor should be the one that makes the entropy $H(\boldsymbol{R}|I)$ larger than the other places. Now we can say that the configuration which owns the maximum $H(\boldsymbol{R}|I)$ is the optimized one. If it is not, we have to adjust the sensors' locations to improve it further. However, for each sensor, no matter where we move it, $H(\boldsymbol{R}|I)$ will decrease since the original one has the maximum value. According to the above theorem, the new configuration is worse than the original one. There is no more space for further improvement. Therefore, it is reasonable to set our objective function as $H(\boldsymbol{R}|I)$ to evaluate the efficiency of the sensor configuration. The optimal configuration should have the largest $H(\boldsymbol{R}|I)$ values. In the designing process, we simultaneously determine the locations for all sensors to maximize $H(\boldsymbol{R}|I)$.

b. Physical explanation

The reason the optimal configuration should have the largest $H(\boldsymbol{R}|I)$ can be intuitively explained by **Fig. 6.1**. An unknown source, the red star, appeared in the target domain. To facilitate understanding, it is assumed in this example that the strength of the source q_s is already limited in a small interval and the STE is operated in a two-dimensional space, which indicate that we only need to estimate its position in a plane here.

If we do not have any sensors, nothing is known. After we deployed one sensor, it can provide measurements D_1, and the corresponding adjoint concentration field R_1 can also be simulated, as shown in **Fig. 6.1**. According to the adjoint relationship in Eq. (2.24), it is reasonable to believe that

R_1 at the source should be close to D_1. The Bayesian inference evaluates the probability of each possible source at different locations using Eq. (2.10). based on the difference between the measurement $\boldsymbol{D}$ and the adjoint concentration $\boldsymbol{R}$. The probability is large when the difference is small. According to the adjoint concentration distribution of sensor No. 1 in **Fig. 6.1**, values of R_1 in the red potential area is similar to that of the true source because they have the same color bar. In this case, the posterior probability of the potential area is similar to that of the true source. The estimation result will show that the source may locate at anywhere in the potential area. Therefore, the estimation results based on one sensor are not accurate. It is still impossible to infer the true parameters of an unknown source.

Then, we can supplement another sensor to improve the estimation. Naturally, we can obtain a measurement vector $\boldsymbol{D} = (D_1, D_2)$, and each place has an adjoint concentration vector $\boldsymbol{R} = (R_1, R_2)$. Owing to the information provided by sensor No. 2, the potential area estimated by one sensor rapidly shrinks because the adjoint concentration R_2 of most parts does not agree with the R_2 of the source or measurement D_2. After removing these areas, the results of Bayesian inference become much better than before that the potential area tightly concentrated around the true source.

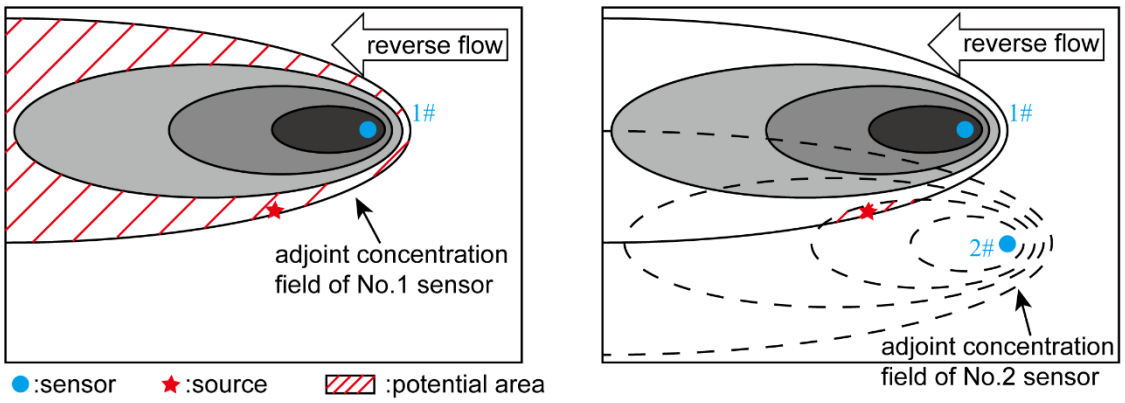


Figure 6.1. Diagram of the proposed objective function $H(\boldsymbol{R}|I)$. Left: Adjoint concentration field of sensor No. 1; Right: Adjoint concentration fields of sensor No. 1 & 2.

A statistical perspective highlights that two sensors collect much more information about the dispersion of the unknown source than one sensor. Hence, the resultant estimations contain considerably fewer uncertainties. The increase in the measured information can be represented by the entropy of the spatial probability distribution of the adjoint concentration in the target domain. When compared with one sensor, two sensors complicate the adjoint concentration field by adding

one dimension in $\boldsymbol{R}$, making the spatial probability distribution wider, and increasing the corresponding entropy. For this reason, there are significantly fewer areas owning a similar adjoint concentration vector with the true source, which can then be successfully identified.

If we want to re-deploy the second sensor, the location of sensor No. 1 should not be considered. Under such circumstances, the adjoint concentration field of sensor No. 2 is the same as that of sensor No. 1, which means that sensor No. 2 will fail to measure any new information regarding pollutant dispersion. Equally, the spatial probability distribution, entropy of the adjoint concentration field, and potential area remain unchanged. The estimation of the STE would not be improved. As a result, it is always desirable for the new sensor to increase the entropy of adjoint concentration; in other words, fewer points have the same adjoint concentration vector as the true source. Moreover, because the unknown source can appear anywhere in the target area, the ideal state is that each place can own a unique adjoint concentration vector in which the spatial probability distribution is close to a uniform distribution and the entropy reaches the maximum value. In other words, the physical meaning of the maximum entropy of spatial distribution of adjoint concentration is that the sensor configuration has unique relationship with all possible sources. When any source appears, the measured concentration vector and the adjoint concentration vector are special enough to eliminate all other options and make sure the STE can identify the true source. Hence, when the number of sensors is fixed and limited, we can find the sensor configuration with the largest entropy of adjoint concentration, which is expressed by the objective function.

6.2.2 Simulated Annealing

After determining the objective function to rank each sensor configuration, it is necessary to quickly finish the optimization process and determine the best configuration with the largest $H(\boldsymbol{R}|I)$. However, the optimization design for a large number of candidates is troublesome because of the heavy calculation burden. If we want to select j sensors out of n candidates to construct the network, the number of possible combinations is $C_n^j = \frac{n!}{(n-j)!j!}$. When j is 8 and n is 100, as a common example, the total number of combinations reaches over one thousand billion. It is impossible to directly calculate the objective function for each combination and select the optimum function. Therefore, this research applied the SA algorithm (Kirkpatrick et al., 1983), which is a heuristic method inspired by the process of cooling a liquid to the lowest possible energy state. The

detailed steps of the SA are as follows:

Initialization: Randomly select *j* sensors to form the first sensor network $\boldsymbol{m}_1$ and calculate its joint entropy of adjoint concentration $H(\boldsymbol{R}_1|I)$

For k = 2 to *N*:

Randomly change a sensor in $\boldsymbol{m}_{k-1}$ to generate a new network $\boldsymbol{m}_k$. Calculate $H(\boldsymbol{R}_k|I)$.

If $H(\boldsymbol{R}_k|I) > H(\boldsymbol{R}_{k-1}|I)$

Accepted

Else

If N [0,1]$< exp\left(-\frac{\frac{H(\boldsymbol{R}_{k-1}|I)-H(\boldsymbol{R}_k|I)}{H(\boldsymbol{R}_{k-1}|I)}}{T_k}\right)$

Accepted

Else

Rejected; $\boldsymbol{m}_k = \boldsymbol{m}_{k-1},\ H(\boldsymbol{R}_k|I) = H(\boldsymbol{R}_{k-1}|I)$

End if

End if

End For

In this algorithm, *N* is the specified loop number for the stop criterion. *N* [0,1] is a random variable with a uniform distribution between 0 and 1, which means that the new configuration has a certain possibility to be accepted even if its entropy is smaller. This setting enables the optimization process to jump out from the local maximum. The critical factor T_k is a virtual temperature defined by $T_k = T_0 \times a^k$, where $0 < a < 1$ is the cooling coefficient. In the beginning, when T_k is large, the probability of jumping to an inferior configuration is higher, and the algorithm is flexible for exploring the entire parameter space. As *k* increases, T_k tends to 0, making it almost impossible to move to an inferior configuration even if the difference is quite small.

T_k adjusts the balance between local and global exploration, thereby affecting the efficiency of the optimization process. Because T_k is closely related to the value of the objective function case by case, there is no general way to set the value of T_k. In this research, T_0 was imposed as 1 and a as 0.9, which is similar to previous research (Fuentes et al., 2007). To remove the influence of the initialization point, eight SA optimization chains were run in parallel with N = 5000. The sensor configuration with the largest entropy among the eight chains was regarded as the optimal configuration.

6.2.3 Calculation of $H(\boldsymbol{R}|I)$

During SA optimization, $H(\boldsymbol{R}|I)$ needs to be calculated for each configuration. It is the joint entropy of probability $p(\boldsymbol{R}|I)$, where $\boldsymbol{R} = (R_{i,j})$, $R_{i,j} = q_{s,i}\overline{C_j^*(\boldsymbol{x}_{s,t})}$, i is the index for all possible sources, and j is the index for sensors. In this definition, $\overline{C_j^*}$ is a three-dimensional adjoint concentration function of sensor No. *j*. Assume that there are *m* possible sources in total and we have *n* sensors to form a configuration, $\boldsymbol{R}$ can be expressed as:

$$\boldsymbol{R} = \begin{bmatrix} R_{1,1} & R_{1,1} & \cdots & & R_{1,n} \\ R_{2,1} & \ddots & & & \\ \vdots & & R_{i,j} & & \vdots \\ & & & \ddots & \\ R_{m,1} & & \cdots & & R_{m,n} \end{bmatrix} \tag{6.12}$$

Here $m \gg n$. Each column represents the adjoint concentration of a sensor No. j corresponding to all possible sources. Each row represents the adjoint concentrations of all sensors for a single source No. *i.* Because the sensor configuration is designed for all possible sources $\boldsymbol{s}$ in the target domain, all rows are included in $p(\boldsymbol{R}|I)$. In this case, we regard each row as a vector, and calculate the probability distribution of these vectors, which is $p(\boldsymbol{R}|I)$.

Each $R_{i,j}$ is a multiple of source strength and adjoint concentration of a spatial point. However, the source strength is unavailable when the sensor configuration is designed. It could have a wide range, which makes m in Eq. (6.12) huge and calculation cost of $p(\boldsymbol{R}|I)$ heavy. To decrease this cost, we notice that in the Bayesian inference for STE, $q_{\boldsymbol{s}}$ is sampled by the MCMC with a uniform prior distribution. Meanwhile, in the process of location optimization for sensor*s*, changing the location of *e* would only affect $\overline{C_j^*}$, so the entropy $H(\boldsymbol{R}|I)$ mainly depends on the $\overline{C_j^*}$. Hence, in the case study of Section 6.3, we calculate $H(\overline{\boldsymbol{C}^*}|I)$ instead of $H(\boldsymbol{R}|I)$ and assume that larger

$H(\overline{\boldsymbol{C}^*}|I)$ is equivalent to larger $H(\boldsymbol{R}|I)$. In the following content, $H(\boldsymbol{R}|I)$ means $H(\overline{\boldsymbol{C}^*}|I)$. In this case, m in Eq. (6.12) equals to the number of spatial grids of the adjoint concentration field $\overline{C_J^*}$.

Future research is still necessary to clarify the effects of this calculation method on the final result. It is still recommended that a wide range of discrete q_s should be multiplied with the adjoint concentration if enough calculation resources is available.

6.2.4 Limitation of the method

Because this method was proposed based on several assumptions, there are some limitations that need to be noted. First of all, this method corresponds with the statistical STE problem of a point source in the target area with stationary meteorological situations. Its effectiveness cannot be ensured when applied to other environmental monitoring. More research is still necessary to evaluate its performance in real urban applications with complex meteorology. However, because numerical simulations of adjoint concentration include the information of meteorology, it is reasonable to say that the proposed method has the potential ability in such scenarios.

Besides, during the mathematical proof, one requirement is that the adjoint concentration should be accurately simulated to ensure that Eq. (2.24) is satisfied. As a result, the performance of this proposed method may depend on the quality of adjoint concentration simulation. What's more, it is useful to note that there is no guarantee that SA would converge at the global optimal, however, it is likely that a 'near optimal' can be obtained (Altinel et al., 2008). Considering the complicated errors in real applications, this imperfection is considered to be acceptable here.

6.3 Case study

To verify the performance of the proposed method, it is used to design an optimum sensor configuration for an ideal urban area: a regular block-arrayed building group model, just the one in **Chapter 4**.

To comprehensively evaluate the quality of the sensor configurations, it is necessary to identify different sources based on the configuration and check the corresponding STE performance. However, it is troublesome to measure the concentrations of many sources in the wind tunnel experiment, let alone that measurements of several sensor configurations are needed, and it is almost

impractical to employ the experiment or field test in this evaluation. Moreover, LES is increasingly gaining credit in the prediction of urban flow fields and dispersion of pollutants (Kikumoto and Ooka, 2012). Some research also relies on the simulation data for sensor network design (Abida et al., 2008) or STE method verification (Mons et al., 2017). As a result, the well-validated LES database produced in **Chapter 4** is applied for the evaluation of sensor configurations.

6.3.1 Numerical simulation

The statistical data of the LES are validated in **Chapter 4** in **Fig. 4.4 & 4.5**. It can be confirmed that the simulated flow field and concentration agree well with the wind tunnel experiment measurements. Therefore, using the LES results as concentration measurements would not damage the credibility of the research.

6.3.2 Unknown sources

A red area is set as the target monitoring domain where an unknown source may appear, as shown in **Fig. 6.2**. The target area is smaller than the entire calculation domain because the measurement information around the outer bounds could be valuable, especially downstream measurements. If the target domain is the same as the calculation domain, the source is difficult to estimate with most sensor configurations when it is on the streamwise boundary.

Then, some sources are set in the target domain to evaluate the quality of each sensor configuration through the STE. It is difficult to obtain the concentration fields of many sources, regardless of experiments or simulations. Because the block configuration is regular in the current case, the concentration field of one point source located at (4H, 7H, 0) simulated in **Chapter 4** was used and copied for the other 24 sources, as shown in **Fig. 6.2**. All these sources are located in the middle of the wake region, where the unsteady turbulent flow is strong, posing a challenging STE problem for sensor configurations. Note that the concentration simulation was well-validated, and the creditability of the concentration database was maintained.

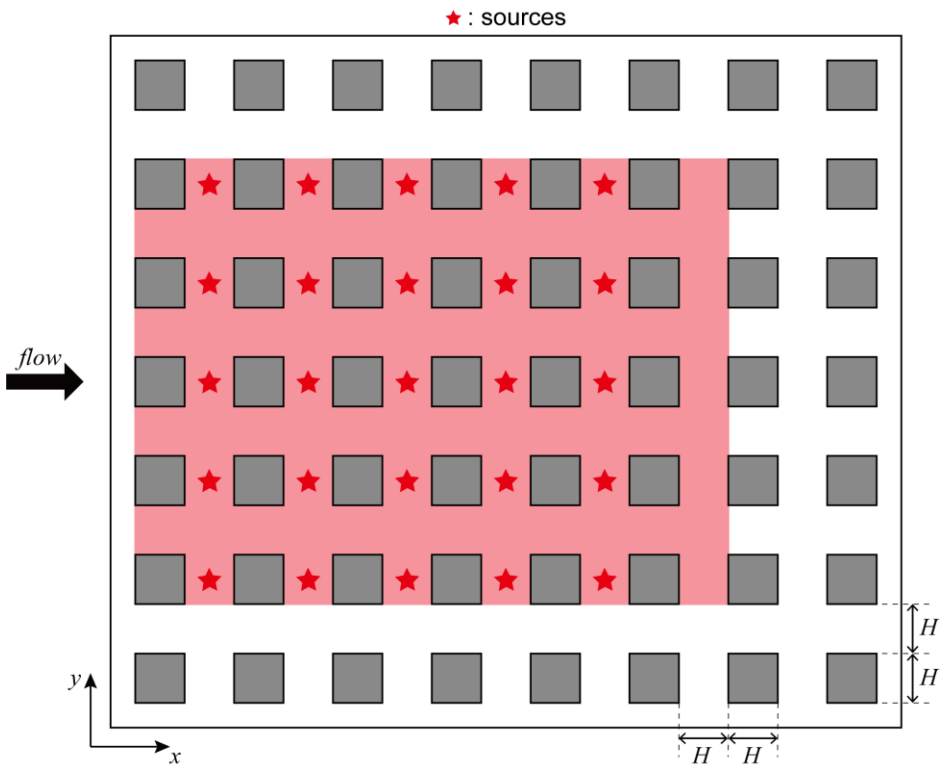


Figure 6.2. Target monitoring domain (red area) and distribution of unknown sources

6.3.3 Sensor candidates

In this part, a number of sensor candidates are prepared, from which the optimization process can select part of sensors to construct the sensor configuration. In real applications, considering the deployment limitation of the stationary sensors and complex terrain in the urban areas, most places cannot be used for reasons like space is not enough or the land property problems. It is more practical to conduct optimization among the candidates than from all positions. Meanwhile, in the proposed method, the adjoint concentration fields of all sensor candidates have to be simulated in advance. If there are too many candidates, the computational cost is very large, and a huge database has to be saved for the entropy calculation later. Therefore, it is difficult to regard all the places as the location candidates. In the previous research, the design is usually selected from the sensor candidates prepared in advance (Kouichi et al., 2019; Ngae et al., 2019). The distribution of the sensor candidates in this case study is illustrated in **Fig. 6.3**. A total of 113 sensors were located on the horizontal plane with $z = H/2$. The objective of the optimization is to select the eight best sensor

combinations among these candidates.

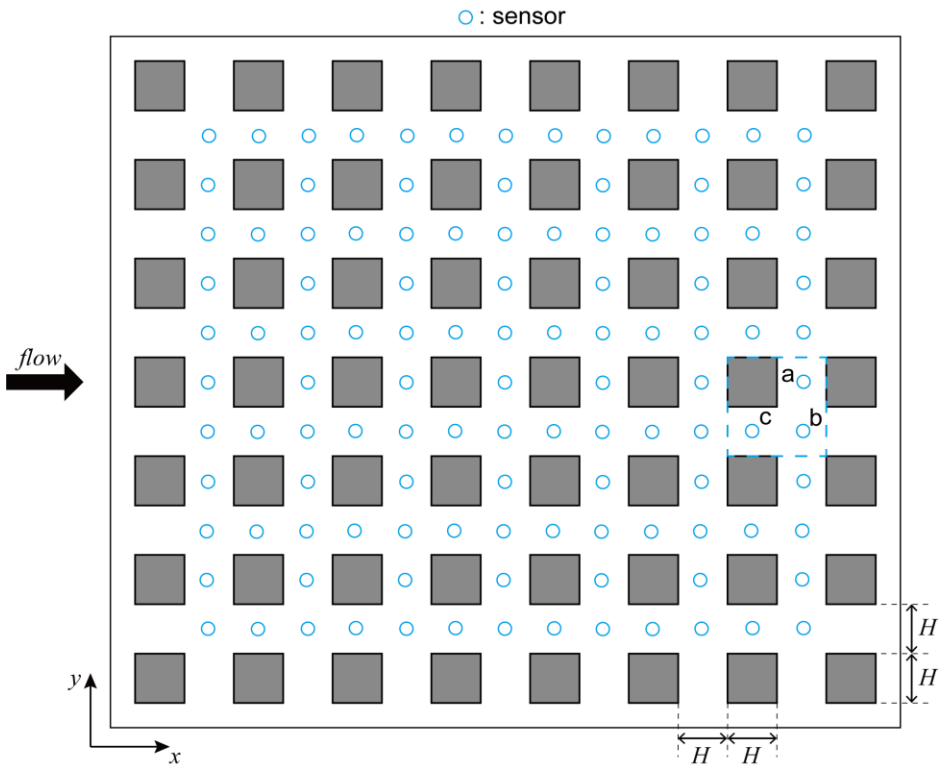


Figure 6.3. Distribution of sensor candidates

Because the objective function is based on the entropy of the adjoint concentration, the heaviest calculation burden is the adjoint equation simulation of all sensors. Once again, the advantage of a regular flow field is used and the adjoint concentration fields for only three sensors (a, b, and c in **Fig. 6.3**) are simulated. The fields of the other sensors are obtained by copying these three fields.

In **Chapter 5**, the performance of Reynolds averaged Navier-Stokes (RANS) and large-eddy simulation (LES) of adjoint equations has been compared and discussed. Although LES has been confirmed to be more accurate, to balance accuracy and calculation costs, the approach proposed by Xue et al. (2018b) is still adopted here, in which the adjoint equation was simulated by a RANS-like model based on the mean velocity field predicted by LES in the forward simulation. It should be noted that this method will bring some modeling errors into the adjoint concentration field and SCO. The obtained optimal configuration may not be the theoretically best one. Meanwhile, the simulated adjoint concentrations are also used in the following STE for all configurations, so it is believed that the modeling errors of adjoint equations simulation will not affect the ranking of different configurations.

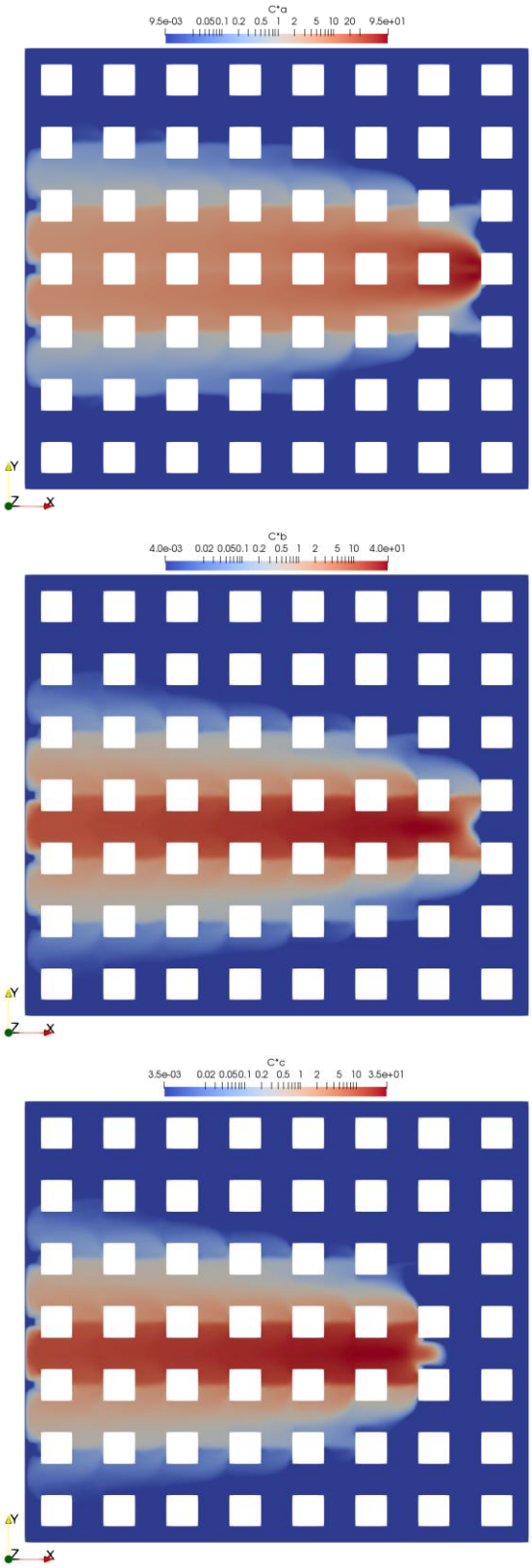


Figure 6.4. Simulated adjoint concentration distribution of sensors a, b, and c in **Fig. 6.3**. (Horizontal plane with $z = 0$; From top to bottom: sensor a, b, c)

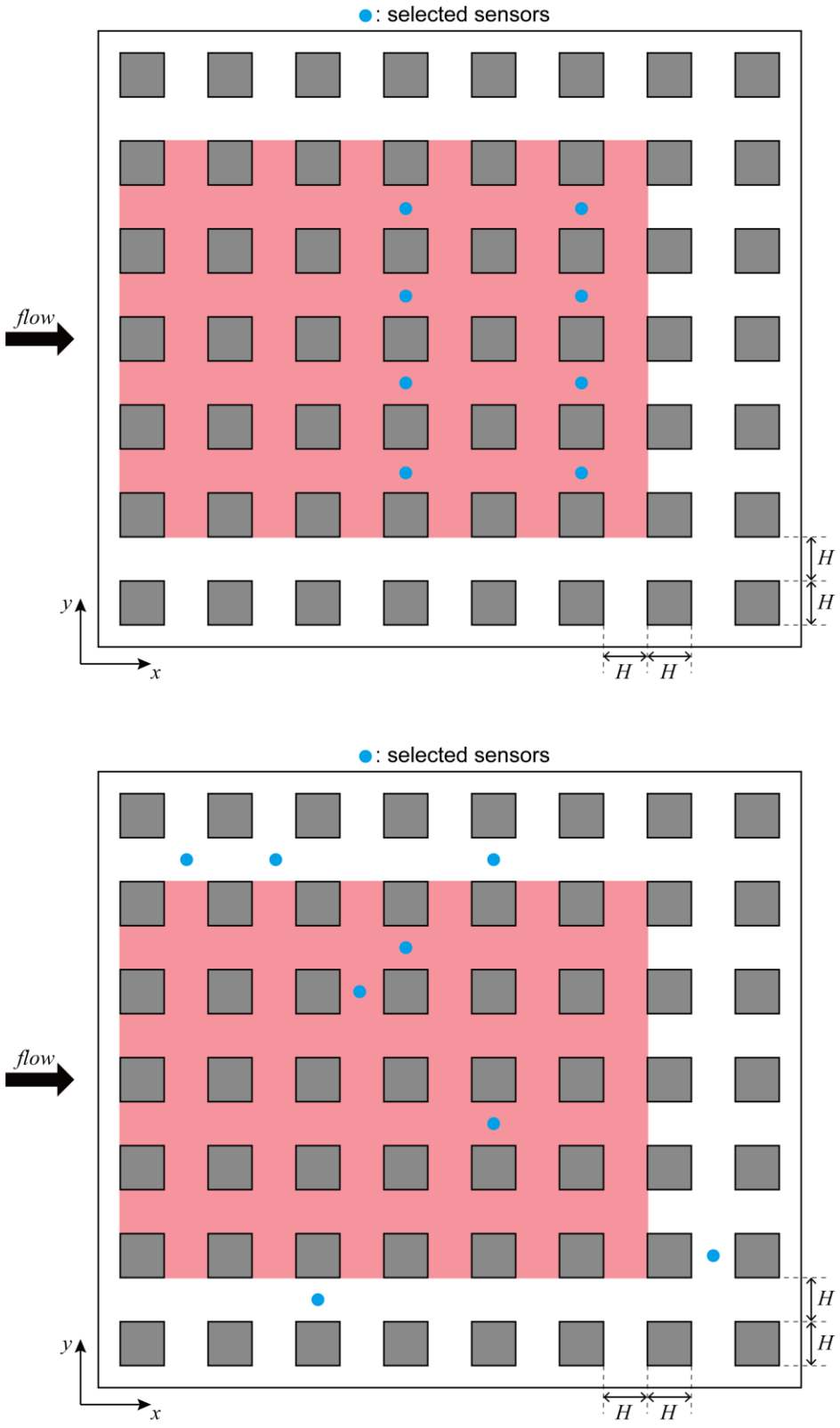


Figure 6.5. Comparison configurations: uniform (top) and random (below)

1

The adjoint concentration fields of three sensors are shown in **Fig. 6.4**.

6.3.4 Comparison configurations

To demonstrate the advantage of the optimum sensor configuration, two comparison configurations that were commonly used in previous STE research are selected, namely, the uniform configuration and random configuration (**Fig. 6.5**). The STE results of these two configurations are compared with those of the optimal configuration.

6.4 Results and Discussion

6.4.1 Optimum sensor configurations

The proposed method is applied to determine the optimum sensor configuration. The value of the objective function of each of the eight sensor combinations $H(\boldsymbol{R}|I)$ was calculated, and the SA quickly adjusted the configuration to reach a larger value. **Fig. 6.6** illustrates the changing process of the objective function value during the eight SA searching chains. Seven chains are transparentized to emphasize the one whose final objective function value is the maximum. The logarithmic horizontal axis was used to clearly illustrate the search process. The SA optimization algorithm was run in MATLAB on a personal computer with Intel® Core™ i7-6700 CPU @ 3.4GHz and 32GB of RAM. The averaged computational time for one SA chain with 5000 optimization steps is about 180s.

The changing patterns of the eight chains were similar. Initially, because the SA chain started from a random configuration, the entropy was small. Before approximately 100 steps, SA moved to a smaller entropy rather than a larger one several times. At this moment, the virtual temperature T_k still has a large value, making the SA flexible to jump into the inferior configuration for global exploration. After, the inferior jumping became strict, and the chain only moved to a larger entropy. Meanwhile, the rate of increase also decreased because it became increasingly difficult to find a better configuration. Finally, the value of the objective function remained unchanged for approximately 1500 steps, yielding the largest value of 11.21 and the corresponding configuration.

As a comparison, the $H(\boldsymbol{R}|I)$ values of the random and uniform configurations are 6.35 and 9.06, respectively. It can be confirmed that 6.35 is almost the smallest value in the chain, which

means that the random configuration is ineffective for measurements. This may challenge the robustness of STE methods and, consequently, should not be used in real applications. The uniform configuration is located at the intermediate level of the chain, with a value of 9.06. It should be noted that the regular block distribution provides some merits to the uniform configuration. Whether the performance can be maintained in a real urban area with a complicated terrain remains unclear.

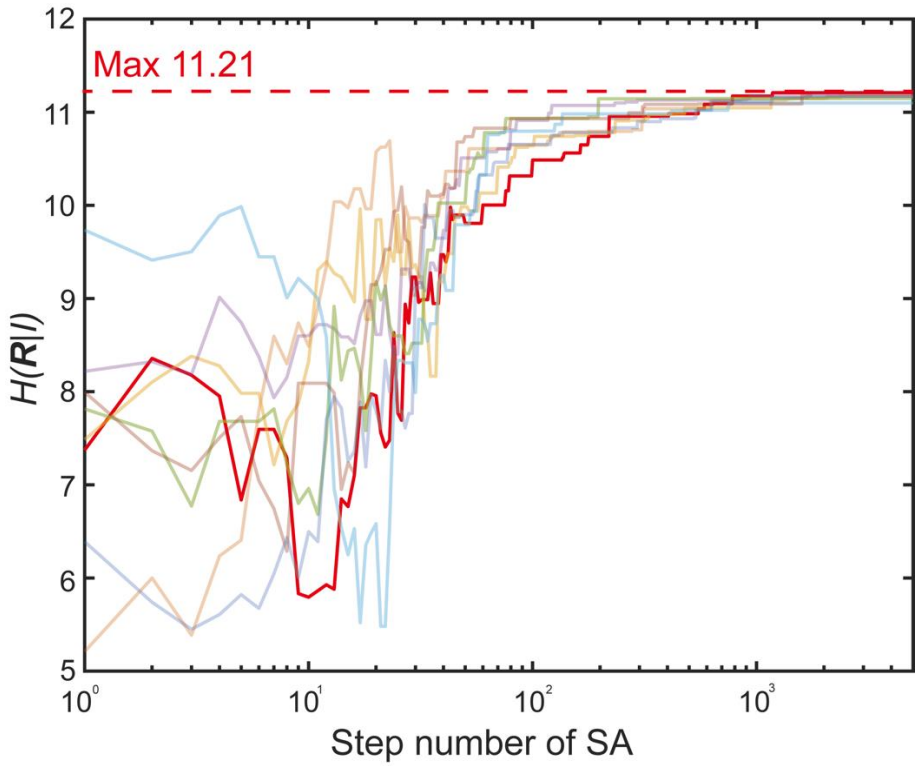


Figure 6.6. Variation in $H(\boldsymbol{R}|I)$ among eight SA chains. The red line indicates the chain whose final objective function value is the maximum. The other chains are shown with faded colors.

The obtained optimum sensor configuration is shown in **Fig. 6.7**. This configuration has several interesting characteristics. First, the configuration is not symmetric as the calculation domain. This could result from the simulation errors of the adjoint equation because the simulated adjoint concentration fields in **Fig. 6.4** were also slightly skewed. It is probable that the optimization result is sensitive to the adjoint concentration field, so the accurate simulation of the adjoint equation is important to the proposed method. From a good perspective, the proposed optimization method captures the information contained in the adjoint concentration field well, which further indicates that the source-receptor relationship is well considered.

Second, most sensors were located in the downstream area of the target domain. This distribution coincides with the intuitive decision that people usually tend to arrange all sensors downstream of sources based on empiricism. If the wind direction is constant, the adjoint concentration fields of these downstream sensors cover a sizable part of the target area, indicating a strong measurement capability. Although the logic seems obvious, it is difficult to quantize it for decision-making in the optimization algorithm. The produced optimal configuration implies that the entropy-based objective function in the proposed method is an appropriate solution.

Moreover, there is one characteristic beyond empiricism that not all sensors gather downstream. The algorithm detected that the downstream area was already crowded with six sensors. It is ineffective to add more sensors there in terms of cost performance because the adjoint concentration field of the new sensor would nearly overlap with existing ones, and the new source–receptor information is limited. Consequently, two sensors were arranged in the middle of the target domain to distinguish the front and back parts of the domain. The importance of these two sensors is easy to ignore when the sensor configuration is designed based on the human experience. The proposed method can deepen the insights from the entropy of the adjoint concentration.

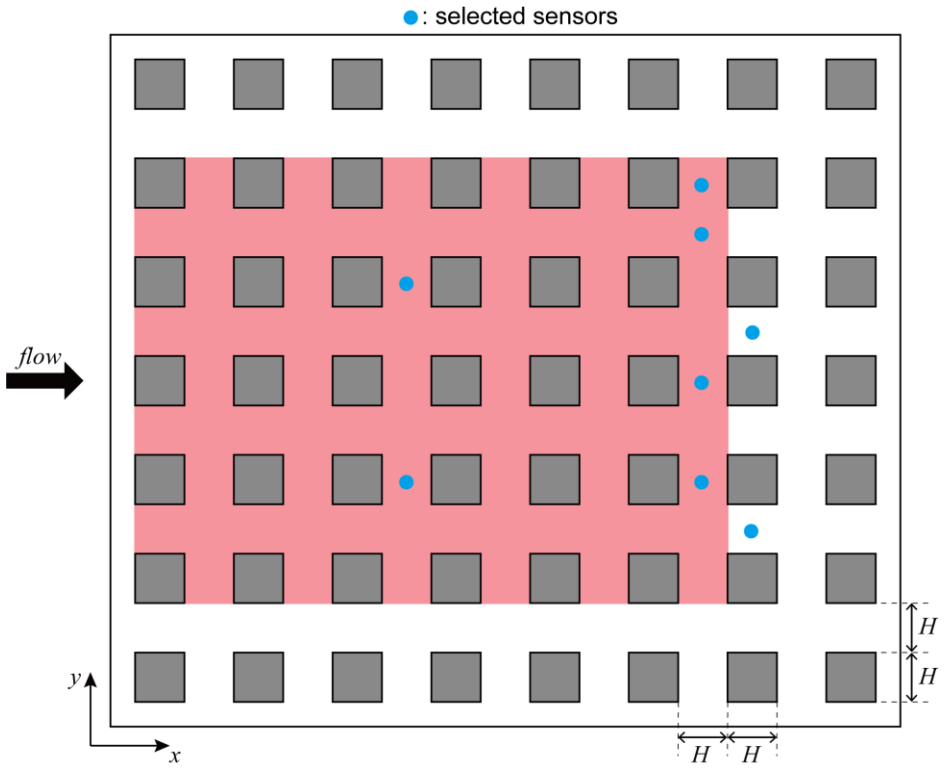


Figure 6.7. Optimum sensor configuration produced by the proposed method

According to the discussion above, the chosen optimum sensor configuration is reasonably designed and seems capable of satisfying the STE performance. We employed this configuration as well as random and uniform configurations to estimate 25 unknown sources. In the Bayesian inference, the variance of errors is set as $\sigma_{d,i}^2 + \sigma_{m,i}^2 = r * D_i$, where $r = 0.3$. The results are as follows.

6.4.2 Estimation results of one source

First, we present the STE results of one source located at (2*H*, 7*H*, 0) based on the measurements of the three sensor configurations. The Bayesian inference of STE was executed with MATLAB on a personal computer with Intel® Core™ i7-6700 CPU @ 3.4GHz and 32GB of RAM. The averaged computational time for estimating one source is about 11s. **Fig. 6.8** summarizes the posterior joint probability distribution $p(x, y|\boldsymbol{D}, I)$, which represents the location estimation of the true source. **Fig. 6.9** shows the posterior marginal probability distribution for the strength estimation.

The optimal configuration produced the best estimations among the three. In **Fig. 6.8(a)**, most of the probability mass is concentrated around the true source. Only a fraction deviates from the truth, and the distance is constrained to approximately 2*H*. The accuracy is better than the STE in **Chapter 5** with the same adjoint equation simulation method. Possible reasons could be that the optimal sensor configuration is more informative. Besides, it can be noticed that the downstream distance between source and sensors is larger than that in **Chapter 5**. According to the previous research (Xue et al., 2017), the simulated adjoint concentration distribution tends to be a narrower plume compared with that in the real scenario. Therefore, a smaller distance may result in a larger difference between $\boldsymbol{D}$ and $\boldsymbol{R}$.

In addition, the emission strength was also well identified, as shown in **Fig. 6.9(a)**. All probability masses lie between 0 and 2. Although the peak value does not perfectly agree with the truth, the error is acceptable. This estimation error mainly results from the simulation method of the adjoint equation. The LES model could effectively mitigate this problem and further improve the STE accuracy. Because this is not relevant in the current study, the details are not expanded upon here.

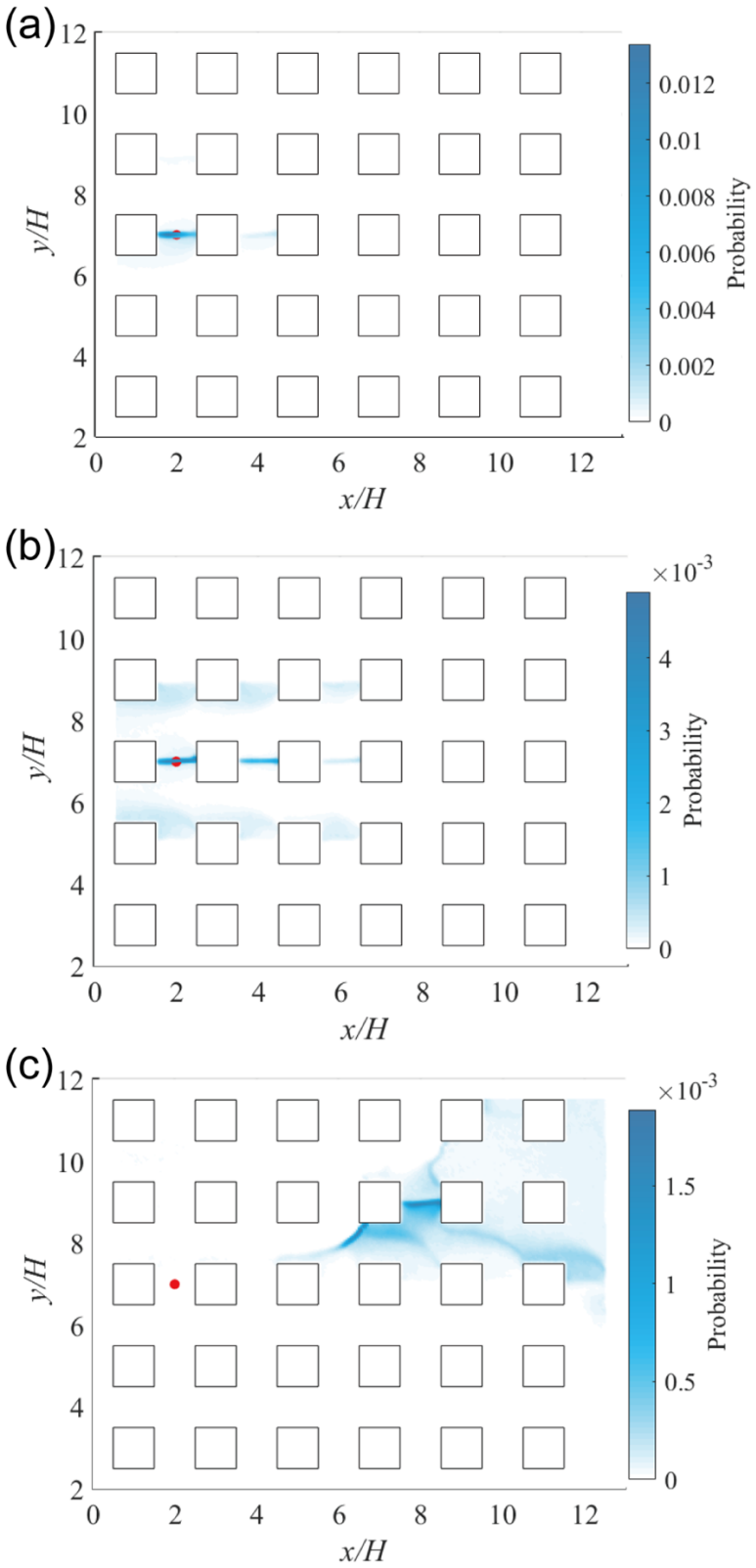


Figure 6.8. Posterior joint probability distribution $p(x, y|\boldsymbol{D}, \boldsymbol{I})$ obtained by three sensor configurations: (a) optimum; (b) uniform; and (c) random. Red point: true source.

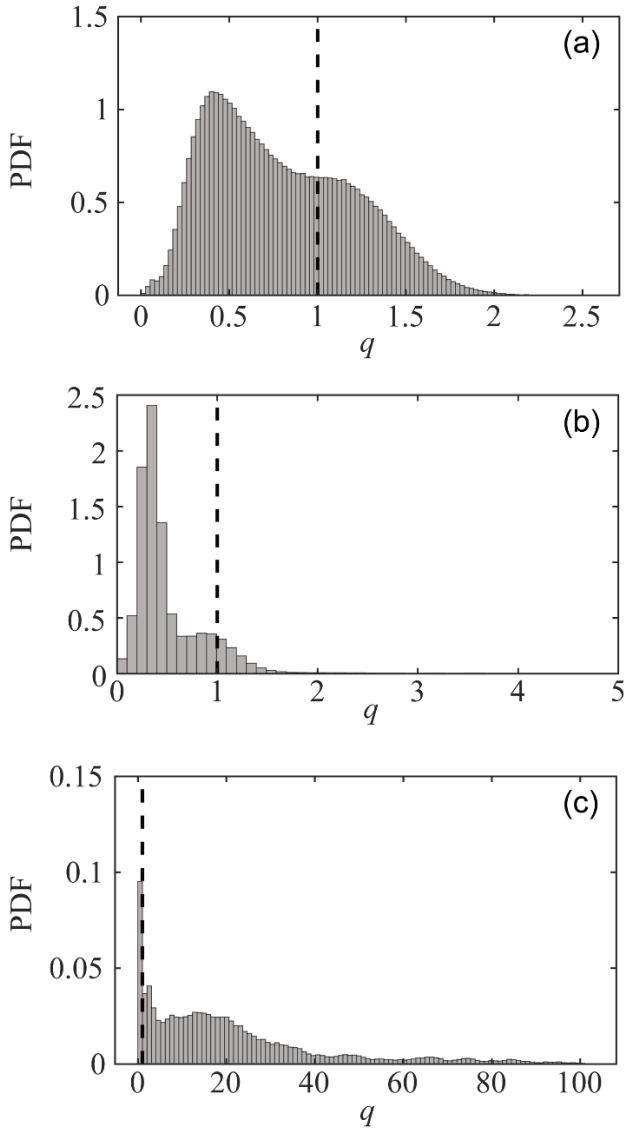


Figure 6.9. Posterior probability distribution $p(q|\boldsymbol{D},\boldsymbol{I})$ obtained by three sensor configurations: (a) optimum; (b) uniform; and (c) random. Dotted line: true value

The uniform configuration failed to provide estimations with comparable accuracy to that of the optimal configuration. Although the true location has a distribution peak, many probability masses diverge broadly. This divergence represents redundant uncertainties caused by low-quality measurements, which is insufficient for STE sampling to firmly adhere to the truth. The same situation can also be confirmed in the strength estimation shown in **Fig. 6.9(b)**. Although the width and the general shape are similar to those of the optimal configuration, more probability masses move from the truth to the wrong peak. In short, the STE estimation contains more noisy

uncertainties than that of the optimal configuration because the measurements of this source provided by the uniform configuration are less informative.

The performance of the random configuration was the worst. All probability mass in **Fig. 6.8(c)** departs from the true source, which reveals that the STE broke down. The information measured by the random configuration cannot support the STE in identifying the true source. Similarly, severe unsteadiness was observed in the strength estimation. Although the peak is close to the truth, the width of the distribution spreads over 100 times, as shown in **Fig. 6.9(c)**. In this case, the estimations have little reliability for risk management.

In general, the performance of these three configurations has the same order as their entropy values, which proves that the objective function design in the proposed method is reasonable.

6.4.3 Estimation results of all sources

To avoid the contingency in STE for only one source, we conducted STE for all 25 sources based on the three configurations. Considering the length of the thesis, it is difficult to present all posterior PDFs here. Two indices $E_{d,m}$ and $E_{q,m}$ introduced by Eq. (5.6 & 5.7) are used to quantify the estimation accuracy of STE for all unknown sources based on the three sensor configurations. m is the index for source $No.m$

Accurate STE always has smaller $E_{d,m}$ and $E_{q,m}$. Because the posterior PDFs are often highly skewed, $E_{d,m}$ and $E_{q,m}$ are first-order statistical moments and cannot comprehensively reflect real situations. The two PDFs may be different, even if their expectations are the same. In the Bayesian inference STE, it is expected that the probability mass is concentrated around the true value with limited distribution rather than a uniform distribution with too many uncertainties. As a result, we introduced the third index to evaluate the concentration degree of the posterior PDF: the joint entropy $H(x,y,q)$ of $p(x,y,q|\boldsymbol{D},\boldsymbol{I})$. A PDF with a small $H(x,y,q)$ and small $E_{d,m}$ & $E_{q,m}$ indicates a good estimation that most probability is close to the truth.

The values of the three indices for STE based on the three configurations are summarized in **Table 6.1**. All values were averaged for 25 unknown sources. The optimal configuration has the smallest values for all three indices. The location error of 2.83H and strength error of 1.5 are at the normal level for STE with a steady simulation of the adjoint equation, which has been confirmed in **Chapter 5** with 16 sensors. When the sensor configuration becomes uniform, the estimation

accuracy becomes inaccurate according to the three indices. The location error increased by approximately 65%, and the strength error significantly increased by approximately 25 times when compared with that of the optimal configuration. The limited amplification of $H(x, y, q)$ demonstrated that considerable deviation was caused in $E_{q,m}$ by ineffective measurements of uniform configuration while the concentration rate of PDF was less damaged. However, the random configuration thoroughly eroded the estimation accuracy in terms of deviation and concentration rate in the PDF. The location error, strength error, and $H(x, y, q)$ increased by more than 80%, 7300%, and 47%, respectively, compared with the optimal configuration. It should be mentioned that most sensors in the random configuration were located in the upper-left area. When the source appears at the lower right corner, only one or two sensors can measure the concentration changes, which is obviously too insufficient and results in STE failures.

Consequently, the efficiency of the optimum sensor configuration in STE was confirmed. The performance of the configurations is positively proportional to the joint entropy of adjoint concentrations $H(\boldsymbol{R}|I)$. Among these three indices, the strength error is the most sensitive to the measurement quality of the sensor configuration. During the Bayesian inference, the exploration space of the location is limited in the current case (approximately 12H for x and 10H for y), while the sampling range is infinite for strength. When the measurements are not sufficient for source identification, Bayesian inference would prefer to adjust the strength rather than the location parameter to meet the gap between the measurements and simulated concentrations. Once a probable parameter combination was found, most sampling would occur in that combination. As a consequence, E_{di} and $H(x, y, q)$ were not promoted comparably to E_{qi} in the random configuration.

Table 6.1. Summarized estimation indices of three sensor configurations

	Mean $E_{d,m}$	Mean $E_{q,m}$	Mean $H(x, y, q)$
Optimal configuration	2.83H	1.54	12.23
Uniform configuration	4.67H	39.54	14.17
Random configuration	5.17H	112.53	18.01

6.5 Conclusions

This chapter proposes an SCO method for identifying an optimum sensor configuration that can provide informative concentration measurements for STE, regardless of where the unknown source appears in the target domain. When compared with other configurations, it is expected that Bayesian inference STE can produce better estimations based on measurements of the optimal configuration, in which the probability mass in the posterior PDF gathers around the true value with a narrow distribution.

The proposed method is composed of an objective function and an SA algorithm. The objective function was set as the joint entropy of the spatial probability distribution of the adjoint concentration of a configuration to evaluate its measurement ability. The reasonability of this setting was explained from mathematical and physical perspectives. The value of the objective function represents the uncertainty about the source that can be measured by the sensor configuration. If more uncertainty has been measured, less probability divergence would appear in the posterior PDF, and the probability mass could concentrate around the truth. Therefore, the optimum sensor configuration should have the largest objective function value. SA was applied to quickly find the optimal configuration using the objective function value among countless possible combinations.

To evaluate the performance of the proposed method, 25 unknown point sources in the regular block-arrayed building group model were estimated based on measurements of optimum, random, and uniform configurations. According to the results, it was found that the main calculation cost in the proposed method results from the adjoint equation simulations for all sensor candidates. By comparing the STE performance of the three configurations, it was proved that the estimation accuracy is positively correlated with the objective function value such that the optimal configuration is the best, the uniform configuration is less accurate, and the random one is the least accurate. Hence, the objective function is appropriately selected. In the presented posterior PDFs for the STE of one source, the probability mass settled around the truth with a limited spread in the optimal configuration, while the estimations totally deviated in the random configuration. This finding was confirmed again when the STE results of 25 unknown sources were quantized using three indices. The average estimation errors of the optimal configuration were limited. The location and strength estimation errors increased by approximately 80% and 7300%, respectively, when the configuration changed from optimum to random.

The measurements in the case study were synthesized from a well-validated LES simulation. Future research with complicated urban terrain and real measurements is necessary to verify the robustness of the proposed method. Besides, the meteorological condition was set stationary by constant inflow boundary in adjoint equation simulation. The optimal sensor configuration may change with different wind directions/speeds. However, it has been presented that the adjoint concentration simulation is able to include the effects of terrain geometry, meteorological situations, and building structures. It is reasonable to expect that the proposed method can be extended to real urban applications with variable conditions.

1 Symbols

C : the concentration field

$C_{\boldsymbol{s}}$: the concentration field of source $\boldsymbol{s}$

C_e^* : the adjoint concentration field of sensor e

C_n^j : the number of possible j combinations out of a set with n elements

$\boldsymbol{D}$: the measurements vector

D_i : the measurement of the sensor with index i

e : the selected sensor for further optimization when other sensors are fixed

$E_{d,m}$: the expectation of the Euclidian distance between the estimated location and the true location of source $No.m$

$E_{q,m}$: the expectation of the normalized difference between the estimated strength and true strength of source $No.m$

H : the edgy length of blocks in the simulation (=60 mm)

$H(x)$: the information entropy of a probability distribution $p(x)$

I : the background information for Bayesian inference

$I_m(\boldsymbol{s};\boldsymbol{D}|I)$: the mutual information representing the information about the source $\mathbf{s}$ provided by sensors with measurements $\boldsymbol{D}$

$\boldsymbol{m}_i$: the ith searching in the Simulated Annealing

N : the specified number of searching steps in Simulated Annealing

$N[\mathrm{a}, \mathrm{b}]$: the uniform distribution bound between a and b

$p(x)$: the probability of event x occurring

$p(A|B)$: conditional probability of event A occurring given that B is true

Q : the gas flow rate at the source

q : strength samplings produced in MCMC for the point source

q_s : the true strength of the point source

r : the ratio between the specified error covariance in the Bayesian inference and the measurements

$\boldsymbol{R}$: adjoint concentration fields

$\boldsymbol{s}$: a vector representing the unknown source

T_0 : the initial virtual temperature for Simulated Annealing

T_k : the virtual temperature of the system at the kth step in Simulated Annealing

$\mathbb{V}$: a parameter space where the unknown source is possibly located

(x, y) : location samplings produced in MHMC for the point source

(x_s, y_s) : the true location of the point source

a : the cooling coefficient in the Simulated Annealing

$\sigma_{d,i}^2$: the variance of error in the measurement of the sensor with index i

$\sigma_{m,i}^2$: the variance of error in the modeling concentration for the sensor with index i

$\bar{}$: Reynolds average operator

1

2

Chapter 7
Conclusion and Future research

7.1 Conclusions

This thesis develops a statistical estimation method for an unknown source of atmospheric pollutants in the complicated urban environment based on Bayesian inference. The main conclusions of each chapter are as follows.

Chapter 1 introduces the research background, objective, and structure of the thesis.

Dispersion emergencies caused by unknown sources occurred in the urban area from time to time and may cause considerable damage to people and the environment. It is necessary to estimate the source parameters as soon as possible after the emergencies happen. Dealing with the characteristics of source term estimation (STE) applications in the urban environment, this thesis proposed three improvements for the statistical STE to realize better estimation performance, and evaluate their effectiveness by a series of case studies.

Chapter 2 provided a short review of the recent progress in STE of atmospheric pollutants, and the basic methodology used in the thesis is introduced.

STE consists of three basic elements: measurements, estimation algorithm, and source-receptor relationship. Different methods for these elements have been proposed in the previous research. Considering the applications in the urban environment, in this thesis, the time-averaged concentrations which can be measured by most sensors are the measurements; the Bayesian inference is selected as the estimation algorithm to assess the noise in STE; the adjoint equation method is applied for the source-receptor relationship simulation to deal with the complicated dispersion phenomenon in the built area.

Chapter 3 proposed a line source estimation method by combining the basic methodology with the super-Gaussian function.

The coefficients of the super-Gaussian function were added into Bayesian inference to realize the geometry estimation. The applicability was first confirmed by a numerical experiment of an ideal urban boundary layer. The proposed method successfully inferred that the source is line-like without any prior knowledge. Based on this case, the effects of different sensor configurations on the line source estimation were discussed. Because the line source contained more geometric information than point sources, the requirements on the sensor configuration become higher that both sensors near the source and null-measurement sensors are indispensable. The conventional sensor configuration may fail in the line source estimation.

To examine the robustness of the proposed method against measurement and modeling errors, the second case of a simplified urban square with wind tunnel experiment measurements was conducted. The line source was successfully identified by the proposed method again. By comparing to the conventional method with ideal point assumption, it is confirmed that the proposed method can not only provide the geometry estimation but also reduce the inference errors caused by the point source assumption. It is meaningful to include the geometry estimation into STE.

Chapter 4 constructed a compression database by wavelet-based compression method for the turbulent flow field of a block-arrayed building group model based on large eddy simulation (LES) raw data. It is a prerequisite content for Chapters 5 & 6.

The compression ability and error control of the wavelet-based compression method was analyzed. The influence of compression on the quality of the data was checked from a single snapshot and time-series perspectives. In the case study, it was found that about 100 times compression can satisfy the requirement of flow field visualization, large-scale turbulent structure preservation, and afterward dispersion simulation. Therefore, it is reasonable to say that the wavelet-based compression method is a powerful tool to construct a portable flow database. The unsteady simulation of the adjoint equation can be realized based on a compressed flow field.

In Chapter 5, to improve the accuracy of STE in complex urban applications, a new method was developed based on Bayesian inference coupled with unsteady adjoint equation modeling via LES.

The performance of the proposed method was evaluated in the block-arrayed urban model with the wind tunnel experiment measurements of continuous dispersion of a point source. The LES of the adjoint equation for the source-receptor relationship was realized based on the compressed flow field constructed in Chapter 4. As a comparison, another STE was also conducted with a conventional method, where steady adjoint equations were simulated with Reynolds averaged Navier-Stokes (RANS) model. The results showed that the modeling of the adjoint equation and STE were significantly improved by the LES. The complicated turbulent flows caused by buildings destroyed the reliability of the gradient diffusion hypothesis in the conventional RANS simulation of the adjoint equation. Although the proposed method needs more computational resources, to effectively perform STE in the complicated urban

environment, it is valuable to apply LES to adjoint equation simulation.

Chapter 6 proposed a sensor configuration optimization method for STE by the design of an objective function and application of the simulated annealing algorithm.

The objective function was set as the information entropy of the spatial distribution of the adjoint concentration field. Its ability to represent the measurement ability of sensor configurations was proved from the views of mathematics and physics. Simulated annealing was applied to find the optimal configuration which owns the largest value of the objective function.

The proposed method was utilized to design an optimal sensor configuration for the block-arrayed urban model in Chapter 4. The performance of the optimal configuration in STE was compared to uniform and random configurations through estimations for 25 unknown sources. The results revealed that the accuracy of STE is related to the entropy contained in the adjoint concentration of the configuration such that the design of the objective function is reliable. The optimal configuration outperforms the other two in STEs. It is valuable to use the proposed method to guide the configuration design in real applications.

7.2 Future research

Although some progress has been achieved in this thesis, there are still several limitations in the statistical STE method which need to be noticed. To ultimately realize effective STE for the urban environment, here are future research summarized based on the author's knowledge.

Firstly, more advanced measurements should be utilized in STE. In all STEs in the thesis as well as a considerable amount of previous research, only the time-averaged concentration of pollutants was used as measurements. Although this information is the simplest and most basic data to measure for sensors, if other advanced measurements can be accounted in, the accuracy of STE could be improved further following the measured information entropy in Chapter 6. For instance, adding the turbulent flux into measurements is likely to better distinguish the wrong estimations and reduce errors in STE further. Research using time-series measurements in STE is still sparse until now. It is believable that much more information can be dug out from time-series data than average value, thus more efforts should be devoted to how to use it in STE.

More importantly, the main limitation of time-averaged concentration is that it should be measured after the dispersion has reached into the statistically steady state by a large enough averaging time scale. It means that plenty of time is cost by measurement before the STE, which potentially brings more risk due to the continuous dispersion of unknown sources during this time. It is meaningful to develop a technique which can identify the unknown source faster based on the sudden change of concentration or time lags between sensors.

Secondly, the estimation target is assumed to be a single source. It means that the proposed methods may suffer from identifying multiple sources that the accuracy cannot be promised any more. Since dispersion emergency with multiple sources is a possible scenario in the real applications, it is necessary to extend the proposed methods to handle with them in the future.

Thirdly, the simulation method for adjoint equations can be improved further. Although the proposed LES of adjoint equations realized accurate prediction of source-receptor relationship, its calculation cost is still large for real applications. Methods with better cost-performance ratios should be developed in the future. According to the results in Chapter 4, the decompressed flow fields only persevering large-scaled turbulent structures can yield almost the same prediction accuracy for the dispersion simulation. It is possible that the unsteady simulation of adjoint equations could be successfully driven only by large-scale turbulent structures. Under such circumstances, bridging models between RANS and LES like unsteady RANS, partially-averaged Navier-Stokes, or detached eddy simulation have the opportunities to simulate adjoint equations. About the data storage for the inverse simulation, apart from the wavelet-compression method in the thesis, the large-scaled turbulent structures can also be stored by low-dimensional modes like proper orthogonal decomposition or dynamic mode decomposition.

Furthermore, the case study in the thesis is a regular block-arrayed building group model, which is an ideal situation for dispersion in the urban environment. Dense buildings with diverse geometries and irregular distribution may cause an unexpected challenge to STE. The moving objects like automobiles, trains, and people will make the source-receptor relationship more complicated to predict in the neighborhood scale. Their influence on the accuracy of STE should be addressed in the future, and if the influence was unignorable, it is also necessary to study how to model them in the source-receptor relationship.

Another critical factor to source-receptor relationship modeling is the meteorological

condition, which was assumed to be stationary during the dispersion emergency in the thesis. However, in the reality, meteorological conditions like wind speed and directions are unsteady at different times. It is likely that the inflow condition would suddenly change during the dispersion emergency. Hence, it is valuable to extend the current method to handle unsteady dispersions.

Meanwhile, the mechanisms of pollutants' dispersion are extremely complicated. The particles may go through processes like collision, sediment, condensation, and so on. Some toxic gases may also react with other gases during the dispersion. For these pollutants, the current method may totally fail because of the passive scalars assumption. Even the model for forward simulation is still on research to predict this kind of complicated dispersion. It is a huge challenge for STE to cover a comprehensive dispersion mechanism.

Last but not least, the proposed methods in the thesis have not been testified by field test. In fact, available databases of field experiments are limited in the literature. Most STE research used the data from wind tunnel experiments or numerical experiments, which are relatively ideal compared to real scenarios. One of the most famous field tests is the Mock Urban Setting Test, where the containers are regularly placed in an open area to design an ideal urban model. However, the field experiment with dispersion in the real urban area is far too sparse, even though it is beneficial to future research like new method development, validation, and noise analysis for STE. It is expected that such a field test database could be built.

The above is a brief description of possible research contents to develop a more effective STE method for the urban environment in the future.

Reference

Abd Razak, A., Hagishima, A., Ikegaya, N., Tanimoto, J., 2013. Analysis of airflow over building arrays for assessment of urban wind environment. Building and Environment 59, 56–65. https://doi.org/10.1016/J.BUILDENV.2012.08.007

Abida, R., Bocquet, M., Vercauteren, N., Isnard, O., 2008. Design of a monitoring network over France in case of a radiological accidental release. Atmospheric Environment 42, 5205–5219. https://doi.org/10.1016/j.atmosenv.2008.02.065

Allen, C.T., Haupt, S.E., Young, G.S., 2007. Source Characterization with a Genetic Algorithm–Coupled Dispersion–Backward Model Incorporating SCIPUFF. Journal of Applied Meteorology and Climatology 46, 273–287. https://doi.org/10.1175/JAM2459.1

Altinel, I.K., Aras, N., Güney, E., Ersoy, C., 2008. Binary integer programming formulation and heuristics for differentiated coverage in heterogeneous sensor networks. Computer Networks 52, 2419–2431. https://doi.org/10.1016/j.comnet.2008.05.002

Araki, S., Iwahashi, K., Shimadera, H., Yamamoto, K., Kondo, A., 2015. Optimization of air monitoring networks using chemical transport model and search algorithm. Atmospheric Environment 122, 22–30. https://doi.org/10.1016/j.atmosenv.2015.09.030

Aubinet, M., Vesala, T., Papale, D. (Eds.), 2012. Eddy Covariance: A Practical Guide to Measurement and Data Analysis. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-007-2351-1

Balczó, M., Lajos, T., 2015. Flow and dispersion phenomena in a simplified urban square. Periodica Polytechnica Civil Engineering 59, 347–360. https://doi.org/10.3311/PPci.7852

BBC, 2019. The Amazon in Brazil is on fire - how bad is it? [WWW Document]. URL https://www.bbc.com/news/world-latin-america-49433767

Berkooz, G., Holmes, P., Lumley, L.J., 1993. The Proper Orthogonal Decomposition in the Analysis of Turbulent Flows. Annual Review of Fluid Mechanics 25, 539–575. https://doi.org/10.1146/annurev.fluid.25.1.539

Blocken, B., 2018. LES over RANS in building simulation for outdoor and indoor applications: A foregone conclusion?, Building Simulation. https://doi.org/10.1007/s12273-018-0459-3

Blocken, B., Stathopoulos, T., Carmeliet, J., 2007. CFD simulation of the atmospheric boundary layer: wall function problems. Atmospheric Environment 41, 238–252. https://doi.org/10.1016/j.atmosenv.2006.08.019

Branford, S., Coceal, O., Thomas, T.G., Belcher, S.E., 2011a. Dispersion of a Point-Source Release of a Passive Scalar Through an Urban-Like Array for Different Wind Directions. Boundary-Layer Meteorology 139, 367–394. https://doi.org/10.1007/s10546-011-9589-1

Branford, S., Coceal, O., Thomas, T.G., Belcher, S.E., 2011b. Dispersion of a Point-Source Release of a Passive Scalar Through an Urban-Like Array for Different Wind Directions. Boundary-Layer Meteorology 139, 367–394. https://doi.org/10.1007/s10546-011-9589-1

Chen, C., Liu, W., Lin, C.-H., Chen, Q., 2015. A Markov chain model for predicting transient particle transport in enclosed environments. Building and Environment 90, 30–36. https://doi.org/10.1016/j.buildenv.2015.03.024

Cheng, H., Castro, I.P., 2002. Near wall flow over urban-like roughness. Boundary-Layer Meteorology 104, 229–259. https://doi.org/10.1023/A:1016060103448

Claus, J., Coceal, O., Thomas, T.G., Branford, S., Belcher, S.E., Castro, I.P., 2012. Wind-Direction Effects on Urban-Type Flows. Boundary-Layer Meteorology 142, 265–287. https://doi.org/10.1007/s10546-011-9667-4

Coceal, O., Dobre, A., Thomas, T.G., 2007. Unsteady dynamics and organized structures from DNS over an idealized building canopy. International Journal of Climatology 27, 1943–1953. https://doi.org/10.1002/joc.1549

Coceal, O., Goulart, E.V., Branford, S., Glyn Thomas, T., Belcher, S.E., 2014a. Flow structure and near-field dispersion in arrays of building-like obstacles. Journal of Wind Engineering and Industrial Aerodynamics 125, 52–68. https://doi.org/10.1016/J.JWEIA.2013.11.013

Coceal, O., Goulart, E. V., Branford, S., Glyn Thomas, T., Belcher, S.E., 2014b. Flow structure and near-field dispersion in arrays of building-like obstacles. Journal of Wind Engineering and Industrial Aerodynamics 125, 52–68. https://doi.org/10.1016/J.JWEIA.2013.11.013

Coceal, O., Thomas, T.G., Castro, I.P., Belcher, S.E., 2006. Mean flow and turbulence statistics over groups of urban-like cubical obstacles. Boundary-Layer Meteorology 121, 491–519. https://doi.org/10.1007/s10546-006-9076-2

Combest, D.P., Ramachandran, P.A., Dudukovic, M.P., 2011. On the Gradient Diffusion Hypothesis and Passive Scalar Transport in Turbulent Flows. Industrial & Engineering Chemistry Research 50, 8817–8823. https://doi.org/10.1021/ie200055s

Constantine, P.G., Kent, C., Bui-Thanh, T., 2016. Accelerating Markov Chain Monte Carlo with Active Subspaces. SIAM Journal on Scientific Computing 38, A2779–A2805. https://doi.org/10.1137/15M1042127

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory. Wiley-Interscience, USA.

Cui, J., Lang, J., Chen, T., Cheng, S., Shen, Z., Mao, S., 2019. Investigating the impacts of atmospheric diffusion conditions on source parameter identification based on an optimized inverse modelling method. Atmospheric Environment 205, 19–29. https://doi.org/10.1016/j.atmosenv.2019.02.035

Daubechies, I., Sweldens, W., 1998. Factoring wavelet transforms into lifting steps. The Journal of Fourier Analysis and Applications 4, 247–269. https://doi.org/10.1007/BF02476026

de Freitas, N., Hojen-Sorensen, P., Jordan, M.I., Russell, S., 2013. Variational MCMC.

Efthimiou, G.C., Kovalets, I.V., Argyropoulos, C.D., Venetsanos, A., Andronopoulos, S., Kakosimos, K.E., 2018a. Evaluation of an inverse modelling methodology for the prediction of a stationary point pollutant source in complex urban environments. Building and Environment 143, 107–119. https://doi.org/10.1016/J.BUILDENV.2018.07.003

Efthimiou, G.C., Kovalets, I. V., Argyropoulos, C.D., Venetsanos, A., Andronopoulos, S., Kakosimos, K.E., 2018b. Evaluation of an inverse modelling methodology for the

prediction of a stationary point pollutant source in complex urban environments. Building and Environment 143, 107–119. https://doi.org/10.1016/J.BUILDENV.2018.07.003

Fuentes, M., Chaudhuri, A., Holland, D.M., 2007. Bayesian entropy for spatial sampling design of environmental data. Environmental and Ecological Statistics 14, 323–340. https://doi.org/10.1007/s10651-007-0017-0

Gelfand, A.E., Sahu, S.K., 1994. On markov chain monte carlo acceleration. Journal of Computational and Graphical Statistics 3, 261–276. https://doi.org/10.1080/10618600.1994.10474644

Gilks, W.R., Richardson, S., Spiegelhalter, D., 1995. Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Graham, J., Kanov, K., Yang, X.I.A., Lee, M., Malaya, N., Lalescu, C.C., Burns, R., Eyink, G., Szalay, A., Moser, R.D., Meneveau, C., 2016. A web services accessible database of turbulent channel flow and its use for testing a new integral wall model for LES. Journal of Turbulence 17, 181–215. https://doi.org/10.1080/14685248.2015.1088656

Gromke, C., Buccolieri, R., Di Sabatino, S., Ruck, B., 2008. Dispersion study in a street canyon with tree planting by means of wind tunnel and numerical investigations – Evaluation of CFD data with experimental data. Atmospheric Environment 42, 8640–8650. https://doi.org/10.1016/J.ATMOSENV.2008.08.019

Hadjidoukas, P., Wermelinger, F., 2019. A Parallel Data Compression Framework for Large Scale 3D Scientific Data.

Hastings, W.K., 1970. Monte carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109. https://doi.org/10.1093/biomet/57.1.97

Hellsten, A., Luukkonen, S.M., Steinfeld, G., Kanani-Sühring, F., Markkanen, T., Järvi, L., Lento, J., Vesala, T., Raasch, S., 2015. Footprint Evaluation for Flux and Concentration Measurements for an Urban-Like Canopy with Coupled Lagrangian Stochastic and Large-Eddy Simulation Models. Boundary-Layer Meteorology 157, 191–217. https://doi.org/10.1007/s10546-015-0062-4

Hutchinson, M., Oh, H., Chen, W.-H., 2017. A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors. Information Fusion

36, 130–148. https://doi.org/10.1016/j.inffus.2016.11.010

Ikegaya, N., Ikeda, Y., Hagishima, A., Razak, A.A., Tanimoto, J., 2017. A prediction model for wind speed ratios at pedestrian level with simplified urban canopies. Theoretical and Applied Climatology 127, 655–665. https://doi.org/10.1007/s00704-015-1655-z

Ikegaya, N., Okaze, T., Kikumoto, H., Imano, M., Ono, H., Tominaga, Y., 2019. Effect of the numerical viscosity on reproduction of mean and turbulent flow fields in the case of a 1:1:2 single block model. Journal of Wind Engineering and Industrial Aerodynamics 191, 279–296. https://doi.org/10.1016/J.JWEIA.2019.06.013

Issartel, J.-P., 2005a. Emergence of a tracer source from air concentration measurements, a new strategy for linear assimilation. Atmos. Chem. Phys. 5, 249–273. https://doi.org/10.5194/acp-5-249-2005

Issartel, J.-P., 2005b. Emergence of a tracer source from air concentration measurements, a new strategy for linear assimilation. Atmospheric Chemistry and Physics 5, 249–273. https://doi.org/10.5194/acp-5-249-2005

Issartel, J.-P., 2003. Rebuilding sources of linear tracers after atmospheric concentration measurements. Atmos. Chem. Phys. 3, 2111–2125. https://doi.org/10.5194/acp-3-2111-2003

Jacob, J., Sagaut, P., 2018. Wind comfort assessment by means of large eddy simulation with lattice Boltzmann method in full scale city area. Building and Environment 139, 110–124. https://doi.org/10.1016/j.buildenv.2018.05.015

Kanov, K., Burns, R., Lalescu, C., Eyink, G., 2015. The Johns Hopkins Turbulence Databases: An Open Simulation Laboratory for Turbulence Research. Computing in Science and Engineering 17, 10–17. https://doi.org/10.1109/MCSE.2015.103

Keats, A., Yee, E., Lien, F.-S., 2010. Information-driven receptor placement for contaminant source determination. Environmental Modelling & Software 25, 1000–1013. https://doi.org/10.1016/j.envsoft.2010.01.006

Keats, A., Yee, E., Lien, F.-S., 2007a. Bayesian inference for source determination with applications to a complex urban environment. Atmospheric Environment 41, 465–479. https://doi.org/10.1016/j.atmosenv.2006.08.044

Keats, A., Yee, E., Lien, F.-S., 2007b. Bayesian inference for source determination with

applications to a complex urban environment. Atmospheric Environment 41, 465–479. https://doi.org/10.1016/j.atmosenv.2006.08.044

Kikumoto, H., Ooka, R., 2012. A numerical study of air pollutant dispersion with bimolecular chemical reactions in an urban street canyon using large-eddy simulation. Atmospheric Environment 54, 456–464. https://doi.org/10.1016/j.atmosenv.2012.02.039

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by Simulated Annealing. Science 220, 671–680. https://doi.org/10.1126/science.220.4598.671

Kolomenskiy, D., Onishi, R., Uehara, H., 2018. Data Compression for Environmental Flow Simulations.

Kopka, P., Wawrzynczak, A., Borysiewicz, M., 2016. Application of the Approximate Bayesian Computation methods in the stochastic estimation of atmospheric contamination parameters for mobile sources. Atmospheric Environment 145, 201–212. https://doi.org/10.1016/J.ATMOSENV.2016.09.029

Kormi, T., Mhadhebi, S., Bel Hadj Ali, N., Abichou, T., Green, R., 2018. Estimation of fugitive landfill methane emissions using surface emission monitoring and Genetic Algorithms optimization. Waste Management 72, 313–328. https://doi.org/10.1016/j.wasman.2016.11.024

Kouichi, H., Ngae, P., Kumar, P., Feiz, A.-A., Bekka, N., 2019. An optimization for reducing the size of an existing urban-like monitoring network for retrieving an unknown point source emission. Geoscientific Model Development 12, 3687–3705. https://doi.org/10.5194/gmd-12-3687-2019

Kumar, P., Feiz, A.-A., Singh, S.K., Ngae, P., Turbelin, G., 2015a. Reconstruction of an atmospheric tracer source in an urban-like environment. Journal of Geophysical Research: Atmospheres 120, 12589–12604. https://doi.org/10.1002/2015JD024110

Kumar, P., Feiz, A.-A., Singh, S.K., Ngae, P., Turbelin, G., 2015b. Reconstruction of an atmospheric tracer source in an urban-like environment. Journal of Geophysical Research: Atmospheres 120, 12589–12604. https://doi.org/10.1002/2015JD024110

Li, F., Liu, X., Liu, J., Cai, H., Wang, H., Zhang, K., Dai, C., 2020. Solutions to mitigate the impact of measurement noise on the air pollution source strength estimation in a multi-

zone building. Building Simulation 13, 1329–1337. https://doi.org/10.1007/s12273-020-0635-0

Li, F., Niu, J., 2005. An inverse approach for estimating the initial distribution of volatile organic compounds in dry building material. Atmospheric Environment 39, 1447–1455. https://doi.org/10.1016/j.atmosenv.2004.11.021

Li, S., 2015. Evaluating the Efficacy of Wavelet Compression for Turbulent-Flow Data Visualization.

Loredo, T.J., 2004. Bayesian Adaptive Exploration. arXiv 707, 330–346. https://doi.org/10.1063/1.1751377

Ma, D., Tan, W., Zhang, Z., Hu, J., 2017. Parameter identification for continuous point emission source based on Tikhonov regularization method coupled with particle swarm optimization algorithm. Journal of Hazardous Materials 325, 239–250. https://doi.org/10.1016/j.jhazmat.2016.11.071

Martin, G.N., 1979. Range encoding: an algorithm for removing redundancy from a digitised message, in: Video & Data Recording Confer- Ence. Southampton, UK.

Martina, M., Masera, G., 2005. Low-complexity, efficient 9/7 wavelet filters implementation, in: IEEE International Conference on Image Processing 2005. IEEE, p. III–1000. https://doi.org/10.1109/ICIP.2005.1530563

Meroney, R.N., Pavageau, M., Rafailidis, S., Schatzmann, M., 1996. Study of line source characteristics for 2-D physical modelling of pollutant dispersion in street canyons. Journal of Wind Engineering and Industrial Aerodynamics 62, 37–56. https://doi.org/10.1016/S0167-6105(96)00057-8

Mons, V., Margheri, L., Chassaing, J.-C., Sagaut, P., 2017. Data assimilation-based reconstruction of urban pollutant release characteristics. Journal of Wind Engineering and Industrial Aerodynamics 169, 232–250. https://doi.org/10.1016/j.jweia.2017.07.007

Naik, A.K., Holambe, R.S., 2014. New Approach to the Design of Low Complexity 9/7 Tap Wavelet Filters With Maximum Vanishing Moments. IEEE Transactions on Image Processing 23, 5722–5732. https://doi.org/10.1109/TIP.2014.2363733

Ngae, P., Kouichi, H., Kumar, P., Feiz, A.A., Chpoun, A., 2019. Optimization of an

urban monitoring network for emergency response applications: An approach for characterizing the source of hazardous releases. Quarterly Journal of the Royal Meteorological Society 145, 967–981. https://doi.org/10.1002/qj.3471

Nicas, M., 2000. Markov Modeling of Contaminant Concentrations in Indoor Air. AIHAJ - American Industrial Hygiene Association 61, 484–491. https://doi.org/10.1080/15298660008984559

Parent, A., Morin, M., Lavigne, P., 1992. Propagation of super-Gaussian field distributions. Optical and Quantum Electronics 24, S1071–S1079. https://doi.org/10.1007/BF01588606

Pudykiewicz, J.A., 1998. Application of adjoint tracer transport equations for evaluating source parameters. Atmospheric Environment 32, 3039–3050. https://doi.org/10.1016/S1352-2310(97)00480-9

Qiu, Z., Lee, C.M., Xu, Z.H., Sui, L.N., 2016. A multi-resolution filtered-x LMS algorithm based on discrete wavelet transform for active noise control. Mechanical Systems and Signal Processing 66–67, 458–469. https://doi.org/10.1016/j.ymssp.2015.05.024

Rajaona, H., Septier, F., Armand, P., Delignon, Y., Olry, C., Albergel, A., Moussafir, J., 2015. An adaptive Bayesian inference algorithm to estimate the parameters of a hazardous atmospheric release. Atmospheric Environment 122, 748–762. https://doi.org/10.1016/j.atmosenv.2015.10.026

Rioul, O., Vetterli, M., 1991. Wavelets and signal processing. IEEE Signal Processing Magazine 8, 14–38. https://doi.org/10.1109/79.91217

Ruttimann, U.E., Pipberger, H. V., 1979. Compression of the ECG by Prediction or Interpolation and Entropy Encoding. IEEE Transactions on Biomedical Engineering BME-26, 613–623. https://doi.org/10.1109/TBME.1979.326543

Sakai, R., Sasaki, D., Obayashi, S., Nakahashi, K., 2013. Wavelet-based data compression for flow simulation on block-structured Cartesian mesh. International Journal for Numerical Methods in Fluids 73, 462–476. https://doi.org/10.1002/fld.3808

Salim, S.M., Buccolieri, R., Chan, A., Di Sabatino, S., 2011a. Numerical simulation of atmospheric pollutant dispersion in an urban street canyon: Comparison between RANS and

LES. Journal of Wind Engineering and Industrial Aerodynamics 99, 103–113. https://doi.org/10.1016/J.JWEIA.2010.12.002

Salim, S.M., Cheah, S.C., Chan, A., 2011b. Numerical simulation of dispersion in urban street canyons with avenue-like tree plantings: Comparison between RANS and LES. Building and Environment 46, 1735–1746. https://doi.org/10.1016/J.BUILDENV.2011.01.032

Saunier, O., Bocquet, M., Mathieu, A., Isnard, O., 2009. Model reduction via principal component truncation for the optimal design of atmospheric monitoring networks. Atmospheric Environment 43, 4940–4950. https://doi.org/10.1016/j.atmosenv.2009.07.011

Schmalzl, J., 2003. Using standard image compression algorithms to store data from computational fluid dynamics. Computers and Geosciences 29, 1021–1031. https://doi.org/10.1016/S0098-3004(03)00098-0

Schmid, P.J., 2010. Dynamic mode decomposition of numerical and experimental data. Journal of Fluid Mechanics 656, 5–28. https://doi.org/10.1017/S0022112010001217

Sharan, M., Singh, S.K., Issartel, J.P., 2012. Least Square Data Assimilation for Identification of the Point Source Emissions. Pure Appl. Geophys. 169, 483–497. https://doi.org/10.1007/s00024-011-0382-3

Skodras, A., Christopoulos, C., Ebrahimi, T., 2001. The JPEG 2000 still image compression standard. IEEE Signal Processing Magazine 18, 36–58. https://doi.org/10.1109/79.952804

Song, M.-S., 2008. Entropy Encoding in Wavelet Image Compression, in: Jorgensen, P.E.T., Merrill, K.D., Packer, J.A. (Eds.), Representations, Wavelets, and Frames. Birkhäuser Boston, Boston, MA, pp. 293–311. https://doi.org/10.1007/978-0-8176-4683-7_14

Thomson, L.C., Hirst, B., Gibson, G., Gillespie, S., Jonathan, P., Skeldon, K.D., Padgett, M.J., 2007. An improved algorithm for locating a gas source using inverse methods. Atmospheric Environment 41, 1128–1134. https://doi.org/10.1016/j.atmosenv.2006.10.003

Tominaga, Y., Mochida, A., Murakami, S., Sawaki, S., 2008a. Comparison of various revised k-ε models and LES applied to flow around a high-rise building model with 1:1:2 shape placed within the surface boundary layer. Journal of Wind Engineering and Industrial

Aerodynamics 96, 389–411. https://doi.org/10.1016/j.jweia.2008.01.004

Tominaga, Y., Mochida, A., Yoshie, R., Kataoka, H., Nozu, T., Yoshikawa, M., Shirasawa, T., 2008b. AIJ guidelines for practical applications of CFD to pedestrian wind environment around buildings. Journal of Wind Engineering and Industrial Aerodynamics 96, 1749–1761. https://doi.org/10.1016/j.jweia.2008.02.058

Tominaga, Y., Stathopoulos, T., 2013. CFD simulation of near-field pollutant dispersion in the urban environment: A review of current modeling techniques. Atmospheric Environment 79, 716–730. https://doi.org/10.1016/j.atmosenv.2013.07.028

Tominaga, Y., Stathopoulos, T., 2012. CFD Modeling of Pollution Dispersion in Building Array: Evaluation of turbulent scalar flux modeling in RANS model using LES results. Journal of Wind Engineering and Industrial Aerodynamics 104–106, 484–491. https://doi.org/10.1016/J.JWEIA.2012.02.004

Tominaga, Y., Stathopoulos, T., 2011a. CFD modeling of pollution dispersion in a street canyon: Comparison between LES and RANS. Journal of Wind Engineering and Industrial Aerodynamics 99, 340–348. https://doi.org/10.1016/J.JWEIA.2010.12.005

Tominaga, Y., Stathopoulos, T., 2011b. CFD modeling of pollution dispersion in a street canyon: Comparison between LES and RANS. Journal of Wind Engineering and Industrial Aerodynamics 99, 340–348. https://doi.org/10.1016/J.JWEIA.2010.12.005

Uehara, K., Murakami, S., Oikawa, S., Wakamatsu, S., 2000. Wind tunnel experiments on how thermal stratification affects flow in and above urban street canyons. Atmospheric Environment 34, 1553–1562. https://doi.org/10.1016/S1352-2310(99)00410-0

Usevitch, B.E., 2001. A tutorial on modern lossy wavelet image compression: Foundations of JPEG 2000. IEEE Signal Processing Magazine 18, 22–35. https://doi.org/10.1109/79.952803

Villasenor, J.D., Belzer, B., Liao, J., 1995. Wavelet Filter Evaluation for Image Compression. IEEE Transactions on Image Processing 4, 1053–1060. https://doi.org/10.1109/83.403412

Vrugt, J.A., ter Braak, C.J.F., Diks, C.G.H., Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with

Self-Adaptive Randomized Subspace Sampling. International Journal of Nonlinear Sciences and Numerical Simulation 10, 1–4. https://doi.org/10.1515/IJNSNS.2009.10.3.273

Wade, D., Senocak, I., 2013a. Stochastic reconstruction of multiple source atmospheric contaminant dispersion events. Atmospheric Environment 74, 45–51. https://doi.org/10.1016/J.ATMOSENV.2013.02.051

Wade, D., Senocak, I., 2013b. Stochastic reconstruction of multiple source atmospheric contaminant dispersion events. Atmospheric Environment 74, 45–51. https://doi.org/10.1016/J.ATMOSENV.2013.02.051

Wang, J., Wang, B., Liu, J., Cheng, W., Zhang, J., 2021. An inverse method to estimate the source term of atmospheric pollutant releases. Atmospheric Environment 118554. https://doi.org/10.1016/j.atmosenv.2021.118554

Wang, Y., Huang, H., Huang, L., Zhang, X., 2018. Source term estimation of hazardous material releases using hybrid genetic algorithm with composite cost functions. Engineering Applications of Artificial Intelligence 75, 102–113. https://doi.org/10.1016/J.ENGAPPAI.2018.08.005

Wilke, C.R., Lee, C.Y., 1955. Estimation of Diffusion Coefficients for Gases and Vapors. Industrial & Engineering Chemistry 47, 1253–1257. https://doi.org/10.1021/ie50546a056

Wilson, J.P., 2002. Wavelet-based lossy compression of barotropic turbulence simulation data, in: Proceedings DCC 2002. Data Compression Conference. pp. 479-. https://doi.org/10.1109/DCC.2002.1000022

Wu, L., Bocquet, M., 2011. Optimal redistribution of the background ozone monitoring stations over France. Atmospheric Environment 45, 772–783. https://doi.org/10.1016/j.atmosenv.2010.08.038

Xie, Z., Castro, I.P., 2006. LES and RANS for turbulent flow over arrays of wall-mounted obstacles. Flow, Turbulence and Combustion 76, 291–312. https://doi.org/10.1007/s10494-006-9018-6

Xue, F., Kikumoto, H., Li, X., Ooka, R., 2018a. Bayesian source term estimation of atmospheric releases in urban areas using LES approach. Journal of Hazardous Materials 349,

68–78. https://doi.org/10.1016/j.jhazmat.2018.01.050

Xue, F., Kikumoto, H., Li, X., Ooka, R., 2018b. Bayesian source term estimation of atmospheric releases in urban areas using LES approach. Journal of Hazardous Materials 349, 68–78. https://doi.org/10.1016/j.jhazmat.2018.01.050

Xue, F., Li, X., Ooka, R., Kikumoto, H., Zhang, W., 2017. Turbulent Schmidt number for source term estimation using Bayesian inference. Building and Environment 125, 414–422. https://doi.org/10.1016/j.buildenv.2017.09.012

Yakhot, V., Orszag, S.A., Thangam, S., Gatski, T.B., Speziale, C.G., 1992. Development of turbulence models for shear flows by a double expansion technique. Physics of Fluids A 4, 1510–1520. https://doi.org/10.1063/1.858424

Yee, E., 2012. Probability theory as logic: Data assimilation for multiple source reconstruction. Pure and Applied Geophysics 169, 499–517. https://doi.org/10.1007/s00024-011-0384-1

Yee, E., Hoffman, I., Ungar, K., 2014. Bayesian Inference for Source Reconstruction: A Real-World Application. International Scholarly Research Notices 2014, 1–12. https://doi.org/10.1155/2014/507634

Zarzhitsky, D., Spears, D.F., 2005. Swarm approach to chemical source localization, in: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics.

Zheng, X., Chen, Z., 2010. Back-calculation of the strength and location of hazardous materials releases using the pattern search method. Journal of Hazardous Materials 183, 474–481. https://doi.org/10.1016/j.jhazmat.2010.07.048

Zubair, L., Sreenivasan, K.R., Wickerhauser, M.V., 1992. Characterization and Compression of Turbulent Signals and Images Using Wavelet-Packets, in: Gatski, T.B., Speziale, C.G., Sarkar, S. (Eds.), Studies in Turbulence. Springer New York, New York, NY, pp. 489–513. https://doi.org/10.1007/978-1-4612-2792-2_37

Appendix

Appendix A. Numerical simulation for turbulence and dispersion of pollution

A.1 Fluid and turbulence

Atmospheric pollution is caused by the dispersion of pollutants in the air, which is a mass transfer behavior in fluids. Fluids like gases and liquids are defined as substances that cannot permanently keep their shapes under certain stress. Fluids also do not have a definite shape. They can be divided into Newtonian or non-Newtonian based on the relationship between the shear stress and shear rate. Most gases are Newtonian whose relationship is linear with the slope of molecular viscosity. Hence, only the Newtonian fluids are discussed in what follows.

Fluid flow has two states: laminar and turbulent. According to the fluid experiments, a key factor to decide the state of the fluid is Reynolds number $Re = \frac{UL}{\nu}$. Here ν is the dynamic viscosity coefficient. U is the standard velocity of the flow. L is the standard length of the flow. Hence, Re measures the relative power of inertia forces and viscous force on the state of fluids. In fact, for each kind of fluid, there exists a critical Reynolds number Re_c. When Re is smaller than Re_c, the flow is smooth and adjacent parts slide past each other in order. If the boundary condition of the flow does not change with time, the flow will keep the steady state everywhere in the domain and it is called laminar flow. If Re gradually increases after it reaches Re_c, the laminar flow will go through an important state called transition, where complicated events happen to change the flow into turbulence. In this state, turbulent structures with different sizes appear spatially and temporally. They include the information of nonlinear correlation among a certain space of the flow. This nonlinearity is mathematically represented by the advection term (see Eq. (A.2)). However, if a single point was monitored, the flow velocity there is totally random and chaotic. It is difficult to build a physical system for such a flow and find the mathematical description of it.

From a microcosmic point of view, fluid is a system consisting of numerous small molecular even though it is a continuum in the human's eyes. In order to describe the system in

mathematics, Eulerian and Lagrangian approaches were used. In the Lagrangian approach, the fluid system is divided into small fluid parcels. When the system is changing through space and time, the location and velocity of each parcel are followed to describe the system. In contrast, the Eulerian approach focuses on the properties of specific locations in the domain. Its view will not follow any fluid parcels, but it cares about the flow velocity somewhere as time changes. Although the two approaches are based on totally different concepts, they are both effective in the mathematical description of fluids. In this thesis, fluids are processed from the Eulerian view, and the corresponding theories are briefly introduced here.

A.2 Governing equations for fluid and dispersion

In this study, the dispersion of pollution is assumed to occur in an incompressible, isothermal fluid flow. Besides, the pollutants are assumed as the passive scalar which will not affect the behavior of fluids. Under this circumstance, the Eulerian descriptions of the fluids and the dispersion are the following three equations: continuity equation, momentum equation (or Navier-Stokes equation), and transport equation.

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{A.1}$$

$$\frac{\partial u_i}{\partial t} + \frac{\partial u_j u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\left\{\nu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right\} \tag{A.2}$$

$$\frac{\partial C}{\partial t} + \frac{\partial u_j C}{\partial x_j} = \frac{\partial}{\partial x_j}\left(D_m \frac{\partial C}{\partial x_j}\right) + S \tag{A.3}$$

Here, u_i is the instantaneous velocity in the x_i direction. ρ is the constant density of fluids. p is the instantaneous pressure. ν is the dynamic viscosity coefficient. C is the concentration. D_m is the molecular diffusion coefficient. S is the source term of the passive scalar. Eq. (A.1) follows the conservation rule of mass. Eq. (A.2) follows the conservation rule of momentum. Eq. (A.3) describes the transport process of passive scalar released from the source.

Although the governing equations for fluids and dispersion have been established, and the number of unknown variables equals that of available equations, it is still one of the most challenging mathematical problems to find the analytical solution for them. The reason is that Navier-Stokes equation is a non-linear partial differential equation. The second term in the left hand side of Eq. (A.2), which is called as advection term, represents the nonlinear coupling of velocities in two directions.

In this case, a compromise approach is to solve it by a numerical method like the finite volume method. In this method, the calculation domain is subdivided into finite small, discrete volumes by meshing, then these partial differential equations are transferred into a huge set of algebraic equations by integration for each volume. The numerical solutions can be calculated by simultaneously solving this set of equations for a short time step with time marching. It is straightforward to notice that the success of numerical simulation needs a considerable amount of computational resources for the iterative calculation of algebraic equations.

In recent years, the development of high performance computers makes numerical simulation realizable. The computational fluid dynamics (CFD) technique has become a powerful tool in industrial applications and research of fluid mechanisms. Therefore, this thesis models the source-receptor relationship and pollution dispersion by CFD.

In theory, CFD can provide a nearly perfect solution for Eqs. (A.1-A.3), and such technique does exist with a name of direct numerical simulation (DNS). However, the reality is not so easy.

On the one hand, in the turbulent flow field of the atmospheric environment, the range of sizes of turbulence structures is considerably wide. During the movement of the flow field, the large turbulence will break into smaller turbulence. The kinetic energy is also transferred into smaller turbulence, which is called the energy cascade. This process will continue to pass the energy to smaller turbulence until it reaches the smallest size and eventually dissipates in the form of thermal energy caused by the molecular viscosity. Therefore, in order to completely simulate the turbulence flow, it is necessary to reproduce all the structures from the largest ones to the smallest ones.

On the other hand, in CFD, the partial differential equations are integrated into algebraic equations for each mesh (volume). In this case, the turbulence structures smaller than the mesh cannot be simulated. The mesh should be made extremely fine to capture all the structures. According to the previous research, the mesh number needs to have the order of $Re^{9/4}$. A common case in wind engineering can be checked here. If the flow field around a building with $L = 15m$ height caused by a coming flow with $U = 1m/s$ is simulated by DNS, with the dynamic viscosity coefficient of air $\nu = 1.5 \times 10^{-6} m^2/s$, Re is 10^6 and the number of mesh astonishingly becomes about 3.1×10^{13}. The resultant calculation burden is unaffordable even with the most advanced computers right now. Therefore, DNS has only been

used to simulate the flow field with a small Reynolds number. Most engineering simulations are constrained.

Since a complete simulation of all turbulence structures with CFD is almost impossible for high Reynolds number, in order to realize the CFD for engineering applications, alternative approaches have been developed by researchers. One common method is that rather than all turbulence, only the turbulence of interest with relatively large sizes are explicitly simulated, while the small turbulence are modeled. This is so-called turbulence modeling. In the past decades, among plenty of modeling methods, two popular turbulence models are the Reynolds averaged Navier-Stokes (RANS) model and the large-eddy simulation (LES) model. These two models are also used in this thesis and will be introduced in the following contents.

A.3 Large eddy simulation

A.3.1 Filtering process

The principal idea behind LES is to reduce the computational cost by modeling the small turbulence structures, which are the most computationally expensive to resolve, via low-passing filtering of the Navier-Stokes equations. In other words, for any wanted physical variables $f(x,t)$, LES will decompose it into the grid scale (GS) component $\tilde{f}(x,t)$ which can be explicitly resolved by the mesh, and the sub-grid scale (SGS) component $f''(x,t)$ which is filtered out and implicitly modeled.

$$f(x,t) = \tilde{f}(x,t) + f''(x,t) \tag{A.4}$$

The separation for GS and SGS components is operated by filter function $G(\xi)$, which is defined by:

$$\tilde{f}(x,t) = \int_{-\infty}^{\infty} G(\xi) f(x-\xi)\, d\xi \tag{A.5}$$

The filter function has to satisfy the following requirements:

$$\lim_{\xi \to \pm\infty} G(\xi) = 0 \tag{A.6}$$

$$\int_{-\infty}^{\infty} G(\xi)\, d\xi = 1 \tag{A.7}$$

Common filter functions in LES include top hat filter, Gaussian filter, and sharp cut-off filter. After the filtering, the filtered momentum equation becomes:

$$\frac{\partial \widetilde{u_i}}{\partial t} + \frac{\partial \widetilde{u_i u_j}}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \tilde{p}}{\partial x_i} + \frac{1}{\nu}\frac{\partial^2 \widetilde{u_i}}{\partial x_j x_j} \tag{A.8}$$

A.3.2 Sub-grid scaled stress modeling

The form of Eq. (A.8) is different from that of Eq. (A.2) because the filtered product $\widetilde{u_i u_j}$ is different from the product of the filtered velocity $\widetilde{u_i}\widetilde{u_j}$. The difference is the residual-stress tensor defined by

$$T_{ij} = \widetilde{u_i u_j} - \widetilde{u_i}\widetilde{u_j} \tag{A.9}$$

This tensor is also called the SGS stress tensor. In this research, it is modeled by the standard Smagorinsky SGS model, according to which, the SGS tensor is split into an isotropic part $\frac{1}{3}T_{kk}\delta_{ij}$ and an anisotropic part $T_{ij} - \frac{1}{3}T_{kk}\delta_{ij}$, where δ_{ij} is the Kronecker delta.

$$\begin{aligned} T_{ij} &= \widetilde{u_i u_j} - \widetilde{u_i}\widetilde{u_j} \\ &= \frac{1}{3}T_{kk}\delta_{ij} + \left(T_{ij} - \frac{1}{3}T_{kk}\delta_{ij}\right) \\ &\approx \frac{1}{3}T_{kk}\delta_{ij} - 2\nu_{sgs}dev(\widetilde{D})_{ij} \\ &= \frac{2}{3}k_{sgs}\delta_{ij} - 2\nu_{sgs}dev(\widetilde{D})_{ij} \end{aligned} \tag{A.10}$$

Here ν_{sgs} is the SGS eddy viscosity. The resolved-scale strain rate tensor $\widetilde{D}_{ij}$ is defined as

$$\widetilde{D}_{ij} = \frac{1}{2}\left(\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i}\right) \tag{A.11}$$

The SGS kinetic energy k_{sgs} is

$$k_{sgs} = \frac{1}{2}T_{kk} = \frac{1}{2}\left(\widetilde{u_i u_j} - \widetilde{u_i}\widetilde{u_j}\right) \tag{A.12}$$

The Smagorinsky SGS model is based on two elementary assumptions: eddy viscosity approximation and local equilibrium. In the first assumption, the anisotropic part in equation (A.10) is approximated by relating it to the resolved rate of the strain tensor $\widetilde{D}_{ij}$.

$$T_{ij} - \frac{1}{3}T_{kk}\delta_{ij} \approx -2\nu_{sgs}dev(\widetilde{D})_{ij} \tag{A.13}$$

The SGS scale viscosity is computed as:

$$\nu_{sgs} = C_k\Delta\sqrt{k_{sgs}} \tag{A.14}$$

Here C_k is a model constant with a default value of 0.094. Δ is the cubic root of the cell volume that defines the sub-grid length scale.

In the second assumption about local equilibrium, there is a balance between the sub-grid scale energy production and dissipation. The SGS kinetic energy k_{sgs} is calculated by this assumption.

$$\begin{gathered}
\widetilde{D}:T_{ij} + C_e\frac{{k_{sgs}}^{1.5}}{\Delta} = 0 \\
\widetilde{D}:(\frac{2}{3}k_{sgs}I - 2\nu_{sgs}dev(\widetilde{D})) + C_e\frac{{k_{sgs}}^{1.5}}{\Delta} = 0 \\
\widetilde{D}:(\frac{2}{3}k_{sgs}I - 2C_k\Delta\sqrt{k_{sgs}}dev(\widetilde{D})) + C_e\frac{{k_{sgs}}^{1.5}}{\Delta} = 0 \\
\sqrt{k_{sgs}}\left(\frac{C_e}{\Delta}k_{sgs} + \frac{2}{3}tr(\widetilde{D})\sqrt{k_{sgs}} - 2C_k\Delta\big(dev(\widetilde{D}):\widetilde{D}\big)\right) = 0 \\
ak_{sgs} + b\sqrt{k_{sgs}} - c = 0 \\
k_{sgs} = (\frac{-b + \sqrt{b^2 + 4ac}}{2a})^2
\end{gathered} \tag{A.15}$$

The operator **:** is a double inner product of two second-rank tensors which is the sum of the 9 products of the tensor components. I is the identity matrix. In Eq. (A.15),

$$a = \frac{C_e}{\Delta}$$
$$b = \frac{2}{3} tr(\widetilde{D}) \tag{A.16}$$
$$c = 2C_k \Delta(dev(\widetilde{D}) : \widetilde{D})$$

In the case of incompressible flow, which is the same as this research, the Equation (A.16) reduces to

$$b = \frac{2}{3} tr(\widetilde{D}) = 0$$
$$c = 2C_k \Delta\left(dev(\widetilde{D}) : \widetilde{D}\right) = C_k \Delta \left|\widetilde{D}\right|^2 \tag{A.17}$$

where

$$\left|\widetilde{D}\right| = \sqrt{2\widetilde{D} : \widetilde{D}} \tag{A.18}$$

By substituting the Eq. (A.17) into Eq. (A.15), we have:

$$k_{sgs} = \frac{c}{a} = \frac{C_k \Delta^2 \left|\widetilde{D}\right|^2}{C_e} \tag{A.19}$$

The following expression can be obtained for the SGS eddy viscosity in the case of incompressible flows by substituting the Equation (A.19) into the Equation (A.14).

$$\nu_{sgs} = C_k \sqrt{\frac{C_k}{C_e}} \Delta^2 \left|\widetilde{D}\right| \tag{A.20}$$

In the literature, the SGS eddy viscosity is commonly expressed as below.

$$\nu_{sgs} = (C_s \Delta)^2 \left|\widetilde{D}\right| \tag{A.21}$$

We can get the relation between the Smagorinsky constant C_s and other coefficients.

$$C_s{}^2 = C_k \sqrt{\frac{C_k}{C_e}} \tag{A.22}$$

The default value of the Smagorinsky constant is 0.173. For the flow field around the solid structure, this constant is commonly set to 0.10~0.15. In this research, it was set as 0.12.

A.3.3 Van driest damping function

In the simulation, if the boundary condition is the non-slip wall, there should be no turbulence at the boundary. However, in the standard SGS model, the eddy viscosity is nonzero, which is contrary to reality. To fix this problem, one way is to add a Van Driest damping function into the length scale. The SGS eddy viscosity changes to:

$$\nu_{sgs} = (C_s f_s \Delta)^2 |\bar{D}| \tag{A.23}$$

Here, the f_s is the Van Driest damping function.

$$f_s = 1 - exp\frac{-y^+}{A^+} \tag{A.24}$$

Here, A^+ is the Van Driest constant, which is set to 26 in this research. $y^+ = yu_\tau/\nu$ is the non-dimensional wall unit. u_τ is the shear velocity. In other words, the characteristic spatial length of the filter in the turbulent boundary layer is not necessarily related to the mesh size, but the minimum value between Δ and the one obtained from the damping function in equation (A.24) is locally adapted in space and time.

A.3.4 Dispersion equation for passive scalar

After the filtering and modeling for the momentum equation, the transport equation for passive scalar in LES can be dealt with in a similar way. By filtering operation, Eq. (3) becomes:

$$\frac{\partial \tilde{C}}{\partial t} + \frac{\partial \widetilde{u_j C}}{\partial x_j} = \frac{\partial}{\partial x_j}\left(D_m \frac{\partial \tilde{C}}{\partial x_j}\right) + S \tag{A.25}$$

The SGS turbulent diffusion term caused by filtering will also appear as:

$$\widetilde{u_i'' C''} = \widetilde{u_i C} - \tilde{u}_i \tilde{C} \tag{A.26}$$

One of the most common ways to model this term is relating it to the gradient of GS concentration:

$$\widetilde{u_i'' C''} \approx -D_{sgs} \frac{\partial(\tilde{C})}{\partial x_j} \approx -\frac{\nu_{sgs}}{Sc_{sgs}} \frac{\partial(\tilde{C})}{\partial x_j} \tag{A.27}$$

where Sc_{sgs} is the SGS turbulent Schmidt number. When the grid is fine enough, most of the turbulent diffusion is directly calculated and the modeled SGS is limited.

A.4 Reynolds averaged Navier-Stokes model

A.4.1 Reynolds average

Another turbulence model RANS can also be regarded as a filtering approach, but unlike LES where the filtering will change with the mesh size, the filtering in RANS is the ensemble average, which is constant in the space. The physical variable of flow f can be divided into ensemble-averaged component $\bar{f}$ and fluctuation component f'.

$$f(x,t) = \bar{f}(x,t) + f'(x,t) \tag{A.28}$$

Meanwhile, for the convenience of mathematical processing, it is required that

$$\bar{\bar{f}} = \bar{f}, \qquad \overline{f'} = 0, \qquad \overline{f'\bar{f}} = 0 \tag{A.29}$$

The ensemble average that satisfies the above relationships is called as Reynolds average.

If Reynolds average is operated on the continuity equation and momentum equation, they change to

$$\frac{\partial \overline{u_\iota}}{\partial x_i} = 0 \tag{A.30}$$

$$\frac{\partial \overline{u_\iota}}{\partial t} + \frac{\partial \overline{u_j}\,\overline{u_\iota}}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \bar{p}}{\partial x_i} + \frac{\partial}{\partial x_j}\left\{\nu\left(\frac{\partial \overline{u_\iota}}{\partial x_j} + \frac{\partial \overline{u_j}}{\partial x_i}\right)\right\} - \frac{\partial \tau_{ij}}{\partial x_j} \tag{A.31}$$

The extra term τ_{ij} is the Reynolds stress resulting from the Reynolds average.

$$\tau_{ij} = \overline{u_\iota' u_j'} = \overline{u_\iota u_j} - \overline{u_j}\,\overline{u_\iota} \tag{A.32}$$

It can be noticed that in order to simulate Eq. (A.30 & A.31) without closure problem, the Reynolds stress needs to be modeled, which is similar to the SGS stress tensor. Several approaches have been developed to model Reynolds stress under the RANS framework, as an instance, here one of the most used models standard $k-\varepsilon$ model is introduced.

A.4.2 Standard $k-\varepsilon$ model

The Standard $k-\varepsilon$ model is based on the assumption that there is an analogy between viscous stress and Reynold stress. In 1877, Boussinesq proposed that the Reynolds stress could be expressed as stress that is proportional to the mean rate of deformation.

$$\tau_{ij} = \nu_t \left(\frac{\partial \overline{u_\iota}}{\partial x_j} + \frac{\partial \overline{u_j}}{\partial x_i} \right) - \frac{2}{3} k \delta_{ij} \tag{A.33}$$

where ν_t is the dynamic turbulent viscosity coefficient. k is the turbulent kinetic energy defined as $\frac{1}{2}\left(\overline{{u_x}'^2} + \overline{{u_y}'^2} + \overline{{u_z}'^2}\right)$ or $\frac{1}{2}\tau_{ii}$. In the next step, it is necessary to provide the modeling expression for ν_t and k.

In fact, the governing equation for k can be derived. First, we multiple the Navier-Stokes equation Eq. (A.2) with the fluctuation velocity components and add them together. Then, the same process can be repeated again on the averaged equations Eq. (A.31). At last, we subtract the resultant two summed equations. Because the process is too long, only the final result is presented here:

$$\frac{\partial k}{\partial t} + \frac{\partial \overline{u_j} k}{\partial x_j} = P_{ij} - \varepsilon_{ij} + \mathbb{T}^{p}_{ijk} + \mathbb{T}^{v}_{ijk} + \mathbb{T}^{R}_{ijk} \tag{A.34a}$$

$$P_{ij} = \tau_{ij} \cdot S_{ij} \tag{A.34b}$$

$$\varepsilon_{ij} = \nu \overline{{s_{\iota j}}' \cdot {s_{\iota j}}'} \tag{A.34c}$$

$$\mathbb{T}^{p}_{ijk} = -\frac{1}{\rho}\left(\overline{p' {u_\iota}'}\delta_{jk} + \overline{p' {u_j}'}\delta_{ik}\right) \tag{A.34d}$$

$$\mathbb{T}^{v}_{ijk} = \nu \frac{\partial \tau_{ij}}{\partial x_k} \tag{A.34e}$$

$$\mathbb{T}^{R}_{ijk} = -\overline{{u_\iota}' {u_j}' {u_k}'} \tag{A.34f}$$

$$S_{ij} = \frac{\partial \overline{u_\iota}}{\partial x_j} + \frac{\partial \overline{u_j}}{\partial x_i}, \qquad s_{ij} = \frac{\partial {u_i}'}{\partial x_j} + \frac{\partial {u_j}'}{\partial x_i} \tag{A.34g}$$

Here, P_{ij} is the rate of production of k. ε_{ij} is the rate of dissipation of k. $\mathbb{T}_{ijk}^{p}$ is the transport of k by pressure. $\mathbb{T}_{ijk}^{v}$ is the transport of k by viscous stresses. $\mathbb{T}_{ijk}^{R}$ is the transport of k by Reynolds stresses.

As for ν_t, it was modeled based on the dimensional analysis.

$$\nu_t = C_\mu \frac{k^2}{\varepsilon} \tag{A.35}$$

where C_μ is a dimensionless constant.

Noticing that $k = \frac{1}{2}\tau_{ii}$ and using Eq. (A.33), the production term for k can be expressed as:

$$P_{ij} = 2\nu_t S_{ij} \cdot S_{ij} \tag{A.36}$$

Besides, the transport effects of pressure and Reynolds stresses can be approximated together by

$$\mathbb{T}_{ijk}^{p} + \mathbb{T}_{ijk}^{R} = \frac{\nu_t}{\sigma_k}\frac{\partial k}{\partial x_j} \tag{A.37}$$

Substituting Eq. (A.36 & A.37) into Eq. (A.34a), the governing equation for k changes to

$$\frac{\partial k}{\partial t} + \frac{\partial \overline{u}_j k}{\partial x_j} = P_{ij} - \varepsilon_{ij} + \frac{\partial}{\partial x_j}\left\{\left(\nu + \frac{\nu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right\} \tag{A.38}$$

The dissipation term ε in the above equation represents the energy-flow rate from the large turbulence to small turbulence in the cascade, which means it is mainly determined in the large-scale motion. The governing equation of ε can also be derived for numerical simulation. However, the derivation process is based on the dissipative range which represents the smallest scale of motion in the flow field. As a result, it is recommended to check this equation from the empirical view that it follows a similar transport form of Eq. (A.38).

$$\frac{\partial \varepsilon}{\partial t} + \frac{\partial \overline{u}_j \varepsilon}{\partial x_j} = \left(C_{\varepsilon 1} P_{ij} - C_{\varepsilon 2}\varepsilon\right)\frac{\varepsilon}{k} + \frac{\partial}{\partial x_j}\left\{\left(\nu + \frac{\nu_t}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_j}\right\} \tag{A.39}$$

In fewer words, the RANS model uses Eq. (A.35, A.38, A.39) to calculate the turbulent viscosity coefficient, and substitute it into Eq. (A.33) to model the Reynolds stress, which satisfies the closure requirement of Eq. (A.30 & A.31). In these equations, in total five dimensionless coefficients exist to tune the performance of RANS. Their standard values are set by Launder and Sharma (1974) by data fitting for different turbulent flows.

$$C_\mu = 0.09, C_{\varepsilon 1} = 1.44, C_{\varepsilon 2} = 1.92, \sigma_k = 1.0, \sigma_\varepsilon = 1.3 \tag{A.40}$$

A.4.3 Dispersion equation for passive scalar

The Reynolds average can also be operated to passive scalar transport equation, Eq. (A.3) changes to

$$\frac{\partial \bar{C}}{\partial t} + \frac{\partial \overline{u_j}\bar{C}}{\partial x_j} = \frac{\partial}{\partial x_j}\left(D_m \frac{\partial \tilde{C}}{\partial x_j}\right) + S - \frac{\partial \overline{u_j{}'C'}}{\partial x_j} \tag{A.41}$$

The extra term $\overline{u_j{}'C'}$ is the Reynolds average of turbulent diffusion term which needs modeling. The most common way in RANS is the gradient diffusion hypothesis, which assumes that the turbulent diffusion is proportional to the gradient of the averaged concentration field.

$$\overline{u_j{}'C'} = -D_t \frac{\partial \bar{C}}{\partial x_j} = -\frac{\nu_t}{Sc_t}\frac{\partial \bar{C}}{\partial x_j} \tag{A.42}$$

where D_t is the turbulent diffusivity, and Sc_t is the turbulent Schmidt number.

A.5 Comparisons of two turbulence model

As two of the most popular approaches for turbulence modeling, RANS and LES have distinguishing features.

RANS focuses on the mean flow due to the Reynolds averaging processing. The unsteady effects of turbulent flow are also removed out from the averaging and approximated by the mean physical quantities. Therefore, the computational cost of RANS is mild and makes RANS widely used in engineering applications for the past decades. Despite that, its disadvantage is also obvious that the unsteady turbulence is unavailable. More importantly, since the effects of turbulence on the mean flow are also implicitly modeled, the accuracy of mean flow is not good enough in the flow fields like strong separation from the corner of the building.

In contrast, LES focuses on the relatively large-scaled turbulent structures with filtering techniques. It only models the SGS behavior of turbulence flow. As a result, when the mesh is fine enough, most dominant turbulence structures can be accurately and explicitly captured, which brings a more reliable simulation of the flow field than RANS. The corresponding cost is the high requirement on the computational resources. In order to promise the performance of LES, the mesh should be carefully designed with fine resolution, which needs additional resources apart from that for unsteady simulation caused by time marching. Meanwhile, due to a large amount of mesh, the storage pressure of LES results is also much heavier than RANS.

Nowadays, with the development of high performance computers and semiconductor products, the calculation burden of LES becomes lighter and the importance of accuracy becomes more desired. This is the reason that more and more engineering applications start to switch from RANS to LES, and the application of LES in the adjoint equation simulation is also part of the main work of this dissertation.

Appendix B. Publications related to dissertation

B.1 Journal publications

(with peer review)

1.1) Hongyuan Jia, Hideki Kikumoto, "Construction of urban turbulent flow database with wavelet-based compression: A study with large-eddy simulation of flow and dispersion in block-arrayed building group model", *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 208, 104433, 2021.

1.2) Hongyuan Jia, Hideki Kikumoto, "Line source estimation of environmental pollutants using super-Gaussian geometry model and Bayesian inference", *Environmental Research*, vol. 194, 110706, 2021.

1.3) Hongyuan Jia, Hideki Kikumoto, "Source term estimation in complex urban environments based on Bayesian inference and unsteady adjoint equations simulated via large eddy simulation", *Building and Environment*, vol. 193, 107669, 2021.

1.4) Hongyuan Jia, Hideki Kikumoto, "Sensor configuration optimization based on the entropy of adjoint concentration distribution for stochastic source term estimation in built environment", *Sustainable Cities and Society*, accepted.

(without peer review)

1.5) 賈 鴻源，菊本 英紀，「スーパーガウス関数とベイズ推定を用いた環境汚染物質線形発生源の同定」，生産研究，2020，72巻，1号，pp.49-55.

1.6) 賈 鴻源，菊本 英紀，「LESによる随伴濃度解析を用いた市環境汚染物質発生源のベイズ推定」，生産研究，2021，73巻，1号，pp.71-76.

B.2 Conference publications

(with peer review)

2.1) Hongyuan Jia, Hideki Kikumoto, "Line source estimation of environmental pollutants using Bayesian inference coupled with super-Gaussian geometry model", 15th RoomVent Conference, Torino, Italy, (2021).

(without peer review)

2.2)賈　鴻源，菊本　英紀 :「スーパーガウス幾何モデルを用いた環境汚染物質線形発生源のベイズ推定」，第 33 回数値流体力学シンポジウム，北海道，2019 年 11 月.

2.3)賈　鴻源，菊本　英紀: 「CFD を用いた市街地気流解析におけるデータ圧縮に関する研究 (その 1)ウェーブレット変換に基づいた圧縮手法の概要」，日本建築学会大会，千葉，2020 年 9 月.

2.4)胡 書媛，賈　鴻源，菊本　英紀: 「CFD を用いた市街地気流解析におけるデータ圧縮に関する研究 (その 2) 圧縮した LES 流れ場データを用いた汚染物質拡散の再解析」，日本建築学会大会，千葉，2020 年 9 月.

2.5)賈　鴻源, 菊本　英紀:「ウェーブレット変換に基づく圧縮手法を用いた市街地気流の LES データベースの構築」，日本流体力学会年会，山口，2020 年 9 月.

2.6)賈　鴻源, 菊本　英紀:「確率的発生源同定におけるセンサー配置最適化手法に関する研究 （その 1）随伴濃度分布のエントロピーに基づく最適化アルゴリズム」，日本建築学会大会，名古屋，2021 年 9 月.

2.7)李　栄茂，賈　鴻源，菊本　英紀: 「確率的発生源同定におけるセンサー配置最適化手法に関する研究 （その 2）室内モデル空間における随伴濃度分布とセンサー配置の最適化」，日本建築学会大会，名古屋，2021 年 9 月.

Acknowledgment

This dissertation is a summary of my research during three years as a doctoral candidate in the graduate school of engineering, the University of Tokyo. In this unforgettable period, I received a great deal of support and help from people around me, to whom I would like to express my sincere gratitude.

First and foremost, I would like to thank my supervisor Associate Professor Hideki Kikumoto. I feel extremely lucky to finish my doctoral degree under your supervision. I missed our long-lasting chatting in the office where I learned about how to start my career as a researcher. You gave me a lot of inspiring suggestions when I encountered difficulties. Any research idea was warmly welcomed by the professor even if it sounds immature. The joy of interest-driven research and insightful discussions provided in the Kikumoto Laboratory are the energetic fuels to this dissertation.

I would also like to express my deepest appreciation to my deputy supervisor Professor Ryozo Ooka. You gave me valuable advice on the general direction of this research. The late lighting of your office always reminds me of the importance of hard-working to success.

I must give my sincere gratitude to my doctoral advising committee member, Professor Shinichi Sakamoto, Associate Professor Yosuke Hasegawa, and Associate Professor Masayuki Mae. Their careful reviewing and helpful suggestions improve this dissertation with no doubt. Encouragement from Prof. Sakamoto gave me confidence to continue my research. Prof. Masayuki guided me from the structure of the dissertation. Prof. Hasegawa made this research more rigorous in mathematics and expressions.

I am also grateful to Mrs. Iizumi and Mrs. Honma, the secretaries of the laboratory. They handled the paperwork of my administrative procedures, which makes me focus on my research at ease.

Special thanks go to Associate Professor Mengtao Han in Huazhong University of Science and Technology. As an OB of the laboratory, he taught me the knowledge about Linux servers and OpenFOAM, which helped me smoothly start the research.

I would like to thank Mr. Chao Lin, who is not only a colleague but also a good friend in daily life. He gave me continuous support starting from my preparation for the entrance

examination. I always enjoyed chatting with him about research or other trivial things around.

With a special mention to Dr. Wonjun Choi, Dr. Wonseok Oh, Dr, Doyun Lee, Dr. Qi Zhou, Dr. Bingchao Zhang, Dr. Shan Gao, Dr. Mingzhe Liu, Dr. Huaiyu Zhong, Dr. Xiaochen Liu, Yuki Matsuda, Ke Wen, Christopher O'Malley, Chenghao Wei, Wenchao Wang, Rongmao Li, Hong Hu, Hiroyuki Ichikawa, Yunchen Bu, Chaoyi Hu, Xiang Wang, Shuyuan Hu, Ken Takahashi, Mayuka Nakai, Sorana Ozaki, Toshiki Nakamura, Eiki Sakamoto, Tomoya Oosaki, Ran Chen, Keisuke Ogasawara, Dun Zhu, Qianwen Guo in general. It was wonderful to have the opportunity to work with such a big group, which exposes me to a wide range of research topics.

I would like to acknowledge my best friend Dr. Zhipeng Zhang for his encouragement to my career pursuit and for accompanying me in the academic journey for over 13 years.

I must express my profound gratitude to my parents for providing me with financial support and understanding about my life choice. This dissertation cannot be finished without their love and care.

Finally, my appreciation goes to my wife, who studies abroad with me, always believes me to be able to accomplish this dissertation and comfort me at depressed moments. Over two of the past three years, I had to work from home because of COVID19, but your love made it the happiest time for both research and life. This dissertation will not be possible without you.