Enhancing CFD-LES air pollution prediction accuracy using data assimilation

Elsa Aristodemou[a,c], Rossella Arcucci[b,*], Laetitia Mottet[c,d], Alan Robins[e], Christopher Pain[c,b], Yi-Ke Guo[b]

[a]*School of Engineering, London South Bank University, UK*

[b]*Data Science Institute, Department of Computing, Imperial College London, UK*

[c]*Department of Earth Science & Engineering, Imperial College London, UK*

[d]*Department of Architecture, University of Cambridge, UK*

[e]*Department of Mechanical Engineering Sciences, University of Surrey, UK*

Abstract

It is recognised worldwide that air pollution is the cause of premature deaths daily, thus necessitating the development of more reliable and accurate numerical tools. The present study implements a three dimensional Variational (3DVar) data assimilation (DA) approach to reduce the discrepancy between predicted pollution concentrations based on Computational Fluid Dynamics (CFD) with the ones measured in a wind tunnel experiment. The methodology is implemented on a wind tunnel test case which represents a localised neighbourhood environment. The improved accuracy of the CFD simulation using DA is discussed in terms of absolute error, mean squared error and scatter plots for the pollution concentration. It is shown that the difference between CFD results and wind tunnel data, computed by the mean squared error, can be reduced by up to three order of magnitudes when using DA. This reduction in error is preserved in the CFD results and its benefit can be seen through several time steps after re-running the CFD simulation. Subsequently an optimal sensors positioning is proposed. There is a trade-off between the accuracy and the number of sensors. It was found that the accuracy was improved when placing/considering the sensors which were near the

*r.arcucci@imperial.ac.uk

pollution source or in regions where pollution concentrations were high. This demonstrated that only 14% of the wind tunnel data was needed, reducing the mean squared error by one order of magnitude.

Keywords: CFD, Data Assimilation, Wind Tunnel, Fluidity, Urban Environment, Pollutant concentration, Sensor positioning

1. Introduction

Climate change and air pollution form one of the grand challenges currently faced by humanity worldwide with still many questions remaining unanswered at the micro-scale/city-scale. The World Health Organisation (WHO) has found that outdoor air pollution in cities has been the primary cause of 4.2 million premature deaths annually worldwide [1]. WHO has subsequently established guidelines on the two main pollutants: the $PM_{2.5}$ and NO_2 [1]. By 2030, reducing deaths and illnesses due to air pollution is also one of the aims of the United Nation sustainable development programme, with one of the goals being "good health and well being" [2]. In Europe, the European Union Commission has also already established procedures for monitoring and advising on air quality, focusing on the five most important air pollutants: ozone (O_3), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), $PM_{2.5}$ and PM_{10} particles [3]. Currently, many areas in the United Kingdom (UK) and London in particular, fail to meet the WHO guidelines on two pollutants, the $PM_{2.5}$ and NO_2. Due to these failings, the UK Government has developed the Clean Air Strategy 2019 [4], which sets out the UK plans for dealing with all sources of air pollution, and ensuring the health of the na-

tion through better air quality. In addition, for London, UK, the Mayor's office has also developed the London Environment Strategy specifically for air pollution problems in the capital [5].

It is therefore clear that serious steps are taken at both international, national and city levels to reduce air pollution levels. Scientific and technological advances are therefore encouraged in the global effort to combat air pollution, with innovative tools being developed to assist in this effort. Computational methods/tools are at the forefront of these efforts, with many researchers worldwide looking at how to most accurately capture the dispersion of pollutants at the micro-scale level, within the urban environment [6, 7, 8]. Many studies have been carried out over the years, employing both simplified Gaussian plume models [8], to the more sophisticated ones using complex computational fluid dynamics (CFD) with turbulence models ranging from the Reynolds Averaged Navier-Stokes (RANS) approach to the more elaborate Large Eddy Simulation (LES) methods [9, 10]. To validate the CFD simulations, comparison of various variables (velocity, pollutant concentration, wind pressure coefficients, Reynolds stresses...) at micro-scale are usually confronted to wind tunnel experiments data [11, 12, 13] and in less extend to full scale experiments [14, 15]. For simple test case, i.e. simple geometric configuration, CFD models are reliable and reproduce with a good agreement data obtained from experiments [14]. However, the success of the comparison becomes mitigated and significant discrepancies between CFD and experiments are locally observed when more complex urban environment set-up

are considered [11, 12, 14]. In the context of predicting accurately the level of pollution at pedestrian level, i.e. at micro-scale, more advance numerical models need to be used in order to improve the reliability of their predictions.

The use of Data Assimilation (DA) technologies is a good candidate to answer this need. DA is an uncertainty quantification technique used to incorporate observational data into a prediction model in order to improve numerical forecasted results [16]. During the last 20 years, data assimilation and its various methodologies [16, 17] have reached a widespread and worldwide interests in many federal research institutes and universities such as the National Center for Atmospheric Research (NCAR, US); the National Centers for Environmental Prediction (NCEP, US); the Deutscher Wetterdienst (DWD, Germany); the Met Office (University of Reading and Imperial College of London, UK); the Japan Meteorological Agency (JMA, Japan); the Canadian Association of Management Consultants (CMC, Canada) and the Euro- Mediterranean Center for Climate Changes (CMCC, Italy). Since 10 years, the Variational DA (VarDA) approaches [18, 19] have gained acceptance for its accuracy and efficiency and thus as a powerful method. VarDA methodology is based on the minimisation of a function which estimates the discrepancy between numerical results and observations assuming that the two sources of information, forecast and observations, have errors that are adequately described by error covariance matrices.

A POD-EnVar DA method to identify pollutant source location and wind pa-

rameters from observations of the gas concentration is described in [20]. The POD-EnVar DA is coupled with a CFD software based on a Lattice Boltzmann Method (LBM) code and V-LES algorithm (PowerFLOW). A sensor placement procedure based on global sensitivity analysis techniques has also been proposed to improve the performances of the assimilation process. Using appropriate sensor placement, the position of the source can be identified with an accuracy of only a few meters. An EnKF method to improve the prediction of air flow (using the OpenFoam libraries as CFD software) in a real urban environment using wind sensors located in Stanford's campus, US, is proposed in [21]. The location as well as the number of sensors are discussed, highlighting that sensors located at roof height allows a better prediction of the velocity field. Moreover, with careful selection of the sensor location, their method is also able to accurately retrieve the probability distribution of the inlet wind velocity and direction. Finally, an Optimal Three Dimensional Variational (3DVar) data assimilation model coupled with a mesh-adaptivity open-source CFD software (Fluidity) is developed in [22, 23]. The method and its parametrisation is fully described and then successfully applied to a real urban environment located in London, showing that the error in the pollutant dispersion and the flow field can be reduced up to one order of magnitude compared to before the VarDA process. Moreover, this reduction in error propagates, as expected, in the next time step of the forcasted model (Fluidity).

As mentioned before, the validation of CFD models (for urban environment

simulations) are usually performed by comparing, more or less successfully, results to wind tunnel experiments, with a trend of higher discrepancy when increasing the complexity of the urban layout. Before going towards a comparison with full-scale experiment, the coupling of DA and CFD has also to be considered as a way to improve the comparison between wind tunnel and CFD results. The assimilation of pressure coefficient from a wind tunnel experiment in open-source CFD software (SU2) is proposed in [24] for the well-known 2D NACA 0012 and RAE 2822 airfoils. The sensitivity of the results depending on the number of observation points is discussed, highlighting that the assimilation works even with a very limited number of measurements (4% of the original data set was used). An Ensemble Kalman Filter (EnKF) method is used in [25] to assimilate values of surface pressure provided by a wind tunnel experiment around a so-called "squared cylinder" which can be assimilated to a single isolated building. They highlight that such a coupling method is promising, however, the 3D effect of the flow is neglected and only a 2D simulation is considered. Finally, the coupling between a Monte Carlo dispersion model (probabilistic model) and an EnKF method is developed in [26] showing that the error in the calculated concentration is reduced when coupling with DA.

The work presented in this paper aims to address the discrepancy between CFD results and wind tunnel data in terms of pollutant concentration prediction in a real urban environment. Thereby, CFD will be coupled with a novel data assimilation approach to show how data assimilation enhances

predictions and reduces the errors between measurements and simulations. In this paper, the Optimal Three Dimensional Variational (3DVar) data assimilation model presented in [22], which has been developed and implemented for improving air pollution prediction, is used. The forcasted model to be improved is the open-source CFD software Fluidity (http://fluidityproject.github.io/) [27], and the observed data are concentration values from a wind tunnel experiment performed in the EnFlo Meteorological Wind Tunnel [12].

The CFD Large Eddy Simulation (LES) method and the Optimal 3DVar DA model are first described in Section 2. The case set-up (wind tunnel experiment and CFD simulation) is then detailed in Section 3. The results using DA to improve the prediction of the pollutant concentration are presented in Section 4. Finally, conclusions are provided in Section 5.

2. Methodology

2.1. The Large Eddy Simulation method and Mesh Adaptivity

Over the last two decades, the Large Eddy Simulation (LES) method has become one of the most popular tool for atmospheric sciences, as it enables a more accurate capturing of the turbulent flows compared to the traditional Reynolds-Averaged Navier-Stokes (RANS) approach [10, 28, 29, 30, 31, 9]. The LES approach, although still complex and computationally demanding is "favoured" because it allows a more accurate representation of turbulence:

it achieves this by separating the flow into resolved and unresolved scales based on a cut-off length scale Δ. For scales greater than Δ, the flow is resolved and numerically solved, whilst for scales smaller than Δ, the flow is unresolved and represented by a sub-grid scale model. The subgrid scale model is crucial in representing the flow of turbulent energy from the large-scale (resolved) scale motions to the smallest (unresolved) scales where energy is dissipated [32]. The importance of the subgrid scale model was very clearly noted and considered in the very early works of the development of the LES methodology - especially the need to address anisotropic filtering and inhomogeneous effects [33, 34].

The LES equations describing turbulent flows are based on the filtered three-dimensional incompressible Navier-Stokes (NS) equations: continuity of mass (equation (1)) and momentum equations (equation (2)) [35].

$$\nabla.\overline{u} = 0 \tag{1}$$

$$\frac{\partial \overline{u}}{\partial t} + \overline{u}.\nabla\overline{u} = -\frac{1}{\rho}\nabla\overline{p} + \nabla.\Big[\left(\nu + \nu_\tau\right)\nabla\overline{u}\Big] \tag{2}$$

where $\overline{u}$ is the resolved velocity (m/s), $\overline{p}$ is the resolved pressure (Pa), ρ is the fluid density (kg/m^3), ν is the kinematic viscosity (m^2/s) and ν_τ is the anisotropic eddy viscosity (m^2/s).

The subgrid-scale model in Fluidity is based on the Smagorinsky model in which the eddy viscosity ν_τ is expressed by equation (3).

$$\nu_\tau = C_S^2 \Delta^2 |\overline{S}| \tag{3}$$

C_S is the Smagorinsky coefficient (taken equal to 0.1), Δ is the Smagorinsky length scale which depends on the local element size and $|\overline{S}|$ is the strain rate expressed as in equation (4).

$$|\overline{S}| = (2\overline{S}_{ij}\overline{S}_{ij})^{1/2} \tag{4}$$

where $\overline{S}_{ij}$ is the local strain rate defined by equation (5).

$$\overline{S}_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_i}{\partial x_i}\right) \tag{5}$$

A novel component in the implementation of the subgrid-scale model within Fluidity is the anisotropic eddy viscosity tensor [35] defined as in equation (6):

$$\nu_\tau = 4C_S^2 \left|\overline{S}\right| \mathcal{M}^{-1} \tag{6}$$

where $\mathcal{M}$ is the length scale metric from the adaptivity process [36] used here to relate eddy viscosity to the local grid size as shown in equation 7.

$$\mathcal{M}^{-1} = V^T \begin{pmatrix} h_\zeta^2 & 0 & 0 \\ 0 & h_\eta^2 & 0 \\ 0 & 0 & h_\xi^2 \end{pmatrix} V \tag{7}$$

with V^T and V the rotational transformations to transform from the local to the global coordinate systems and (h_ζ, h_η, h_ξ) the local element sizes. The factor of 4 arises because the filter width separating resolved and unresolved scales is assumed to be twice the local element size, which is squared in the viscosity model. It has been shown that an anisotropic eddy viscosity gives better results for flow simulations on unstructured grids [35].

The transport of a scalar field c (i.e, a passive tracer) in kg/m^3 is expressed using a classic advection-diffusion equation having a source term as in equation (8):

$$\frac{\partial c}{\partial t} + \nabla.(\mathbf{u}c) = \nabla.\left(\overline{\overline{\kappa}}\nabla c\right) + F \tag{8}$$

where $\mathbf{u}$ is the velocity vector (m/s), $\overline{\overline{\kappa}}$ is the diffusivity tensor (m^2/s) and F represents the source terms (kg/m^3/s).

The source term F is expressed by equation (9):

$$F = \frac{Q\rho}{V} \tag{9}$$

where Q is a volumetric flow rate expressed in m^3/s and V is the volume of the source in m^3.

The behaviour of the atmospheric boundary layer in Fluidity is represented using a turbulent inlet velocity based on a synthetic eddy method [37, 38]. The turbulent inlet velocity is controlled by: the turbulence length scales profiles (L_u, L_v, L_w), the mean velocity profiles $(\overline{u}, \overline{v}, \overline{w})$ as well as the Reynolds stresses profiles $(\overline{u'u'}, \overline{v'v'}, \overline{w'w'})$.

The need for combining the LES approach with adaptive meshes has been tackled as a way of overcoming the large range of length scales that exist in turbulent flows [39]. The challenge of combining the LES approach with 3D, adaptive, unstructured meshes was first undertaken and implemented within the Fluidity software [36, 35, 27]. Hence, one of the key and innovative aspects of Fluidity is its mesh-adaptivity capability on unstructured meshes. The adaptivity process allows: (i) the addition or reduction of the number of nodes and elements, leading subsequently to refining or coarsening of the mesh depending on the area of interest; (ii) smoothing of the mesh by moving nodes whilst keeping the overall number of elements and nodes the same. A-posteriori error estimates are made, aiming at certain targets for error [27]. Adaptivity options can be field-specific, i.e. different computed fields can be configured with their own specific adaptivity options. This process allows to have fine mesh in region where small-scale and important physical processes occur, while keeping a coarser mesh elsewhere, and then allowing to

considerably reduce the total computation time [36].

2.2. Data Assimilation

Let n be a fixed time level and let c^n be the state variable c as described in equation (8) at the fixed time level n. Let v^n be an observation of the state variable at time n and let consider a mapping H as in equation 10.

$$H : c^n \mapsto v^n. \tag{10}$$

Let $d^n = v^n - H(c^n)$ be the misfit. In this section, we introduce a Data Assimilation process in which the solution of the forecasting model (Fluidity) obtained from equation (8) is combined with information provided by a wind tunnel experiment in order to improve the accuracy of the solution c^n, i.e. to reduce d^n. The aim of the Data Assimilation problem is to find an optimal trade-off between the prediction made based on the Fluidity system state c^n (background) defined in equation (8) and the available observation v^n provided by the wind tunnel.

For a fixed time step n, given c^n and v^n, the DA process consists in finding c^{DA} as an inverse solution of equation (11) subject to the constraint given by equation (12).

$$v^n = H\left(c^{DA}\right), \tag{11}$$

$$c^{DA} = c^n. \tag{12}$$

Since H is typically rank deficient, the equation (11) is an ill-posed inverse problem [40, 41]. The Tikhonov formulation [42] leads to an unconstrained least squares problem, where the term in equation (12) provided by Fluidity ensures the existence of a unique solution of equation (11). The DA process can be then described as following [43]:

$$c^{DA} = argmin_c \left\{ \|c - c^n\|^2_{\mathbf{B}^{-1}} + \|v^n - H\left(c\right)\|^2_{\mathbf{R}^{-1}} \right\} \tag{13}$$

where $\mathbf{R}$ and $\mathbf{B}$ are the observation and model error covariance matrices respectively defined by equation (14) and equation (15):

$$\mathbf{R} := \sigma_0^2 \mathbf{I} \tag{14}$$

with $0 \leq \sigma_0^2 \leq 1$ representing the variance value of the distribution of the instruments errors and $\mathbf{I}$ the identical matrix;

$$\mathbf{B} = \mathbf{V}\mathbf{V}^T \tag{15}$$

where $\mathbf{B}$ is the background error covariance matrix as defined in Definition 1 associated with the state c since the true state will differ from the simulated

state by random or systematic errors.

Definition 1 (Variance-Covariance Matrix)**.** *Let* $\mathbf{X}$ *be a matrix of measurements of* pv *physical variables at spatial locations* $\mathcal{D} = \{x_j\}_{j=1,\dots,np}$ *for a correlation time window* $[0, T_1] = \{\tau_k\}_{k=1,\dots,M}$*:*

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_{NP} \end{bmatrix} \in \Re^{NP \times M} \tag{16}$$

where each of NP *row is a time series for a given location and* $NP = [pv] \cdot np$*. Let's assume that each row* X_i *of* $\mathbf{X}$ *has a mean* $E[X_i] = \{m_i\}_{i=1,\dots,NP}$ *and let's define* $\mathbf{m} = (m_i)_{i=1,\dots,NP}$*. Hence, the deviation matrix is:*

$$\mathbf{V} = \mathbf{X} - \mathbf{m} \in \Re^{NP \times M}, \tag{17}$$

If each vector X_i *has a distribution with probability density function* P*, then the expected value of* X_i *is defined by:*

$$E(X_i) = \frac{1}{M-1} \sum_{j=1,\dots,M} x_{ij} P(X_j) \quad . \tag{18}$$

The variance-covariance matrix $\mathbf{B} \in \Re^{NP \times NP}$ *of* $\mathbf{X}$ *(equation (19)) is then defined via the expected value of the outer product:*

$$\mathbf{B} = \mathbf{V}\mathbf{V}^T \quad . \tag{19}$$

If equation (13) is linearised around the background state [44], it yields:

$$c = c^n + \delta c \tag{20}$$

where $\delta c = c - c^n$ denotes the increments. The DA problem can then be re-formulated by the following form:

$$\delta c^{DA} = argmin_{\delta c} \left\{ \frac{1}{2} \delta c^T \mathbf{B}^{-1} \delta c + \frac{1}{2} (\mathbf{H}\delta c - d^n)^T \mathbf{R}^{-1} (\mathbf{H}\delta c - d^n) \right\} \tag{21}$$

where

$$d^n = v^n - H\left(c^n\right) \tag{22}$$

is the misfit between the observation and the solution computed by Fluidity and

$$H\left(c\right) \simeq H\left(c^n\right) + \mathbf{H}\delta c \tag{23}$$

denotes the linearised observational and model operators evaluated at $c = c^n$ where $\mathbf{H}$ is the Hessian of H.

In equation (21), the minimisation problem is defined on the field of increments [45]. In order to avoid the inversion of $\mathbf{B}$, as $\mathbf{B} = \mathbf{V}\mathbf{V}^T$ (see equation (19)), the minimisation can be computed with respect to a new variable $w = \mathbf{V}^+ \delta c$ [44], where $\mathbf{V}^+$ denotes the generalised inverse of $\mathbf{V}$, yielding to:

$$w^{DA} = argmin_w \left\{ \frac{1}{2} w^T w + \frac{1}{2} (\mathbf{HV}w - d^n)^T \mathbf{R}^{-1} (\mathbf{HV}w - d^n) \right\} \tag{24}$$

As the background error covariance matrix is ill-conditioned [41], in order to improve the conditioning, only Empirical Orthogonal Functions (EOFs) of the first largest eigenvalues of the error covariance matrix are considered. Since its introduction to meteorology by Edward Lorenz [46], EOFs analysis has become a fundamental tool in atmosphere, ocean, and climate science for data diagnostics and dynamical mode reduction. Each of these applications exploits the fact that EOFs allow a decomposition of a data function into a set of orthogonal functions, which are designed so that only a few of these functions are needed in lower-dimensional approximations [47]. Furthermore, since EOFs are the eigenvectors of the error covariance matrix [48], its condition number is reduced as well. Nevertheless, the accuracy of the solution obtained by truncating EOFs exhibits a severe sensibility to the variation of the value of the truncation parameter, so that a suitably choice of the number of EOFs is strongly recommended. This issue introduces a severe drawback to the reliability of EOFs truncation, hence to the usability of the operative software in different scenarios [48, 49]. In this paper, we set the optimal choice of the truncation parameter as a trade-off between efficiency and accuracy of the DA algorithm as introduced in [22].

The Optimal 3DVar data assimilation model as implemented in this paper is summarised in Algorithm 1.

Algorithm 1 : A VarDA algorithm to assimilate Wind Tunnel data into Fludity.

1: Input: α, $\{v_k\}_{k=0,\dots,m}$, c_0^M
2: Define $\mathbf{H}$
3: Compute $d_k \leftarrow v - \mathbf{H}c_0^M$ ▷ compute the misfit
4: Define $\mathbf{R}$ starting from the wind tunnel data v
5: Compute $\mathbf{V}$ ▷ deviance matrix defined in (17)
6: Compute $\mathbf{V}_\tau = EOFs(\mathbf{V}, \tau)$ ▷ reduced space computed by EOFs
7: Define the initial value of $\delta\mathbf{u}^{DA}$
8: Compute $w \leftarrow \mathbf{V}_\tau^+ \delta c^{DA}$ ▷ from the physical to reduced space
9: repeat ▷ start of the L-BFGS steps
10: Compute $J \leftarrow J(w)$
11: Compute $gradJ \leftarrow \nabla J(w)$
12: Compute new values for w
13: until (Convergence on w is obtained) ▷ end of the L-BFGS steps
14: Compute $\delta c^{DA} \leftarrow \mathbf{V}_\tau w$ ▷ from the reduced to physical space
15: Compute $c^{DA} \leftarrow c_0^M + \delta c^{DA}$

3. Case Set-up

Initial validations of Fluidity have already been reported in which comparisons of velocity, mean pollutant concentration predictions and surface pressures with wind tunnel data were carried out [11, 12, 13, 15, 35]. However, comparisons between experiments and simulations are most of the time plagued by discrepancies. In [12], the comparison of mean pollutant concentrations at 81 detector locations was carried out and it was observed that the errors between simulations and measurements ranged between 3% to over 50%. Thereby, the same test case than in [12] is used in this paper and is coupled with the Optimal Three Dimensional Variational (3DVar) data assimilation model presented in Section 2.2 (and fully described in [22]) in order

Building	Height (cm)
N	14.28
1	13.15
2	12.38
3	11.52
4	3.15
5	9.71
7	12.28

Table 1: Buildings heights, used in the LES simulation, based on the wind tunnel configuration. The buildings labels refer to the ones given in Figure 1b.

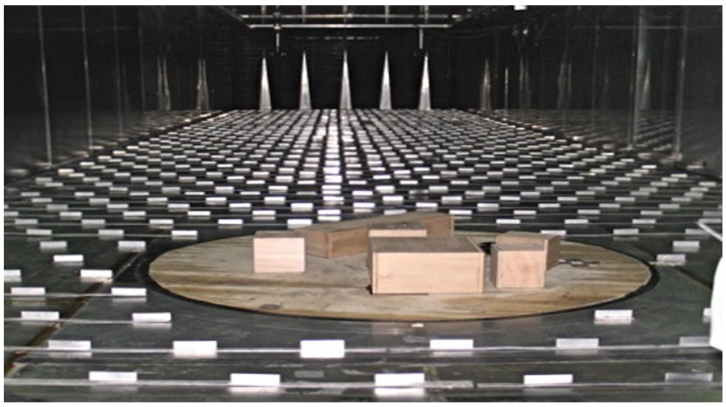
to improve the accuracy of the results predicted by Fluidity.

3.1. Geometry

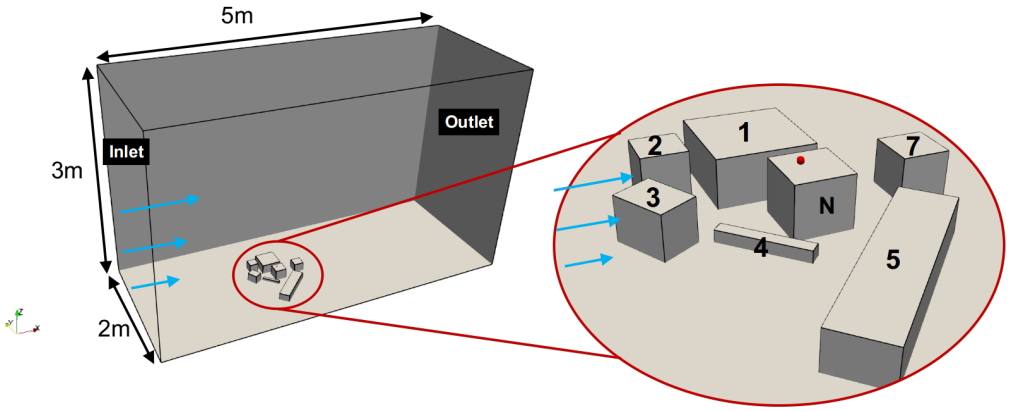
A 7-buildings configuration is considered in this paper as shown in Figure 1. The buildings represent a real, small neighbourhood area in central London, UK (51°30'00.0"N, 0°12'00.9"W), at the scale of 1:200 (wind tunnel scale). The heights of the seven buildings are given in Table 1, where the labels of each building refer to the ones given in Figure 1b.

3.2. Wind tunnel data

A set of experiments were carried out at the EnFlo wind tunnel [12] for the 7-buildings configuration (Figure 1a). The geometry represented is at 1:200 scale. The experiments were carried out in a fully developed, 1m-deep, simulated atmospheric boundary layer with a reference wind velocity U_{ref} of 2 m/s. The experimental atmospheric boundary layer represents neutral



(a) Wind tunnel set-up



(b) CFD Computational Domain

Figure 1: The 7-buildings configuration (a) in the wind tunnel experiment and (b) in the CFD simulation. In (b) the location of the source is denoted by the red sphere at the top of Building N and the wind direction is shown by the blue arrows.

atmospheric conditions and is initiated by a set of Irwin spires (vorticity-generators) at the inlet of the wind tunnel working section, with roughness elements on the floor to maintain the surface roughness condition. The surface roughness length z_0 and the friction velocity u^* are equal to 1.5 mm and 0.057 m/s, respectively, with u^* being the air velocity at the edge of the boundary layer.

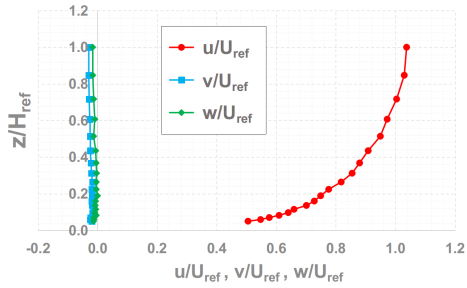
A passive tracer (propane) was emitted from a horizontal source, having a diameter of 20 mm, above Building N (Figure 1b) at 15.08 cm height from the ground of the test section, i.e 0.8 cm above Building N (having an height of 14.28 cm). It has to be noted that the source is not centred on the top of Building N. The tracer release flow rate in experiments was equal to $Q_{WT} = 2.4 \times 10^{-7} m^3/s$. The assumption could be made that there is no density difference between the emission gas and the surrounding fluid (air) [50]. Indeed, the propane gas (the trace gas) is diluted into the surrounding air such that the percentage proportion of propane/trace gas ranges between 0.99% to 2.1% of the total released gas. This mixture is considered neutrally buoyant and is released at a point source. These proportions and this gas are commonly used in wind tunnel experiment as non-reactive and non-depositing tracer gas, so that it disperses as a passive tracer in the flow [50, 13]. Due to the large amount of air mass, it is considered that the trace particle number is small so that the trace particles do not significantly influence the density. The density of the emission is then considered to be the same as of the surrounding air.

Mean tracer concentrations were measured using Combustion Fast Flame Ionisation Detectors (FFIDs) carried on a three-dimensional traverse system and each point measurement is an average over an acquisition period of 2 minutes. Measurements were taken for varying wind directions and model configuration, however only one configuration and one wind direction is considered in this paper. The tracer concentration was obtained at 738 different locations, located downstream the source.

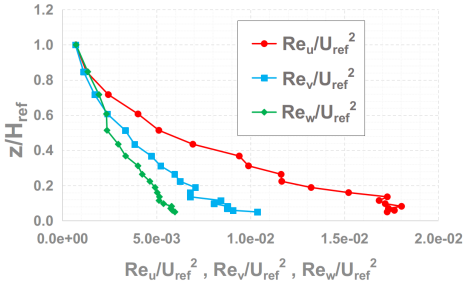
Figure 2 shows the inflow profiles of the velocity, the Reynolds stresses and the turbulence length scales used in the wind tunnel experiment, where the reference height H_{ref} is the boundary layer height and U_{ref} the reference velocity. The Reynolds number based on the mean building height $H_{mean} = 10.9cm$ is approximately equal to 1.4×10^4.

3.3. The LES Computational set-up

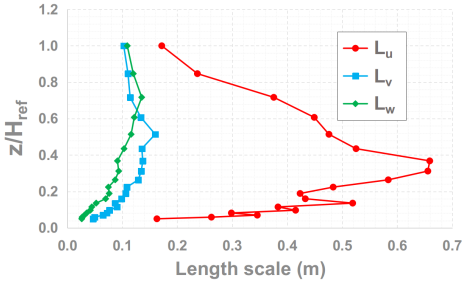
The Fluidity-LES software was used to model the flow field and the passive tracer concentrations within the 7-buildings configuration. The dimensions of the computational domain covered a volume of 5.0 m$\times$2.0 m$\times$3.0 m ($x \times y \times z$ -direction) as shown in Figure 1b, allowing a relatively long-development section for the formation of a deep boundary layer in the LES simulation. The blockage ratio is equal to 2.3%, below the maximum value recommended of 3% [51, 52]. The height, the width and the length of the domain are more than 5 times higher than the taller building (Building N) and/or the diameter



(a) Velocity profiles



(b) Reynolds stresses profiles



(c) Length scale profiles

Figure 2: Inflow profiles of the three components of (a) the velocity, (b) the Reynolds stresses and (c) the turbulence length scales used in both the wind tunnel experiments and the CFD simulations. H_{ref} is the boundary layer depth and U_{ref} is the reference velocity.

of the buildings area, hence following the guidance rules for CFD in urban environment [51, 52].

The location of the inlet and the outlet of the domain are shown in Figure 1b. A turbulent velocity is prescribed at the inlet, based on a synthetic eddy method [37] and the blue arrows in Figure 1b shows the wind direction. The wind direction is directly perpendicular to the front façades of Buildings 1, 2 and N. The mean velocities, the turbulence length scales as well as the Reynolds stresses profiles are set-up using the profiles provided by the wind tunnel experiments as shown in Figure 2. In a real urban dispersion problem, wind direction and velocity are constantly changing, however this is not taken into account here as the application is proposed for wind tunnel test cases only. Indeed, in wind tunnel, the experiments are done in controlled environments where the wind direction and velocity are fixed. The downstream boundary (outlet) is left as pressure boundary, whilst the remaining boundary conditions consisted of: (i) the "no slip" condition for the solid walls of buildings and "floor" of the domain, and (ii) the "slip/no shear" condition for the free surfaces (sides and top of the domain).

The emission source was placed at the top of the central building, i.e Building N, at the same location and height than in the wind tunnel as shown by the red sphere in Figure 1b. The diameter of the source is equal to 20 mm and the diffusion coefficient of propane in an excess of air is set to $1 \times 10^{-5} m^2/s$. The propane is considered as non-reactive and non-depositing

tracer gas [50]. Thus, the propane emission in the simulations is considered as a passive tracer, i.e. no density effect/variation with the surrounding fluid and travel with the air flow velocity such that the classic advection-diffusion with a source term (equation 8) is used. The source term F, expressed by equation (9), is set equal to $F = 1kg/m^3/s$, leading to a volumetric flow rate Q_F equal to $2.5 \times 10^{-6} m^3/s$, i.e. one order of magnitude higher than in experiments. In order to be compared, the concentration c from wind tunnel experiment and the ones obtained from Fluidity, the concentrations are commonly normalised using equation (25) [53, 54]:

$$c^* = \frac{cU_{ref}H_{mean}^2}{Q} \tag{25}$$

where c^* is the normalised concentration, U_{ref} is the reference velocity (m/s) at the top of the boundary layer and H_{mean} is the mean building height. U_{ref} and H_{mean} are the same in both experiment and simulation. Hence, the concentrations from the wind tunnel c_{WT} are converted into their equivalent "Fluidity" values v^n using equation (26):

$$v^n = c_{WT}\frac{Q_F}{Q_{WT}} \tag{26}$$

where Q_F and Q_{WT} stand for the volumetric flow rate in Fluidity and in wind tunnel, respectively. In the DA process, this modified wind tunnel concentration corresponds to the observed data, i.e v^n.

All the equations are solved using second order schemes in time and space. The NS equations are discretised using a continuous Galerkin discretisation, while the advection-diffusion is discretised using a second order upwind scheme. An adaptive time step is used and the CFL number is equal to 0.9, leading to an average time step equal to 1×10^{-3}, while the Crank-Nicholson scheme is used for the time discretisation. Note that the time step is not constant in the simulation because of the use of mesh adaptivity. Absolute and relative convergence errors were set to 10^{-12} and 10^{-7}, respectively for all fields (pressure, velocity and tracer).

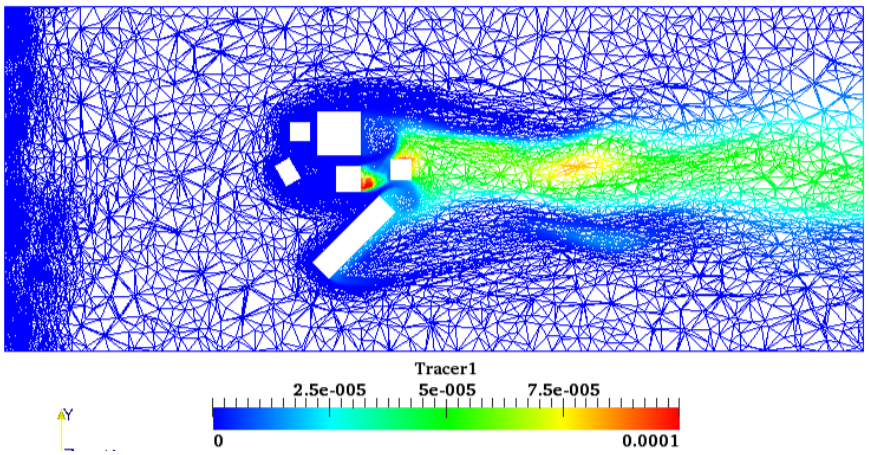
3.4. Mesh adaptivity and Supermesh

For the LES simulation presented in this work, field-specific adaptivity options were assigned to the velocity field and the tracer field. For both fields, mesh resolution was also controlled by specifying the maximum and the minimum sizes of the elements in the domain. They are respectively taken equal to 1 cm and 15 cm. Moreover, to resolve the source, the mesh is locally controlled around the source location by setting the minimum edge length to be 3 mm, and allowing the maximum element size to be 4 mm. The mesh was adapted every 15 time steps, and anisotropic gradation was also allowed in the simulation. The maximum number of nodes was set to 400,000. An example of the adaptivity effect on the computational mesh can be seen in Figure 3 for the instantaneous tracer field on two horizontal planes. The resolution of the mesh is fine near the inlet to capture accurately all the eddies

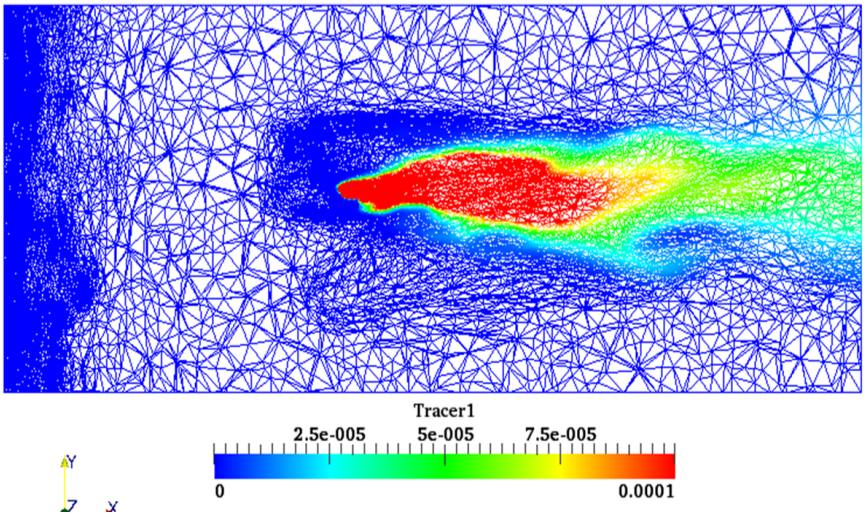
coming into the domain: this is a direct effect of the mesh adaptivity. It has to be noted that the mesh is changing every 15 time steps, and the meshes shown in Figure 3 is an example of instantaneous mesh.

In order to compute $\mathbf{B}$, the background error covariance matrix, (equation (15)), the number of nodes in the mesh has to be "fixed", i.e always the same at every time step. Thereby, Fluidity is running with mesh adaptivity for 34 sec (real time), which is sufficient for the flow statistics to reach a quasi-steady state. From this point onwards, the mesh obtained has a minimum and maximum edge lengths equal to 1.4 mm and 25 cm, respectively; while the number of nodes in the mesh is equal to 170,775. This mesh will be referred as the *supermesh* in the following and is shown in Figure 3. The *supermesh* is considered as an optimal mesh, as the simulation has run long enough to have fine elements in areas where important physical processes occur repeatedly. Fluidity results are then projected onto that *supermesh* in order to compute $\mathbf{B}$, the background error covariance matrix, (equation (15)). It has to be mentioned that this process (projection of all data) has to be done only once, as $\mathbf{B}$ has to be computed only once. The mesh adaptivity process can then be used normally when Fluidity is running: the projection of Fluidity data onto the *supermesh* is then done only for the time step at which observed data want to be assimilated.

The wind tunnel data v^n, i.e the observed data, has also to be projected on the *supermesh*. The location of sensors in wind tunnel does not necessarily



(a) $z = 6.5$ cm



(b) $z = 14.8$ cm

Figure 3: Instantaneous tracer concentration, i.e. pollutant concentration, at t = 34 sec for horizontal planes (xOy) at heights (a) $z = 6.5$ cm and (b) $z = 14.8$ cm, obtained from Fluidity. The mesh shown in these figures also corresponds to the mesh used as the *supermesh*. The tracer concentration ranges between $0kg/m^3$ (blue colour) and $1 \times 10^{-4}kg/m^3$ (red colour).

lie on a *supermesh* node. Therefore, using interpolation method, the sensor value is distributed to the four nodes of the tetrahedron in which lies the sensor. As one mesh node can be part of several tetrahedrons in which lie different sensors, the number of nodes in the mesh where sensors data are assigned is smaller than four times the number of sensors. This process leads to a number of nodes equal to 1391, i.e. values from wind tunnel experiments are assigned to 1391 nodes in the *supermesh*.

4. Results and Discussion

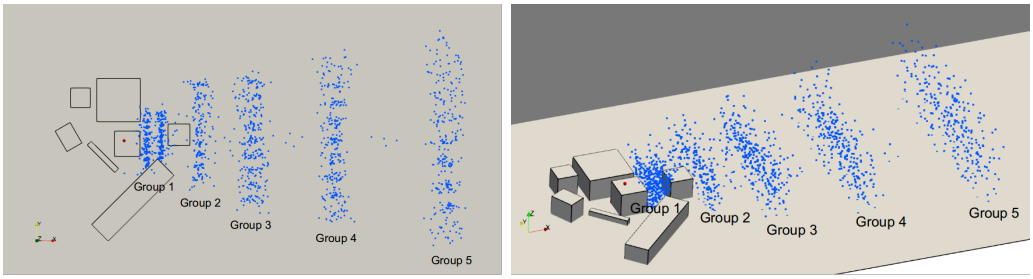
A comparison between Fluidity results and wind tunnel data for 81 detectors was carried out in [12], with differences/errors between simulations and measurements ranging between 3% to over 50%. The results presented here aim to reduce these errors using the DA method described in Section 2.2. In this section, 1391 observation points, located downstream of the pollutant source, are considered and their locations are shown in Figure 4.

4.1. Accuracy evaluation

The accuracy of the DA results are evaluated using:

- the absolute error

$$E(c) = |c - v^n| \tag{27}$$



(a) (b)

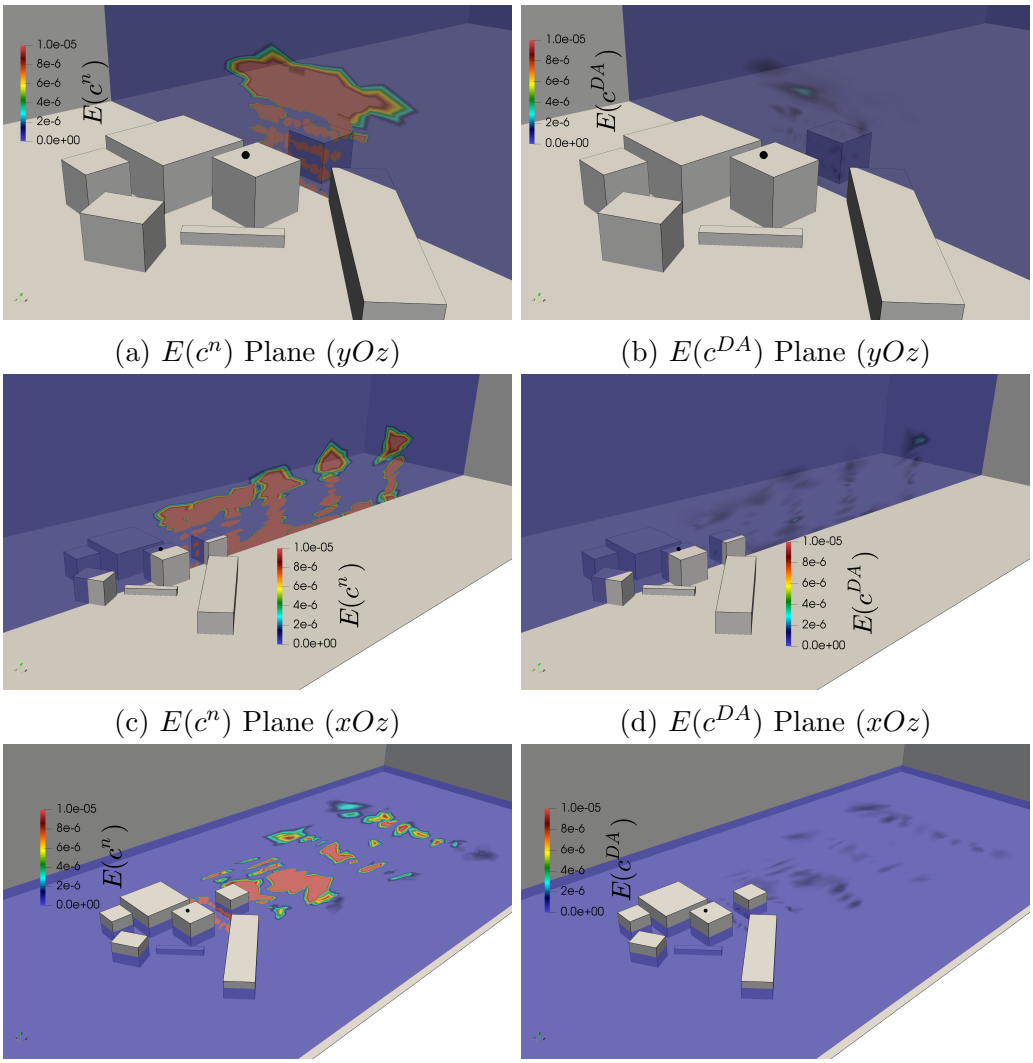
Figure 4: Location (blue dots) of the sensors in the domain. Five groups of sensors can be identified, based on their distances from the source. The red sphere denotes the location of the source.

- the mean squared error

$$MSE(c) = \frac{\|c - v^n\|_{L^2}}{\|v^n\|_{L^2}} \tag{28}$$

where c is either c^n the Fluidity concentration at time step n or c^{DA} the corrected concentration using DA (Algorithm 1) and v^n is the wind tunnel observed data.

Figure 5 shows the values of the absolute errors $E(c^n)$ and $E(c^{DA})$ on three different slices: through the oriented planes (xOy), (xOz) and (yOz). After the DA process, the absolute error is visibly reduced by almost one order of magnitude everywhere in the domain. The absolute error $E(c^n)$ ranges between 1×10^{-5} and 3×10^{-6}, with error values decreasing as the distance in the y-direction from the source increases (Figure 5e). After the DA process, the absolute error $E(c^{DA})$ becomes lower than 2.5×10^{-6} at every sensor location.



(a) $E(c^n)$ Plane (yOz) (b) $E(c^{DA})$ Plane (yOz)

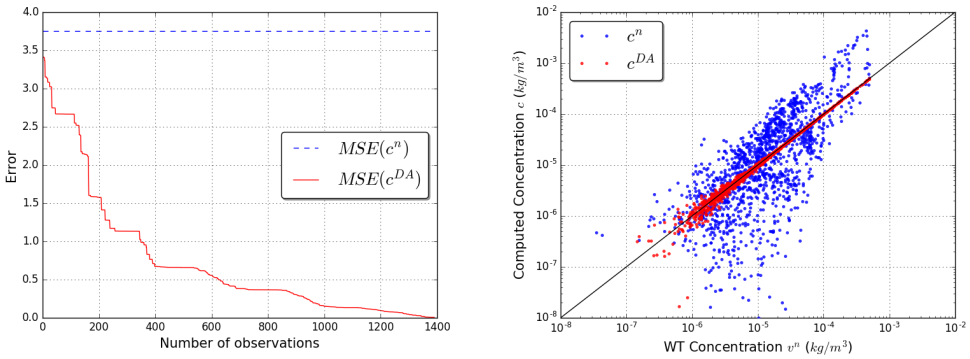
(c) $E(c^n)$ Plane (xOz) (d) $E(c^{DA})$ Plane (xOz)

(e) $E(c^n)$ Plane (xOy) (f) $E(c^{DA})$ Plane (xOy)

Figure 5: Values of the absolute error $E(c)$ (equation (27)) through three slices: through a plane (xOy), a plane (xOz) and a plane (yOz). The absolute errors shown are computed before $(E(c^n))$ and after $(E(c^{DA}))$ the assimilation process. The scale of $E(c)$ ranges between 0 (blue colour) and 1×10^{-5} (red colour) in all sub-figures. The black sphere denotes the source location.

Figure 6a shows the variation of the mean squared errors $MSE(c^n)$ and $MSE(c^{DA})$ as a function of the number of assimilated observations. The $MSE(c^n)$ does not depend on the number of observations and is then constant and equal to 3.749. $MSE(c^{DA})$ decreases as a function of the number of observations, reaching a value of 5.6×10^{-3} for 1391 observations assimilated: the DA process allows a reduction of the mean squared error by almost three order of magnitudes. $MSE(c^{DA})$ is reduced by one order of magnitude (3.75×10^{-1}) and two order of magnitudes (3.75×10^{-2}) assimilating 722 and 1312 observations, respectively. Moreover, $MSE(c^{DA})$ is approximately divided by two for 164 observations assimilated. Indeed, as shown in Figure 6a, while the number of observations starts to increase, the $MSE(c^{DA})$ firstly decreases very sharply, exhibiting a value of 6.7×10^{-1} for 400 number of observations. After what, the $MSE(c^{DA})$ continues to be reduced as the number of observations raises, but less quickly. The observed values v^n are assimilated in "ascending order" in terms of distance from the source, i.e from the sensor being the closer of the pollutant source to the sensor being the farthest. In other words, during the assimilation process, while the number of observations increases, more and more sensors located far way from the source are taken into account. The trend of $MSE(c^{DA})$ shown in Figure 6a tends to highlight that the closest sensors have an higher impact on the error reduction.

A scatter plot of the computed concentrations c^n and c^{DA} (using 1391 observations) as a function of the observed data v^n is shown in Figure 6b. Ideally,



(a) $MSE(c)$ at n. (b) c at n.

Figure 6: Values of (a) the mean squared error $MSE(c)$ (equation (28)) as a function of the number of assimilated observations and (b) the pollutant concentration c as a function of the wind tunnel data v^n (logarithm scale for both axis). In (b), the black line corresponds to the ideal matching between data and c^{DA} is obtained assimilating 1391 observations. Results at n, i.e before re-running Fluidity.

values should match the black line shown in Figure 6b. However, Fluidity concentration c^n are spread above and below it (blue dots in Figure 6b) with a tendency of larger spread towards low concentrations. After using the DA process, the corrected concentrations c^{DA} exhibit an obvious better agreement with wind tunnel data (red dots in Figure 6b), with corrected values much closer to the ideal matching (black line). The DA process performs very well in correcting Fluidity results for high concentrations, i.e concentrations higher than $1 \times 10^{-5} kg/m^3$; while, even if obviously a better agreement exists, small discrepancies still subsist for lower concentrations, i.e. concentrations lower than $1 \times 10^{-5} kg/m^3$.

4.2. Impact of DA on Fluidity results

In this section, the impact of DA on Fluidity results are discussed: the corrected concentrations c^{DA} are used to re-run Fluidity such that these new values are used as initial condition for the tracer.

Let $M_{n,n+i}$ denote the Fluidity software from the time step n to the time step $n+i$ such that $c^{n+i} = M_{n,n+i}(c)$. The mean squared error at time step $n+i$ is then defined as in equation (29):

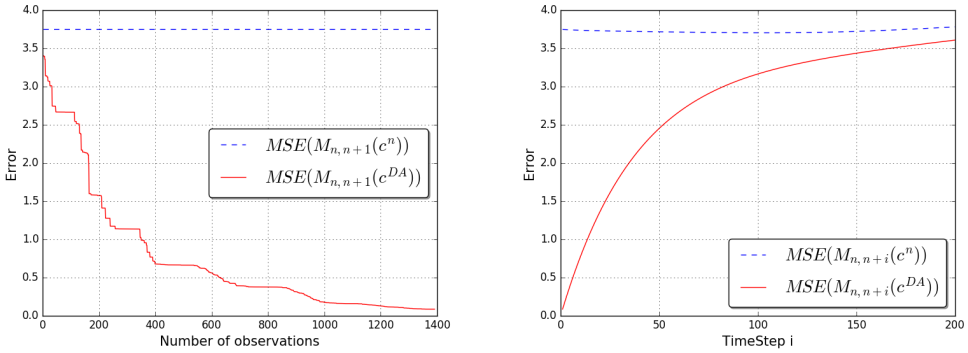
$$MSE(M_{n,n+i}(c)) = \frac{\|M_{n,n+i}(c) - v^{n+i}\|_{L^2}}{\|v^{n+i}\|_{L^2}} \tag{29}$$

where c is either c^n the Fluidity concentration at time step n or c^{DA} the corrected concentration using DA (Algorithm 1) and v^{n+i} is the wind tunnel observed data.

Figure 7a shows the variation of the mean squared errors $MSE(M_{n,n+1}(c^n))$ and $MSE(M_{n,n+1}(c^{DA}))$ obtained after re-running Fluidity for one more time step. As for $MSE(c^n)$, $MSE(M_{n,n+1}(c^n))$ does not depend of the number of observations and is equal to 3.747. Figure 7a confirms that the error $MSE(M_{n,n+1}(c^{DA}))$ also decreases as the number of observed data increases with almost the same trend than the reduction of $MSE(c^{DA})$. For 1391 observations, $MSE(M_{n,n+1}(c^{DA}))$ is equal to 8.7×10^{-2}, i.e the error is reduced by two order of magnitudes compared to $MSE(M_{n,n+1}(c^n))$. It has to be noted that the minimum value of $MSE(M_{n,n+1}(c^{DA}))$ is one order of

magnitude higher than the minimum of $MSE(c^{DA})$: this is not surprising as the Fluidity software $M_{n,n+1}$ introduces intrinsically new errors. Noteworthy values can be mentioned: as for $MSE(c^{DA})$, the error is almost divided by two for 164 observations assimilated ($MSE(M_{n,n+1}(c^{DA})) = 1.95$) and $MSE(M_{n,n+1}(c^{DA}))$ is reduced by one order of magnitude (3.7×10^{-1}) when 742 observations are considered.

Fluidity is re-run for 200 more time steps in order to see how the reduction in error gained by the DA process at time step n propagates into the model through time. The values of $MSE(M_{n,n+i}(c))$ as a function of the time step i is shown in Figure 7b. $MSE(M_{n,n+i}(c^n))$ slightly changes over time but stays however more or less constant with an average value of 3.724. $MSE(M_{n,n+i}(c^{DA}))$ increases while the time step increases, tending to reach the value of $MSE(M_{n,n+i}(c^n))$ after a long enough time: this behaviour is expected as the model introduces new errors. This is because the physical system does not change after the assimilation process as this only affects the state, the boundaries and initial conditions. These are errors intrinsic to the forecasting model problem which propagate on time steps. These are the approximation errors introduced by the linearisation, the discretisation, the model reduction... These occur when infinite-dimensional equations are replaced by a finite dimensional system (that is the process of discretisation), or when simpler approximations to the equations are developed (e.g., by model reduction). Finally, given the numerical problem, the algorithm is developed and implemented as a mathematical software. At this stage, the



(a) (b)

Figure 7: Values of the mean squared error $MSE(M_{n,n+i}(c))$ (equation (29)) after re-running Fluidity. (a) Variation of $MSE(M_{n,n+1}(c))$ as a function of the number of assimilated observations for $i = 1$, i.e after one time step. (b) Variation of $MSE(M_{n,n+i}(c))$ as a function of the time step i. c^{DA} is obtained assimilating 1391 observations.

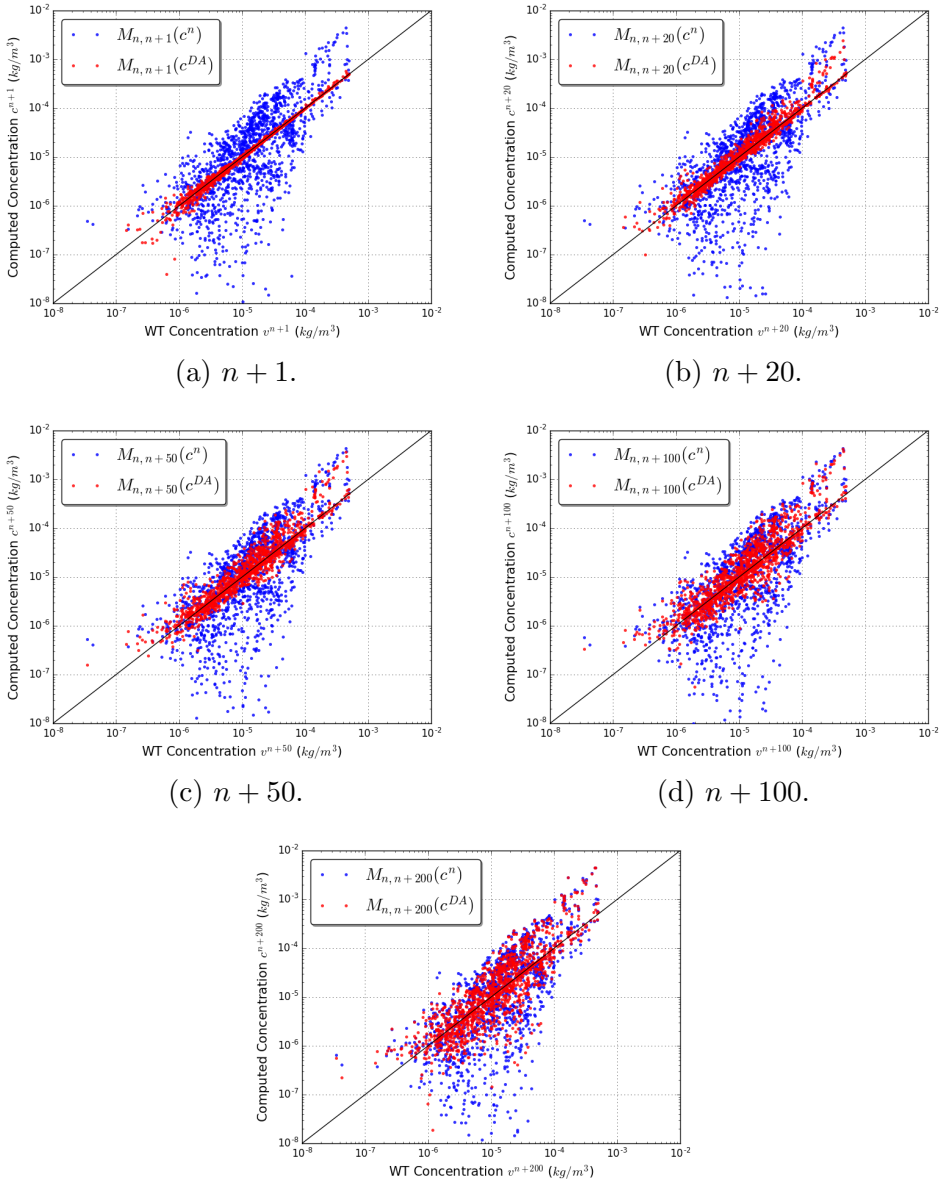
inevitable rounding errors introduced by working in finite-precision arithmetic occurs. These errors cannot be controlled but, after few time steps, the DA process can be run again to maintain the forecasting error under a fixed value. It can be observed that the error $MSE(M_{n,n+i}(c^{DA}))$ stays smaller than $MSE(M_{n,n+i}(c^n))$ for the 200 time steps shown in Figure 7b, highlighting that the reduction in error gained with the use of DA travels through the model and then benefit positively to the accuracy of the results predicted by Fluidity. In particular, the error $MSE(M_{n,n+31}(c^{DA}))$ at time step $i = 31$ still exhibits a value twice smaller than $MSE(M_{n,n+31}(c^n))$, with a value equal to 1.86.

Figure 8 shows a scatter plot of the computed concentrations $M_{n,n+i}(c^n)$ and $M_{n,n+i}(c^{DA})$ (using 1391 observations) as a function of the observed data v^{n+i} for i equal to 1, 20, 50, 100 and 200. Figure 8 shows how the

pollutant concentration at sensors location evolves through time. It can be seen that $M_{n,n+i}(c^{DA})$ deviates from the ideal values (black line) starting from the highest concentrations, i.e. concentrations higher than $1 \times 10^{-4} kg/m^3$, as shown in Figure 8a. The model tends to recover the higher computed concentrations very quickly through time. Then, the points having medium concentration, ranging between $1 \times 10^{-4} kg/m^3$ and $1 \times 10^{-6} kg/m^3$, start to deviate from the ideal values as i increases, but still keeping a reasonable spread (Figure 8b, Figure 8c and Figure 8d). Finally, at time step $i = 200$, the benefit of DA has more or less vanished and the values $M_{n,n+200}(c^{DA})$ tend to recover the ones obtained from $M_{n,n+200}(c^n)$ (Figure 8e). An interesting point that can be noted from Figure 8 is that the positive impact and benefit introduced by the DA process for sensors where the concentration was underestimated by Fluidity is preserved through time, i.e the impact is clear even after 200 time steps (Figure 8e).

4.3. Location of assimilating sensors

The values of the mean squared errors $MSE(c^{DA})$ and $MSE(c^n)$ are here used to choose an optimal sensors positions which add a positive benefit when they are assimilated: a trade-off between the number of sensors available in reality and the gain obtained from the DA process has to be considered. Several tests were performed to find the optimal sensors positioning and the cases considered are summarised in Table 2. As a reminder and for comparison, $MSE(c^n)$ is equal to 3.749.



(a) $n+1$.

(b) $n+20$.

(c) $n+50$.

(d) $n+100$.

(e) $n+200$.

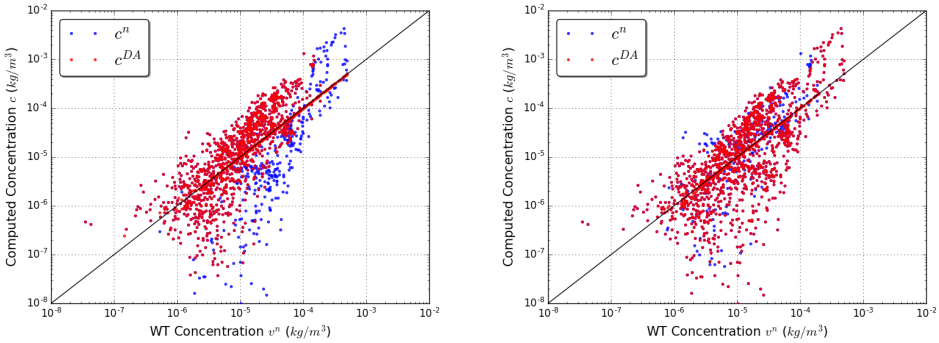
Figure 8: Values of $M_{n,n+i}(c)$ as a function of the wind tunnel data v^{n+i} at time steps (a) $i=1$, (b) $i=20$, (c) $i=50$, (d) $i=100$ and (e) $i=200$. The black lines correspond to the ideal matching between data. c^{DA} is obtained assimilating 1391 observations. Logarithm scale is used for all axis.

Test	**Sensors location**	**Number of observations**	$MSE(c^{DA})$
Group 1	Figure 4	341	1.13
Group 2	Figure 4	194	3.63
Group 3	Figure 4	300	3.71
Group 4	Figure 4	283	3.73
Group 5	Figure 4	273	3.746
High concentration $c^n \geq 1.5 \times 10^{-4} kg/m^3$	Figure 10a	140	3.89×10^{-1}
Low concentration $c^n \leq 1.5 \times 10^{-6} kg/m^3$	Figure 11a	295	3.746
Group 1 + High Concentration $c^n \geq 1.5 \times 10^{-4} kg/m^3$	Figure 12a	425	3.16×10^{-1}

Table 2: Summary of the cases considered to propose an optimal sensor positioning. As a reminder and for comparison, $MSE(c^n)$ is equal to 3.749.

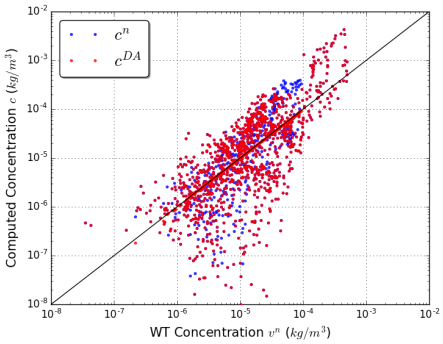
As a first attempt, as it can be seen in Figure 4, five different groups of sensors can be identified, based on their distances from the source location. Hence, every sensors group is assimilated separately in order to highlight and quantify their impacts on $MSE(c^{DA})$: results are shown in Table 2 and Figure 9. Assimilating sensors in Group 1 divides the error $MSE(c^{DA})$ by two compared to $MSE(c^n)$, while assimilating any other sensors group lead to a very poor, to not say negligible, reduction in error. As shown in Figure 9, Group 1 is also almost the only one having a positive impact on reducing the spread observed at low concentrations (spread compared to ideal values, i.e. black line). It is then obvious that sensors near the source should be prioritised in order to improve the accuracy of Fluidity.

The next two tests considered are as follows: only sensors having the high-

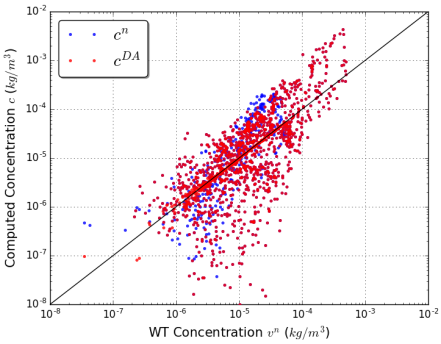


(a) Group 1

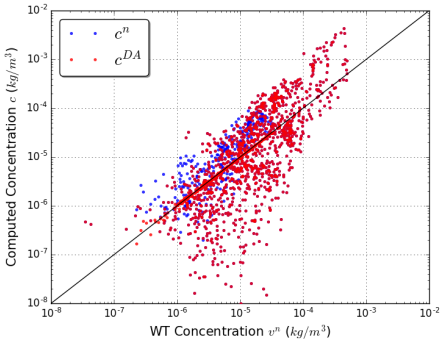
(b) Group 2



(c) Group 3



(d) Group 4

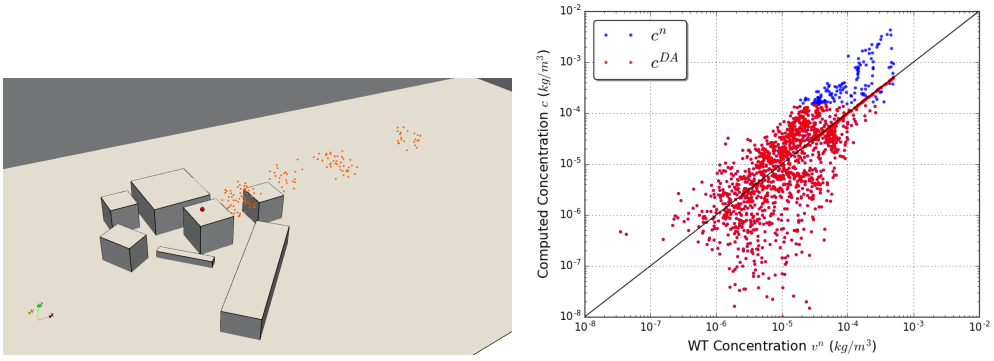


(e) Group 5

Figure 9: Values of the pollutant concentration c as a function of the wind tunnel data v^n. Value of c^{DA} are obtained assimilating different group of sensors from (a) Group 1, the closest to the source to (e) Group 5, the farthest from the source. For the group labels, see Figure 4. The black lines correspond to the ideal matching between data. Logarithm scale is used for all axis. Data obtained at n, i.e before re-running Fluidity.

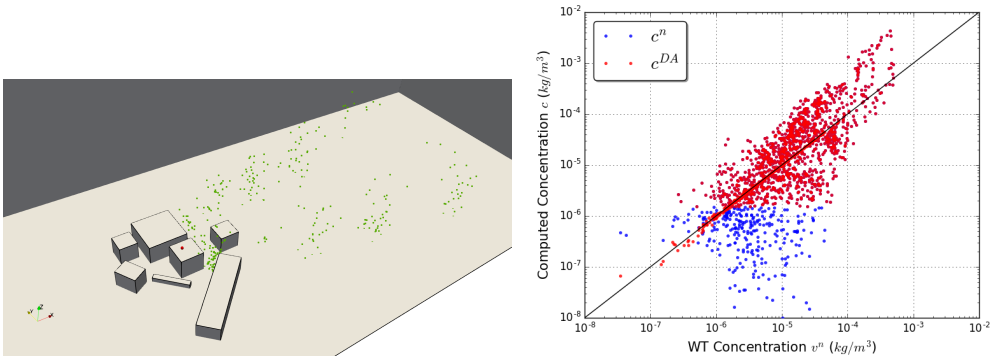
est Fluidity concentrations ($c^n \geq 1.5 \times 10^{-4} kg/m^3$) are assimilated, then only the ones having the lowest concentrations ($c^n \leq 1.5 \times 10^{-6} kg/m^3$) are considered. Figure 10a and Figure 11a show the locations of the 140 observation points having the highest concentrations and the 295 observation points having the lowest concentrations, respectively. Not surprisingly, the high concentrations are located downstream and in the alignment of the source. Results of $MSE(c^{DA})$ are reported in Table 2. The error is reduced by one order of magnitude if the higher concentrations locations are assimilated ($MSE(c^{DA}) = 3.89 \times 10^{-1}$), while assimilating the 295 lowest concentrations locations lead to a non-significant reduction of error ($MSE(c^{DA}) = 3.746$). Even if the number of assimilated sensors is twice higher, the error is not significantly reduced when assimilating low concentrations. Looking at the scatter plots in Figure 10b and Figure 11b, c^n exhibit a large spread around the ideal value (black line) for low concentrations and assimilating them sounds a legitimate choice. However, the higher concentrations play a more determining role in the model error and should then be preferred as location for sensors. Moreover, talking about air pollution in general, the spots of high concentration are usually of primary interest and are the ones that need to be accurately predicted.

The last test proposed, as an ultimate optimised case, consists in assimilating sensors located near the source, i.e. in Group 1, as well as the sensors exhibiting the highest concentrations only ($c^n \geq 1.5 \times 10^{-4} kg/m^3$) for all the other groups. This case leads to 425 observation points as shown in Fig-



(a) (b)

Figure 10: (a) Location of the sensors assimilated, i.e. the 140 sensors exhibiting the higher Fluidity concentrations ($c^n \geq 1.5 \times 10^{-4} kg/m^3$). The red sphere denotes the location of the source. (b) Values of the pollutant concentration c as a function of the wind tunnel data v^n when the highest concentrations are assimilated. The black line corresponds to the ideal matching between data.



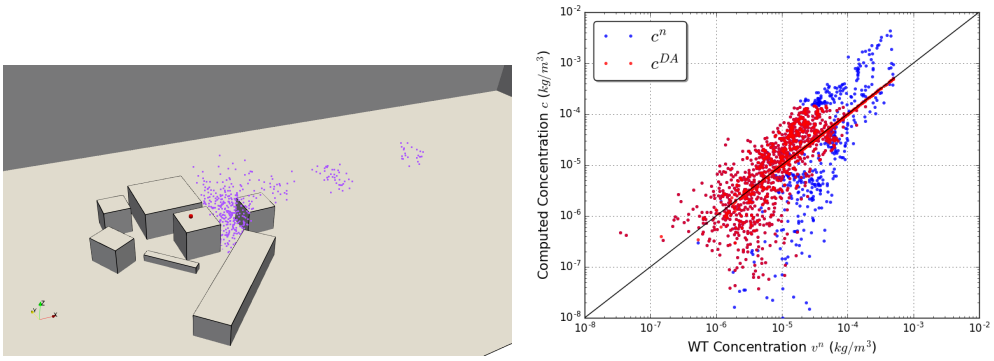
(a) (b)

Figure 11: (a) Location of the sensors assimilated, i.e. the 295 sensors exhibiting the lower Fluidity concentrations ($c^n \leq 1.5 \times 10^{-6} kg/m^3$). The red sphere denotes the location of the source. (b) Values of the pollutant concentration c as a function of the wind tunnel data v^n when the lowest concentrations are assimilated. The black line corresponds to the ideal matching between data.

ure 12b. Table 2 and Figure 12 show the results for this optimised case: the $MSE(c^{DA})$ is equal to 3.16×10^{-1}. Compared to the value obtained when assimilating only the highest concentration location (3.89×10^{-1}), adding sensors in Group 1 in the assimilation process brings a relatively small improvement of results. However, looking at Figure 12b, this set of sensors positioning remains the optimal one in terms of error reduction: the higher concentrations are properly corrected, thus decreasing the MSE; while the discrepancies seen at low concentrations are satisfyingly reduced. Hence, the optimal sensors locations recommended to improve the accuracy of Fluidity results is a trade-off between being close to the source independently of the concentration values and being in region where the concentration is high. In this particular optimal case, 425 observations points are used: as the wind tunnel data are projected onto the *supermesh*, this approximately corresponds to 106 wind tunnel sensors. Compared to the 738 points data provided by the experiment, only 14% of the data need to be used to improve the accuracy of Fluidity.

5. Conclusion

In this paper an Optimal Three Dimensional Variational (3DVar) data assimilation model to reduce the discrepancy between CFD results and wind tunnel data in terms of pollutant concentration prediction in urban environment was presented. Wind tunnel experiments were performed in the



(a) (b)

Figure 12: (a) Location of the optimal sensors assimilated, i.e. 425 sensors (Group 1) and sensors having high Fluidity concentrations ($c^n \geq 1.5 \times 10^{-4} kg/m^3$)). The red sphere denotes the location of the source. (b) Values of the pollutant concentration c as a function of the wind tunnel data v^n when the optimal sensors are assimilated. The black line corresponds to the ideal matching between data.

607 EnFlo Meteorological Wind Tunnel and the forcasted model was Fluidity,
608 an open-source CFD software using mesh adaptivity. The mesh adaptivity
609 technology was used during CFD simulations and then generate an optimal
610 *supermesh*. The *supermesh* was used in the variational DA process, as well
611 as to interpolate the Wind Tunnel data.

612 The improvement of Fluidity accuracy, in terms of pollutant concentration
613 prediction, was discussed using the absolute errors, the mean squared errors
614 and scatter plots. Using the DA process presented in this paper, the error in
615 the results between Fluidity and wind tunnel data can be reduced by three
616 order of magnitudes if all the wind tunnel sensor values are assimilated. It
617 has been shown that this reduction in error gained using DA is preserved by
618 the model Fluidity and its benefit can still be observed through several time

steps. In particular, it has been observed that high concentration are the one deviating quickly from ideal values, while corrections on low concentrations are fully preserved through time.

Finally, an optimal sensors location were proposed taking into account the improvement of Fluidity accuracy while having a limited number of wind tunnel sensors. The optimal sensors locations is a trade-off between being close to the source independently of the concentration values and being in regions where the concentration is high. In the particular case presented in this paper, which used 738 points data from the wind tunnel experiment, only 14% of the data points were needed to reduce the errors by one order of magnitude and improve the accuracy of results predicted by Fluidity in terms of pollutant concentration.

Acknowledgments

This work is supported by the EPSRC Grand Challenge grant Managing Air for Green Inner Cities (MAGIC) EP/N010221/1, by the EPSRC Centre for Mathematics of Precision Healthcare EP/N0145291/1 and by the EPSRC Low Carbon Climate-Responsive Heating and Cooling of Cities (LoHCool) EP/N009797/1.

References

[1] World Heath Organization. Ambient air pollution: A global assessment of exposure and burden of disease. http://www.who.int/airpollution/ambient/health-impacts/en/. Accessed: 2019-04-16.

[2] United Nations Sustainable Development Goals. Ambient air pollution: A global assessment of exposure and burden of disease. https://www.un.org/sustainabledevelopment/health/. Accessed: 2019-05-15.

[3] European Commission. Air quality. http://ec.europa.eu/environment/air/quality/index.htm. Accessed: 2019-05-15.

[4] Department of Environment, Food and Rural Affairs, UK. The UK Clean Air Strategy 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/770715/clean-air-strategy-2019.pdf. Accessed: 2019-04-16.

[5] London Environment Strategy. The UK Clean Air Strategy 2019. https://www.london.gov.uk/what-we-do/environment/london-environment-strategy. Accessed: 2019-05-15.

[6] Y. Tominaga and T. Stathopoulos. CFD simulation of near-field pollutant dispersion in the urban environment: A review of current modeling techniques. *Atmospheric Environment*, 79:716–730, 2013.

[7] M. Lateb, R.N. Meroney, M. Yataghene, H. Fellouah, F. Saleh, and M.C. Boufadel. On the use of numerical modelling for near-field pollutant

dispersion in urban environments - A review. *Environmental Pollution*, 208:271–283, 2016.

[8] F. Gronwald and S.Y. Chang. Evaluation of the precision and accuracy of multiple air dispersion models. *Journal of Atmospheric Pollution*, 6:1–11, 2018.

[9] N. Mazarakis, E. Kaloudis, A. Nazos, and K.S Nikas. LES and RANS comparison of flow and pollutant dispersion in urban environment. *International Journal of Environmental Studies*, 73:48–65, 2016.

[10] M.S. Salim, R. Buccolieri, A. Chan, and S. Di Sabatino. 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, 2011.

[11] E. Aristodemou, T. Bentham, C. Pain, R. Colvile, A. Robins, and H. ApSimon. A comparison of mesh-adaptive LES with wind tunnel data for flow past buildings: Mean flows and velocity fluctuations. *Atmospheric Environment*, 43(39):6238–6253, 2009.

[12] E. Aristodemou, L.M. Boganegra, L. Mottet, D. Pavlidis, A. Constantinou, C. Pain, A. Robins, and H. ApSimon. How tall buildings affect turbulent air flows and dispersion of pollution within a neighbourhood. *Environmental pollution*, 233:782–796, 2018.

[13] J. Song, S. Fan, W. Lin, L. Mottet, H. Woodward, M. Davies Wykes, R. Arcucci, D. Xiao, J.-E. Debay, H. ApSimon, E. Aristodemou, D. Birch, M. Carpentieri, F. Fang, M. Herzog, G. R. Hunt, R. L. Jones, C. Pain, D. Pavlidis, A. G. Robins, C. A. Short, and P. F. Linden. Natural ventilation in cities: the implications of fluid mechanics. *Building Research & Information*, 46(8):809–828, 2018.

[14] H. Gough, M.-F. King, P. Nathan, C.S.B. Grimmond, A. Robins, C.J. Noakes, Z. Luo, and J.F. Barlow. Influence of neighbouring structures on building façade pressures: Comparison between full-scale, wind-tunnel, CFD and practitioner guidelines. *Journal of Wind Engineering and Industrial Aerodynamics*, 189:22–33, 2019.

[15] L. Mottet, M.-F. King, H. Gough, C.J. Noakes, J.F. Barlow, and C. Pain. Evaluating the capability of a mesh adaptivity LES-CFD software (Fluidity) to investigate the influence of staggered array on the Silsoe cube façade pressures: a comparative study between two CFD software and Full-Scale experiment. *Journal of Building and Environment.*, –:—, 2019.

[16] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, Cambridge, MA, 2003.

[17] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering.*, 82:35–45, 1960.

[18] E. Andersson, J. Haseler, P. Undén, P. Courtier, G. Kelly, D. Vasiljevic, C. Brancovic, C. Cardinali, C. Gaffard, A.Hollingsworth, C. Jakob, P. Janssen, E. Klinker, A. Lanzinger, M. Miller, F. Rabier, A. Simmons, B. Strauss, J-N.Thepaut, and P. Viterbo. The ECMWF implementation of three dimensional variational assimilation (3DVar). Part III: Experimental results. *Quarterly Journal of the Royal Meteorological Society.*, 124(550):1831–1860, 1998.

[19] D. M. Baker, W. Huang, Y.R. Guo, J. Bourgeois, and Q.N. Xiao. Three-Dimensional Variational Data Assimilation System for MM5: Implementation and Initial Results. *Monthly Weather Review*, 132:897–914, 2004.

[20] V. Mons, L. Margheri, J.-C. Chassaing, and P. Sagaut. Data assimilation-based reconstruction of urban pollutant release characteristics. *Journal of Wind Engineering and Industrial Aerodynamics*, 169:232 – 250, 2017.

[21] Jorge Sousa, Clara Garca-Snchez, and Catherine Gorl. Source apportionment and data assimilation in urban air quality modelling for no2: The lyon case study. *Building and Environment*, 132:282 – 290, 2018.

[22] R. Arcucci, L. Mottet, C. Pain, and Y.-K. Guo. Optimal reduced space for variational data assimilation. *Journal of Computational Physics*, 379:51–69, 2019.

[23] R. Arcucci, C. Pain, and Y.-K. Guo. Effective variational data as-

simulation in air-pollution prediction. *Big Data Mining and Analytics*, 1(4):297–307, 2018.

[24] Z. Belligoli, R. Dwight, and G. Eitelberg. Assessment of a Data Assimilation Technique for Wind Tunnel Wall Interference Corrections. In *AIAA Scitech 2019 Forum*, page 939, 2019.

[25] H. Kato and S. Obayashi. Integration of CFD and Wind Tunnel by Data Assimilation. *Journal of Fluid Science and Technology*, 6(5):717–728, 2011.

[26] D.Q. Zheng, J.K.C. Leung, and B.Y. Lee. An ensemble Kalman filter for atmospheric data assimilation: Application to wind tunnel data. *Atmospheric Environment*, 44:1699–1705, 2010.

[27] Imperial College London AMCG. Fluidity Manual v4.1.12. https://figshare.com/articles/Fluidity_Manual/1387713, 2015.

[28] P. Gousseau, B. Blocken, T. Stathopoulos, and G.J.F. Van Heijst. CFD simulation of near-field pollutant dispersion on a high-resolution grid: A case study by LES and RANS for a building group in downtown Montreal. *Atmospheric Environment*, 45:428–438, 2011.

[29] B. Blocken. LES over RANS in building simulation for outdoor and indoor applications: A foregone conclusion? *Building Simulation*, 11:821–870, 2018.

[30] T. Van Hooff, B. Blocken, and Y. Tominaga. On the accuracy of CFD simulations of cross-ventilation flows for a generic isolated building: Comparison of RANS, LES and experiments. *Building and Environment*, 114:148–165, 2017.

[31] Y. Tominaga and T. Stathopoulos. Ten questions concerning modeling of near-field pollutant dispersion in the built environment. *Building and Environment*, 105:390–402, 2016.

[32] F.T.M. Nieuwstadt, J. Westerweel, and B.J. Boersma. *Turbulence: introduction to theory and applications of turbulent flows.* Springer, 2016.

[33] J.W. Deardorff. A numerical study of three-dimensional turbulent channel flow at large Reynolds numbers. *Journal of Fluid Mechanics*, 41(2):453–480, 1970.

[34] J. Bardina, J. Ferziger, and W. Reynolds. Improved subgrid-scale models for large-eddy simulation. In *13^{th} Fluid and Plasma Dynamics Conference*, page 1357, 1980.

[35] T. Bentham. *Microscale modelling of air flow and pollutant dispersion in the urban environment.* PhD thesis, Imperial College London, 2004.

[36] C. Pain, A.P. Umpleby, C.R.E. De Oliveira, and A.J.H. Goddard. Tetrahedral mesh optimisation and adaptivity for steady-state and transient finite element calculations. *Computer Methods in Applied Mechanics and Engineering*, 190:3771–3796, 2001.

[37] D. Pavlidis, G.J. Gorman, J.L.M.A. Gomes, C. Pain, and H. ApSimon. Synthetic-Eddy Method for Urban Atmospheric Flow Modelling. *Boundary-Layer Meteorology*, 136:285–299, 2010.

[38] N. Jarrin, S. Benhamadouche, D. Laurence, and R. Prosser. A synthetic-eddy-method for generating inflow conditions for large-eddy simulations. *International Journal of Heat and Fluid Flow*, 27(4), 2006.

[39] S.B. Pope. *Turbulent Flows.* Cambridge University Press, 2000.

[40] H. K. Engl, M. Hanke, and A. Neubauer. Regularization of Inverse Problems. *Kluwer*, 1996.

[41] N. Nichols. *Data Assimilation - Chapter Mathematical concepts in data assimilation.* Springer, 2010.

[42] P.C. Hansen. *Rank Deficient and Discrete Ill-Posed Problems.* Society for Industrial and Applied Mathematics, 1998.

[43] L. D'Amore, R. Arcucci, L. Marcellino, and A. Murli. A Parallel Three-dimensional Variational Data Assimilation Scheme. In *AIP Conference Proceedings*, volume 1389, pages 1829–1831, 2011.

[44] A.C. Lorenc. Development of an operational variational assimilation scheme. *Journal of the Meteorological Society of Japan*, 75:339–346, 1997.

[45] J.P. Courtier. A strategy for operational implementation of 4D-VAR, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994.

[46] E.N. Lorenz. Empirical orthogonal functions and statistical weather prediction., 1956.

[47] A. Hannachi, I.T. Jolliffe, and D.B. Stephenson. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 27:1119–1152, 2007.

[48] A. Hannachi. A Primer for EOF Analysis of Climate Data. *Department of Meteorology, University of Reading, UK*, 2004.

[49] R. Arcucci, L. D'Amore, J. Pistoia, R. Toumi, and A. Murli. On the variational data assimilation problem solving and sensitivity analysis. *Journal of Computational Physics*, pages 311–326, 2017.

[50] AG Robins. Experimental model techniques for the investigation of the dispersion of chimney plumes. *Proceedings of the Institution of Mechanical Engineers*, 189(1):44–54, 1975.

[51] W.J. Coirier and K. Sura. CFD modeling for urban area contaminant transport and dispersion: model description and data requirements. In *Sixth Symposium on the Urban Environment, The 86th AMS annual meeting*, volume 163, pages 175–185, 2006.

[52] J. Franke, A. Hellsten, H. Schlnzen, and B. Carissimo. *Best practice guideline for the CFD simulation of flows in the urban environment, COST Action 732*. COST Office, 2007.

[53] M. Carpentieri, P. Salizzoni, A. Robins, and L. Soulhac. Evaluation of a neighbourhood scale, street network dispersion model through comparison with wind tunnel data. *Environmental Modelling & Software*, 37:110–124, 2012.

[54] V. Fuka, Z.-T. Xie, I.P. Castro, P. Hayden, M. Carpentieri, and A. Robins. Scalar Fluxes Near a Tall Building in an Aligned Array of Rectangular Buildings. *Boundary-Layer Meteorology*, 167(1):53–76, 2018.