BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 04/06 pp297–307
Volume 1, Number 4, December 2018
DOI: 10.26599/BDMA.2018.9020025

Effective Variational Data Assimilation in Air-Pollution Prediction

Rossella Arcucci*, Christopher Pain, and Yi-Ke Guo

Abstract: Numerical simulations are widely used as a predictive tool to better understand complex air flows and pollution transport on the scale of individual buildings, city blocks, and entire cities. To improve prediction for air flows and pollution transport, we propose a Variational Data Assimilation (VarDA) model which assimilates data from sensors into the open-source, finite-element, fluid dynamics model Fluidity. VarDA is based on the minimization 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. The conditioning of the numerical problem is dominated by the condition number of the background error covariance matrix which is ill-conditioned. In this paper, a preconditioned VarDA model is presented, it is based on a reduced background error covariance matrix. The Empirical Orthogonal Functions (EOFs) method is used to alleviate the computational cost and reduce the space dimension. Experimental results are provided assuming observed values provided by sensors from positions mainly located on roofs of buildings.

Key words: data assimilation; reduced order space; big data; preconditioning

1 Introduction and Motivation

Numerical simulations are widely used as a predictive tool to better understand complex air flows and pollution transport on the scale of individual buildings, city blocks, and entire cities. For these complex phenomena, knowledge about the state of a system and the governing physical processes is often incomplete, inaccurate, or both. The current approach in numerical modeling (which includes air pollution predictions) consists in simulating explicitly only the largest-scale phenomena, while taking into account the smaller-scale ones by means of physical parameterizations. Due to the inability to resolve the full spectrum of physical mechanisms involved as well as the fundamentally stochastic nature of the turbulent processes, all numerical models introduce uncertainty through the selection of scales and parameters that are somewhat ambiguous. Additionally, any computational methodology contributes to uncertainty due to discretization, finite precision, and the consequent accumulation and amplification of round-off errors. Taking into account these uncertainties is essential for the acceptance of any numerical simulation.

Uncertainty quantification is then permeating the science workload. The demand for predictive science results is driving the development of improved approaches for establishing levels of confidence in computational predictions using Data Assimilation (DA) methodologies. Data Assimilation is an uncertainty quantification technique used to incorporate observed data into a prediction model in order to improve numerical forecasted results[1].

There are many DA methods which have been mostly custom-developed on the forecasting model

• Rossella Arcucci and Yi-Ke Guo are with the Data Science Institute, Department of Computing, Imperial College London, London, SW7 2AZ, United Kingdom. E-mail: r.arcucci@imperial.ac.uk; y.guo@imperial.ac.uk.
• Christopher Pain is with the Department of Earth Science & Engineering, Imperial College London, London, SW7 2AZ, United Kingdom. E-mail: c.pain@imperial.ac.uk.
* To whom correspondence should be addressed.
Manuscript received: 2018-03-16; accepted: 2018-03-20

with which they are combined. Those which have gained acceptance as powerful methods in the last ten years are the variational DA approaches[2, 3] based on the minimization 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.

Variational approaches essentially implement a standard Tikhonov (or L^2) regularization[4]. To solve a VarDA problem means to compute the minimum of a Tikhonov function which includes the choice of the Tikhonov regularization parameter. The most popular DA software, which implements a VarDA model, is used to fix the regularization parameter equal to one. It means that the forecasted and the observed data have the same weight. In operational forecasting, real-time utilization of DA to improve predictions is needed. As there is insufficient time to restart a run from the beginning with new data, the information provided by observations must be incorporated on the fly. Data assimilation has to enable real-time utilization of data to improve predictions. In DA, one makes repeated corrections to data during a single run, to bring the code output into agreement with the latest data. The most popular DA software computes the minimum of the DA function by a Conjugate Gradient (CG) algorithm. The main computational kernel is then the solution of a linear system[1,5]. Caused by the background error covariance matrices, this system is strongly ill conditioned[5,6]. This mandates the introduction, in a DA software, preconditioning methods.

In summary, the necessity to run DA in real-time mandates a proper choice of numerical algorithms to regularize the ill posed problem, to compute the minimum as well as to introduce preconditioning.

2 Related Work and Contribution of the Present Work

During the last 20 years, algorithms for DA have been investigated by a number of federal research institutes and universities. Up to now, the main efforts towards the development of Variational DA systems were achieved in numerical weather prediction applications, namely by the ECMWF (European Center for Medium-Range Weather Forecasts), in Reading (UK) and by the NCAR (National Center for Atmospheric Research), in Colorado (USA). Also, variational DA models, namely IS4DVAR and NEMOVAR, have been developed for the most used ocean general circulation models: the Regional Ocean Modeling System (ROMS)[7] and the Nucleus for European Modeling of the Ocean (NEMO), respectively. These software have been mostly custom-developed on the forecasting model with which they are combined. The strong dependencies of the codes from the application domains, the data, and the type of assimilated observations do not allow a simple use of these codes in general cases.

In this paper, the problem to assimilate data to improve prediction for air flows and pollution transport is faced by a Variational DA model for the first time. Simulations are here performed using the open-source, finite-element, fluid dynamics model Fluidity (http://fluidityproject.github.io/). The details of the equations solved and their implementation can be found in Refs. [8, 9]. The state variable consists of values of pressure and velocities. Observed values of the state variable from positions mainly located on the roof of the buildings were assimilated. In operational forecasting, there is insufficient time to restart a run from the beginning with new data, then data assimilation should enable real-time utilization of data to improve predictions. This mandates the choice of an efficient method to compute the minimum of the data assimilation function. Here we adopt the L-BFGS (Limited Broyden-Fletcher-Goldfarb-Shanno) method which has been proved to be the fastest for large scale optimization problems[10]. L-BFGS method is a Quasi Newton method[11] that can be viewed as extension of conjugate-gradient methods in which the addition of some modest storage serves to accelerate the convergence rate. The convergence rate of L-BFGS depends on the conditioning of the numerical problem[10] which is dominated by the condition number of the background covariance matrix[12]. In order to reduce the ill conditioning of the background covariance matrix and remove the statistically less significant modes which could add noise to the data assimilation estimate, we use here the Empirical Orthogonal Functions (EOFs) method. EOFs implement a Truncated Singular Value Decomposition (TSVD) method. In order to improve the conditioning, only the Empirical Orthogonal Functions of the first largest eigenvalues of the error covariance matrix are considered. The EOFs (introduced by Edward

Lorenz[13]) are the eigenvectors of the error covariance matrix, its condition number is reduced as well. Even if the employment methods as the TSVD, which strongly reduce the dimension, alleviate the computational cost as they make the running less expensive, nevertheless, a consequence is that important informations are missed[14]. This issue introduces a severe drawback to the reliability of the EOFs truncation if the truncation parameter is not properly chosen. We face the problem concerning the selection of an optimal truncation parameter picked to minimize both the condition number of the problem after the preconditioning and a relative Preconditioning Error we define to provide an estimate of how much the preconditioned problem differs from the starting problem.

In summary, in developing our DA model and algorithm both Efficiency and Accuracy are been required.

- *Efficiency*: In order to alleviate the computational cost we use in this paper the EOFs method based on a TSVD method. EOFs allow to strongly reduce the dimension of the problem making the running less expensive.
- *Accuracy*: The use of EOFs reduces the ill conditioning and remove the statistically less significant modes which could add noise to the data assimilation estimate. The proper choice of the truncation parameter is usually related to a reference exact solution[15] or, it is statistical related to the variance of the full spectrum of the error covariance matrix. However, in operational Data Assimilation, the knowledge of this reference solution represents a strong condition. Also, as the error covariance matrix has a dimension $(N \times N)$ related to the size N of the domain, to compute the spectrum for big domains which require $\mathcal{O}\left((N \times N)^3\right)$ is often a too expensive operation. Here we face the problem concerning the selection of an optimal truncation parameter picked to minimize both:
  - condition number of the problem after the preconditioning;
  - a relative Preconditioning Error defined to provide an estimate of how much the preconditioned problem differs from the starting problem.

The rest of the paper is structured as follows. In Section 3 some preliminary definitions are introduced. In Section 4, the Data Assimilation problem is described and the definition of Variational approaches to solve it is presented. The Reduced order space and the preconditioning by EOFs are introduced and described in Section 5. Experimental results are shown in Section 6. Conclusion and future works are drawn in Section 7.

3 Preliminary Definitions

Here we assume some definitions we will use on next sections.

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

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_{NP} \end{bmatrix} \in \mathbf{R}^{NP \times M} \tag{1}$$

where each of NP row is a time series for a given location and $NP = [pv] \cdot np$. Let assume that each row $\boldsymbol{X}_i$ of $\boldsymbol{X}$ has mean $E[X_i] = \{m_i\}_{i=1,\dots,NP}$ and let $\boldsymbol{m} = (m_i)_{i=1,\dots,NP}$. Let

$$\boldsymbol{V} = \boldsymbol{X} - \boldsymbol{m} \in \mathbf{R}^{NP \times M} \tag{2}$$

be the deviation matrix. The variance-covariance matrix $\boldsymbol{B} \in \mathbf{R}^{NP \times NP}$ of $\boldsymbol{X}$ is defined via the expected value* of the outer product:

$$\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \tag{3}$$

Definition 2 (Singular Value Decomposition, SVD) Let $\boldsymbol{A} \in \mathbf{R}^{N \times M}$ where $M \geqslant N$ and let

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^{\mathrm{T}} \tag{4}$$

be the SVD of $\boldsymbol{A}$ where $\boldsymbol{U} \in \mathbf{R}^{N \times N}$ and $\boldsymbol{W} \in \mathbf{R}^{M \times M}$ are orthogonal (or orthonormal) matrices and

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_j)_{j=1,\dots,N} \tag{5}$$

where singular values σ_j appear in decreasing order:

$$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_N > 0 \tag{6}$$

Definition 3 (Condition number) Let $\boldsymbol{A} \in \mathbf{R}^{N \times M}$ where $M \geqslant N$ and let

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^{\mathrm{T}} \tag{7}$$

be the SVD of $\boldsymbol{A}$ in Definition 2. Then the condition number of $\boldsymbol{A}$ is such that

$$\mu(\boldsymbol{A}) = \frac{\max\{\sigma_j\}_{j=1,\dots,N}}{\min\{\sigma_j\}_{j=1,\dots,N}} \tag{8}$$

*If each vector $\boldsymbol{X}_i$ has a distribution with probability density function P, then the expected value of $\boldsymbol{X}_i$ is defined by

$$E(\boldsymbol{X}_i) = \frac{1}{M-1} \sum_{j=1,\dots,M} x_{ij}\, P(\boldsymbol{X}_j).$$

as singular values σ_j appear in decreasing order, from Formula (6), it is

$$\mu(\boldsymbol{A}) = \frac{\sigma_1}{\sigma_N} \tag{9}$$

If $\boldsymbol{A}$ is a matrix of an over-determined linear system then the discrete problem is ill posed, it is needed to filter out the contribution to the solution corresponding to the smallest singular values[15, 16]. Filtering can be introduced by recurring to the Truncated Singular Value Decomposition as given in the following definitions:

Definition 4 (Truncated Singular Value Decomposition) Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^{\mathrm{T}}$ be the SVD of $\boldsymbol{A}$ as in Formula (7). Let $\boldsymbol{\Phi}_\tau \in \mathbf{R}^{N\times N}$ be a matrix such that

$$\boldsymbol{\Phi}_\tau = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_\tau, 0, \ldots, 0) \tag{10}$$

with $1 \leqslant \tau \leqslant N$. Then the matrix

$$\boldsymbol{A}_\tau := \boldsymbol{U}\boldsymbol{\Phi}_\tau\boldsymbol{W}^{\mathrm{T}} \tag{11}$$

is the Truncated SVD (TSVD) matrix for $\boldsymbol{A}$.

4 DA Problem and the VarDA Formulation

The method we describe here is the most general VarDA method. It is called four-dimensional (4D) VarDA because it takes into account observations that are distributed in space and over an interval of time $[t_i, t_{i+\Delta t}]$. If $\Delta t = 0$, i.e., the time window is reduced to one instant, the method is called three-dimensional (3D) VarDA[1, 14, 17]. Let us give the mathematical settings describing the VarDA problem.

4.1 DA model: Set-up and problem definition

Let $\Omega \subset \mathbf{R}^3$ be a spatial domain and let

$$\begin{cases} u(t_i, x) = \mathcal{M}[u(t_j, x)], & \forall\, x \in \Omega,\ t_i, t_j \in [0, T], \\ & (t_i > t_j > 0); \\ u(t_0, x) = u_0(x), & t_0 = 0,\ \ x \in \Omega \end{cases} \tag{12}$$

be a description of the forecasting model of interest where

$$u : (t, x) \in [0, T] \times \Omega \mapsto u(t, x) \tag{13}$$

is the state function of $\mathcal{M}$. Let

$$v : (t, x) \in [0, T] \times \Omega \mapsto v(t, x) \tag{14}$$

be the observations function and

$$\mathcal{H} : \quad u(t, x) \mapsto v(t, x), \qquad \forall\, (t, x) \in [0, T] \times \Omega \tag{15}$$

denote the nonlinear observations mapping. According to the applications of model-based assimilation of observations, we will use the following discrete formulation for the VarDA problem. Given

(1) NP points of $\Omega \subset \mathbf{R}^3$: $\{x_j\}_{j=1,\ldots,NP}$;
(2) $nobs$ points of $\Omega \subset \mathbf{R}^3$, where $nobs << NP$: $\{y_j\}_{j=1,\ldots,nobs}$;
(3) N points of $[0, T]$: $\{t_k\}_{k=0,1,\ldots,N-1}$;
(4) the background estimate, i.e., vector

$$\boldsymbol{u_0} = \{u_0^j\}_{j=1,\ldots,NP} \equiv \{u(t_0, x_j)\}_{j=1,\ldots,NP} \in \mathbf{R}^{NP} \tag{16}$$

which is the state at time t_0;
(5) the operator

$$\boldsymbol{M}_{k-1,k} \in \mathbf{R}^{NP\times NP}, \quad k = 1, \ldots, N,$$

representing a discretization of a first order approximation of $\mathcal{M}$ from t_{k-1} to t_k;
(6) the vector

$$\boldsymbol{v}_k \equiv \{v(t_k, y_j)\}_{j=1,\ldots,nobs} \in \mathbf{R}^{N\times nobs}$$

consisting of the observations at t_k, for $k = 0, \ldots, N-1$;
(7) the linear operator

$$\boldsymbol{H}_k \in \mathbf{R}^{nobs\times NP},\ k = 0, \ldots, N-1$$

representing a linear approximation of the Jacobian of $\mathcal{H}$;
(8) a block diagonal matrix $\boldsymbol{G} \in \mathbf{R}^{(N\times nobs)\times(NP\times N)}$ such that

$$\boldsymbol{G} = \begin{cases} \mathrm{diag}[\boldsymbol{H}_0, \boldsymbol{H}_1\boldsymbol{M}_{0,1}, \ldots, \boldsymbol{H}_{N-1}\boldsymbol{M}_{N-2,N-1}], \\ \qquad\qquad\qquad\qquad \text{if } N > 1; \\ \boldsymbol{H}_0, \qquad\qquad\qquad\qquad \text{if } N = 1 \end{cases} \tag{17}$$

(9) the measurements error covariance matrix $\boldsymbol{R} \in \mathbf{R}^{(N\times nobs)\times(N\times nobs)}$ which describes the probability distribution function (pdf) of measurement errors. Here we assume $\boldsymbol{R}$ to be defined as follows:

$$\boldsymbol{R} = \mathrm{diag}\,(\boldsymbol{R}_k)_{k=0,\ldots,N-1},\ \boldsymbol{R}_k := \sigma_0^2\boldsymbol{I} \tag{18}$$

with $0 \leqslant \sigma_0^2 \leqslant 1$ and $\boldsymbol{I} \in \mathbf{R}^{nobs\times nobs}$ be the identical matrix.
(10) the background error covariance matrix $\boldsymbol{B} \in \mathbf{R}^{NP\times NP}$ which describes the pdf of background errors. Here we assume that $\boldsymbol{B}$, defined as in Definition 1 where $T_1 > T$, is such that

$$\boldsymbol{B} = \sigma_b^2\boldsymbol{C} \tag{19}$$

where the matrix $\boldsymbol{C}$ denoting the correlation structure of the background error, is homogeneous, and the correlations depend only on distance between states and not position, i.e.,

$$\boldsymbol{C}_{(NP,h,L)} = (c_{ij}) \tag{20}$$
$$c_{ij} = \exp\left(-\frac{1}{2}(j-i)^2 \cdot \|x_j - x_{j-1}\|_\infty^2\right),$$

with length scale $L = NP \cdot \|x_j - x_{j-1}\|_\infty$.

Given the DA problem set-up, we now define the DA inverse problem.

Definition 5 (The DA inverse problem) Given the vectors

$$\boldsymbol{v} = (\boldsymbol{v}_k)_{k=0,\dots,N-1} \in \mathbf{R}^{N \times nobs}, \quad \boldsymbol{u}_0 \in \mathbf{R}^{NP}$$

and the block diagonal matrix

$$\boldsymbol{G} \in \mathbf{R}^{(N \times nobs) \times (NP \times N)},$$

a DA problem concerns the computation of

$$\boldsymbol{u}^{DA} = (\boldsymbol{u}_k^{DA})_{k=0,\dots,N-1} \in \mathbf{R}^{NP \times N},$$

such that

$$\boldsymbol{v} = \boldsymbol{G} \cdot \boldsymbol{u}^{DA} \tag{21}$$

subject to the constraint that

$$\boldsymbol{u}_0^{DA} = \boldsymbol{u}_0 \tag{22}$$

Since $\boldsymbol{G}$ is typically rank deficient, the DA is an ill posed problem[5, 18]. In next section we define the variational formulation which leads to an unconstrained least square problem, where the term in Formula (22) ensures the existence of a solution of Formula (21).

4.2 VarDA model

In this section, descriptions of the VarDA model of the incremental VarDA model and of the preconditioned VarDA formulation are provided.

Definition 6 (The VarDA problem) VarDA problem can be described as follows:

$$\boldsymbol{u}^{DA} = \mathrm{argmin}_{\boldsymbol{u} \in \mathbf{R}^{NP \times N}} J(\boldsymbol{u}) \tag{23}$$

with

$$J(\boldsymbol{u}) = \alpha \|\boldsymbol{u} - \boldsymbol{u}_0\|^2_{\boldsymbol{B}^{-1}} + \|\boldsymbol{G}\boldsymbol{u} - \boldsymbol{v}\|^2_{\boldsymbol{R}^{-1}} \tag{24}$$

where, for any vector $\boldsymbol{w} \in \mathbf{R}^{NP}$ and $\boldsymbol{q} \in \mathbf{R}^{N \times nobs}$, $\|\boldsymbol{w}\|_{\boldsymbol{B}^{-1}} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{w}$ and $\|\boldsymbol{w}\|_{\boldsymbol{R}^{-1}} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{w}$. Parameter $\alpha > 0$ denotes the regularization parameter. In general, operational DA software assumes $\alpha = 1$. Choosing $\alpha = 1$ can be considered as giving the same relative weight to the observations in comparison to the background state.

If Formula (24) is linearized around the background state[19], the VarDA problem is formulated by the following form.

Definition 7 (The incremental VarDA problem) The incremental VarDA problem can be described as follows:

$$\delta\boldsymbol{u}^{DA} = \mathrm{argmin}_{\delta\boldsymbol{u} \in \mathbf{R}^{NP \times N}} J(\delta\boldsymbol{u}) \tag{25}$$

with

$$J(\delta\boldsymbol{u}) = \frac{1}{2}\alpha\delta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{B}^{-1}\delta\boldsymbol{u} + \frac{1}{2}(\boldsymbol{G}\delta\boldsymbol{u} - \boldsymbol{d})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{G}\delta\boldsymbol{u} - \boldsymbol{d}) \tag{26}$$

where $\boldsymbol{d} = [\boldsymbol{v} - \boldsymbol{G}\boldsymbol{u}_0]$ is the misfit, $\boldsymbol{G}$ is here the linearized observational and model operators evaluated at $\boldsymbol{u} = \boldsymbol{u}_0$ and $\delta\boldsymbol{u} = \boldsymbol{u} - \boldsymbol{u}_0$ are the increments.

In Formula (26), the minimization problem is defined on the field of increments. In order to avoid the inversion of $\boldsymbol{B}$ and to precondition the minimization of the cost function it is assumed that $\boldsymbol{B}$ can be written in the form $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$ (see Formula (3)) and the cost function can be minimized using a new variable[19].

Definition 8 (The preconditioned VarDA problem) The preconditioned VarDA problem can be described as follows:

$$\boldsymbol{w}^{DA} = \mathrm{argmin}_{\boldsymbol{w} \in \mathbf{R}^{NP \times N}} J(\boldsymbol{w}) \tag{27}$$

with

$$J(\boldsymbol{w}) = \frac{1}{2}\alpha\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \frac{1}{2}(\boldsymbol{G}\boldsymbol{V}\boldsymbol{w} - \boldsymbol{d})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{G}\boldsymbol{V}\boldsymbol{w} - \boldsymbol{d}) \tag{28}$$

where $\boldsymbol{w} = \boldsymbol{V}^{+}\delta\boldsymbol{u}$ and $\boldsymbol{V}^{+}$ denotes the generalized inverse of $\boldsymbol{V}$.

5 Reduced Order Space and Preconditioning

Some of the relevant DA operative software[2, 3] adopt the EOFs method in order to reduce the ill conditioning and remove the statistically less significant modes which could add noise to the data assimilation estimate. EOFs implement a TSVD method. In order to improve the conditioning, only the Empirical Orthogonal Functions of the first largest eigenvalues of the error covariance matrix are considered. The EOFs (introduced by Edward Lorenz[13]) are the eigenvectors of the error covariance matrix, its condition number is reduced as well. By the EOFs method, the matrix $\boldsymbol{V}$ in Eq. (28) is replaced with the matrix $\boldsymbol{V}_\tau$ which is obtained by the TSVD of $\boldsymbol{V}$ as in Eq. (11).

Even if the employment methods as the EOFs which strongly reduce the dimension, alleviate the computational cost, nevertheless, a consequence is that important informations are missed[14]. This issue introduces a severe drawback to the reliability of the EOFs truncation if the truncation parameter is not properly chosen.

The problem concerning the selection of an optimal truncation parameter is here faced. As it is known that the numerical error which propagates into the DA solution is influenced by the condition number[12], a proper value of the truncation parameter should *minimize the condition number*. However, to be sure that the preconditioned problem does not differ too much from the original problem, the optimal truncation parameter should also *minimize a Relative*

Preconditioning Error (RPE) which provides an estimate of how much the preconditioned problem differs from the starting problem as defined in Definition 9.

Definition 9 (RPE) Let E_τ be the relative Preconditioning Error which provides an estimate of how much the preconditioned problem differs from the starting problem and defined as

$$E_\tau = \frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Phi}_\tau\|_\infty}{\|\boldsymbol{\Sigma}\|_\infty} \tag{29}$$

The proper choice of the truncation parameter is usually related to a reference exact solution[15]. However, in operational Data Assimilation, the knowledge of this reference solution represents a strong condition. Here we provide an estimation of the truncation parameter which is independent from a knowledge of an exact solution.

An optimal truncation parameter σ_{opt} should be picked to minimize both:

- the condition number of V after the preconditioning[16]

$$\mu(V_\tau) \simeq \frac{\sigma_1}{\sigma_\tau} \tag{30}$$

- the Relative Preconditioning Error (Formula (29)).

In examining the asymptotic behaviour, for the condition number in Formula (30), it is

$$\lim_{\sigma_\tau \to 0} \mu(V_\tau) = +\infty, \quad \lim_{\sigma_\tau \to +\infty} \mu(V_\tau) = 0 \tag{31}$$

for the Relative Preconditioning Error, it is instead

$$\lim_{\sigma_\tau \to 0} E_\tau = \frac{\|\Sigma - I\|_\infty}{\|\Sigma\|_\infty} \simeq 1 - \frac{1}{\sigma_1},$$

$$\lim_{\sigma_\tau \to +\infty} E_\tau = \frac{\|\Sigma\|_\infty}{\|\Sigma\|_\infty} = 1 \tag{32}$$

As σ_τ is subject to the constraints[15] $\sigma_1 \geqslant \sigma_\tau \geqslant \sigma_N$, we have that, from Formula (31), the smallest value of the condition number is obtained for $\sigma_\tau \simeq \sigma_1$. From Formula (32), however, the smallest error is obtained for $\sigma_\tau \simeq \sigma_N$.

Due this difference in the asymptotic behaviour of the two functions E_τ and $\mu(V_\tau)$, an optimal value $\sigma_\tau = \sigma_{opt}$ is such that

$$\sigma_{opt} \simeq mean(\sigma_1, \sigma_N) \tag{33}$$

where $mean(\cdot,\cdot)$ denotes the mean values function. This assumption will be also experimentally validated on a consistent test case in Section 6.

6 Experimental Results

The VarDA model presented in the previous sections is applied to the pollutant dispersion within an urban environment. Hence, the VarDA model is coupled with Fluidity, an open-source, finite-element, fluid dynamic software (http://fluidityproject.github.io/). The basic Large Eddy Simulation (LES) equations describing the turbulent flows are based on the filtered incompressible Navier-Stokes equations (momentum equations and continuity of mass). The dispersion of the pollution is described by the classic advection-diffusion equation such that the concentration of the pollution is seen as a passive scalar. The equations are solved using second order schemes in time and space. Details of the equations solved and their implementations can be found in Refs. [20–22].

We consider a 2D scenario and, for our studies, we consider a domain with three buildings as we know this does not affect the generality. We set-up the problem and we face all the computational issues concerning the ill conditioning of the background covariance matrices and the distribution of the observed data.

6.1 Set-up of the test cases

A 2D case is presented in this paper and the geometry is shown in Fig. 1. This 2D case represents an idealized case used to test the ability of Fluidity to be coupled with the VarDA model and to evaluate the improvements in accuracy provided by the use of a reduced background error covariance matrix. The 2D case represents 3 buildings and the mesh includes 852 nodes (Fig. 1). The mesh is unstructured as it is implemented in operational simulations. The inlet boundary condition is a constant velocity equal to 1 m/s. No-slip boundary conditions are applied on all building façades and the bottom surface of the domain. The outlet boundary condition is defined by a zero pressure.

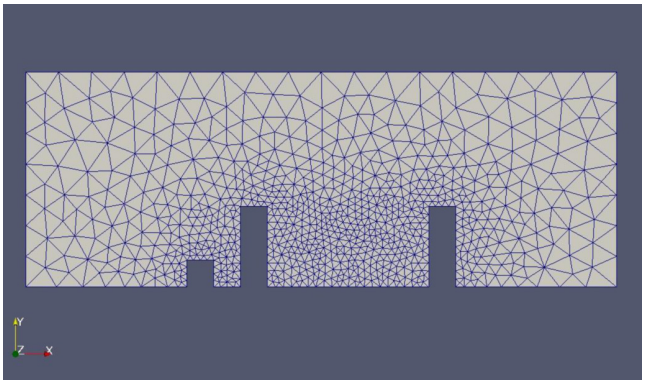

Fig. 1 2D case representing 3 buildings. The mesh is unstructured as it is implemented in operational simulations. It includes 852 nodes.

A background of pollution is set up as a sinusoidal function as expressed by Formula (34):

$$C(t) = \frac{1}{2}\left(\sin\left(\frac{2\pi t}{T}\right) + 1\right) \tag{34}$$

where C is the pollutant concentration, t is the time (in seconds) and T is the period (in seconds). Even if this background pollution is not based on real data, it mimics waves of pollution in an urban environment. The kinematic viscosity is set equal to $1.10^{-5}\,\mathrm{m}^2/\mathrm{s}^{-1}$.

We set-up our Data Assimilation problem following the points described in Section 4.1, then we have fixed:

- NP which is the number of grid points such that $NP = 852$, then the complexity of our test case is $\mathcal{O}(10^8)$;
- we assume the observations given by sensors on the roofs of the three buildings and we consider $nobs$ (which is the number of observed data) such that
  - $nobs = 6$, i.e., we have chosen few data from sensors, just two grid points on each roof of the three buildings;
  - $nobs = 60$, i.e., a reasonable number of data from sensors, twenty grid points on each roof of the three buildings;
  - $nobs = 852$, i.e., we assume data from sensors in all the grid points;
- for the time steps, we have assumed $N = 1$ and $M = 300$;
- the operator $\mathcal{M}$ is provided by FLUIDITY;
- the background u_0 is obtained by truncating the resulting data from FLUIDITY;
- the error covariance matrix $\boldsymbol{R} = \bar{\sigma_o}\boldsymbol{I}$ with $\bar{\sigma_o} = 0.5$;
- the background error covariance matrix such that $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$ and we have computed matrix $\boldsymbol{V}$. by considering a temporal sequence of data collected by FLUIDITY. Then we have applied the EOFs regularization method and we have computed the condition number of V_τ as function of τ.

In this section an evaluation of the results has been provided in term of

- Section B: Improvement in conditioning by using the background error deviance matrix $\boldsymbol{V}$ instead of the background error covariance matrix $\boldsymbol{B}$ into the VarDA formulation and introducing the reduced dimension matrix $\boldsymbol{V}_\tau$;
- Section C: the trend of the error defined as distance of the solution computed by the VarDA with $\boldsymbol{V} = \boldsymbol{V}_\tau$ and a control variable u_C:

$$u^{DA} - u_C \tag{35}$$

The error is evaluated for different numbers of observed data $nobs = 6$, $nobs = 60$, and $nobs = 852$.

6.2 Reduced background error covariance matrix and choice of the truncation parameter τ

Figure 2 shows the spectrum of the background error covariance matrix, such that, the computed condition number is $\mu(\boldsymbol{B}) = 1.191\,599\,809\,890\,142\mathrm{e}+17$. Figure 3, instead, shows the spectrum of the background error deviance matrix and Fig. 4 shows the strong improvement in conditioning for the background error deviance matrix $\boldsymbol{V}$ with respect the background error covariance matrix $\boldsymbol{B}$.

The trend of the computed condition number in Fig. 4

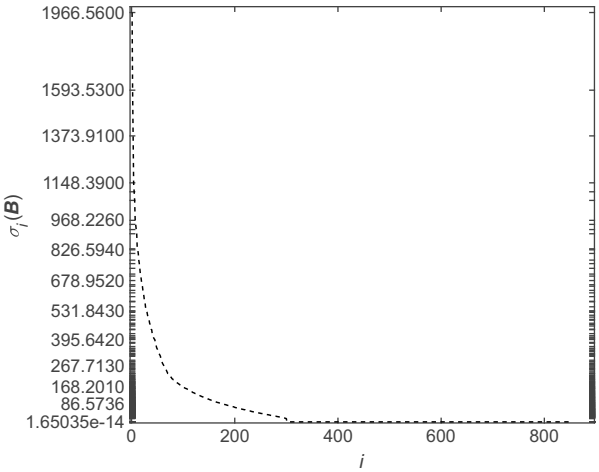

Fig. 2 Spectrum of the background error covariance matrix $\boldsymbol{B}$.

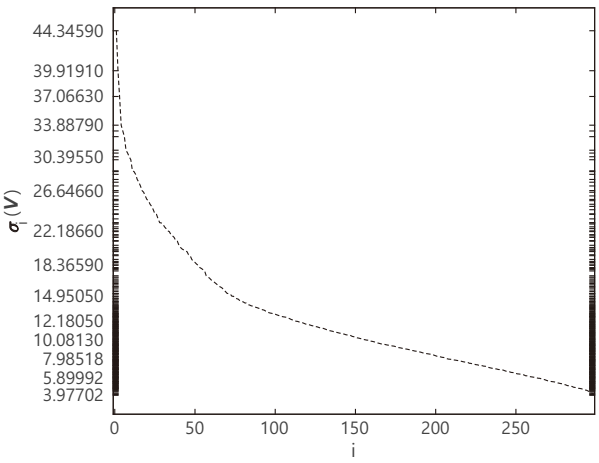

Fig. 3 Spectrum of the background error deviance matrix $\boldsymbol{V}$.

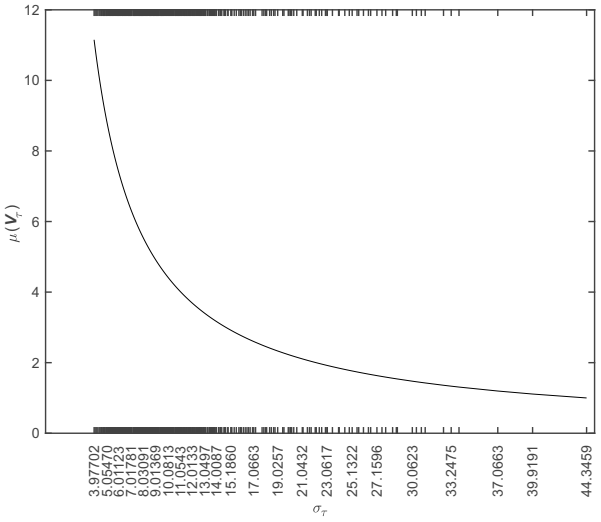

Fig. 4 Condition numbers of the reduced dimension matrices V_τ obtained by EOFs for different values of σ_τ.

confirms the qualitative evaluation of the asymptotic behaviour provided in Formula (31). Figure 5, instead, shows the values of the relative Preconditioning Error defined in Formula (29). Also in this case, the trend of the values confirms the qualitative evaluation of the asymptotic behaviour provided in Formula (32).

Even if the employment methods as the EOFs which strongly reduce the dimension, alleviate the computational cost, nevertheless, a consequence is that important informations are missed[14]. This issue introduces a severe drawback to the reliability of the EOFs truncation if the truncation parameter is not properly chosen.

Figure 6 and Fig. 7 show that, for small value of τ,

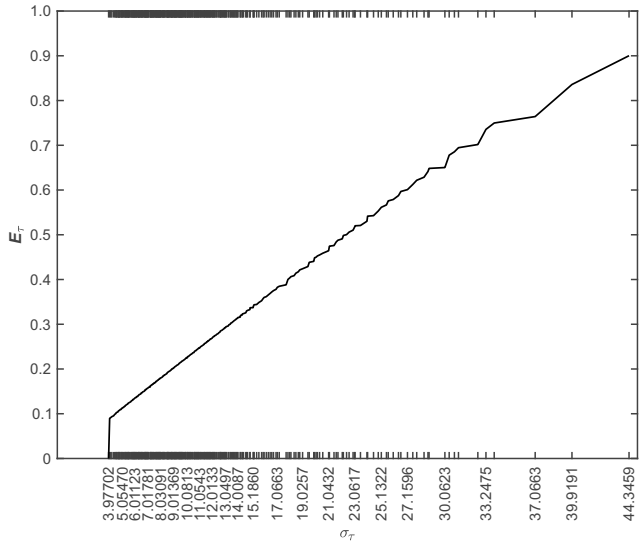

Fig. 5 Relative error EOFs.

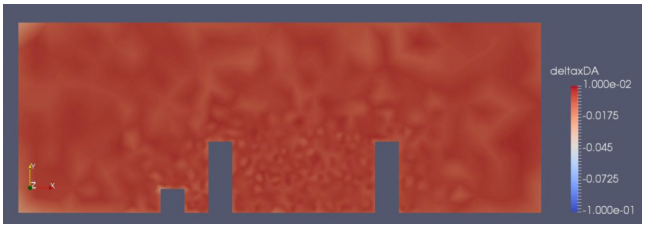

Fig. 6 Results of the VarDA algorithm with τ=5 EOFs for Pressure field.

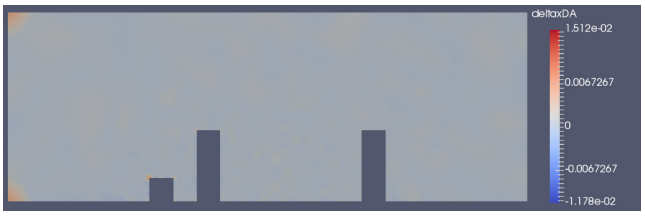

Fig. 7 Results of the VarDA algorithm with τ=5 EOFs for Velocity field.

the numerical error propagates into the solution such that we do not have any impact of the observed data. Comparing results shown in Figs. 8 and 9 with Figs. 10 and 11, we observe that the choice of τ which satisfy

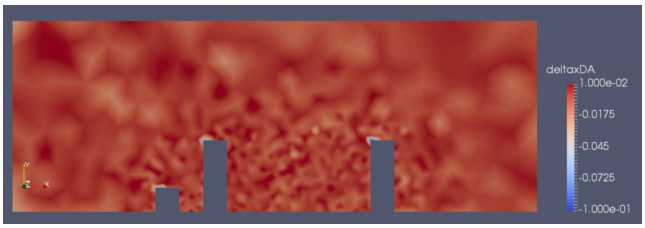

Fig. 8 Results of the VarDA algorithm with τ=155 EOFs for Pressure field.

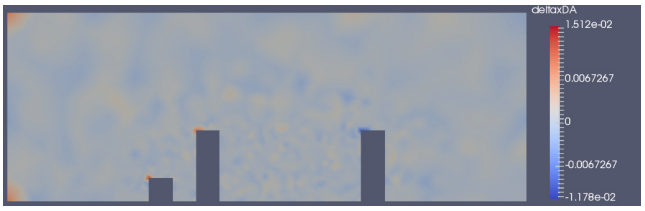

Fig. 9 Results of the VarDA algorithm with τ=155 EOFs for Velocity field.

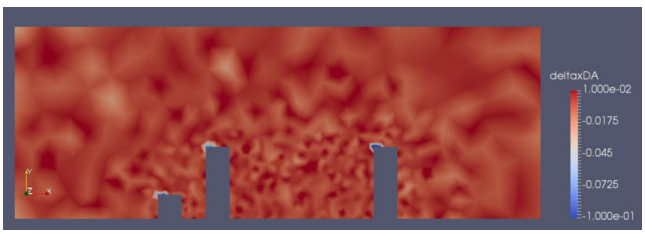

Fig. 10 Results of the VarDA algorithm with τ =295 EOFs for Pressure field.

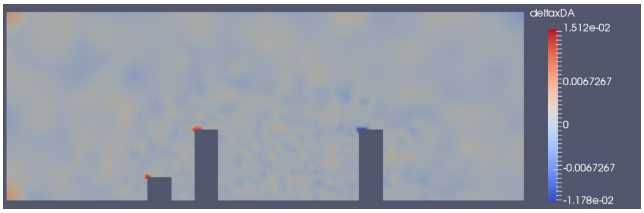

Fig. 11 Results of the VarDA algorithm with τ =295 EOFs for Velocity field.

the condition in Formula (33) allows us to properly assimilate observed data even if the problem is solved in a reduced dimension space, i.e., by alleviating the computational cost.

6.3 Results

From the evaluations provided in the previous section, we assume here the value of the parameter τ such that the condition (Formula (33)) is satisfied, i.e., $\tau = 155$. We evaluate the error as defined in Formula (35) for different numbers of observed data: $nobs = 6$, $nobs = 60$, and $nobs = 852$ as described in Figs. 12–14. Figures 15–17 confirm our expectation, i.e., they

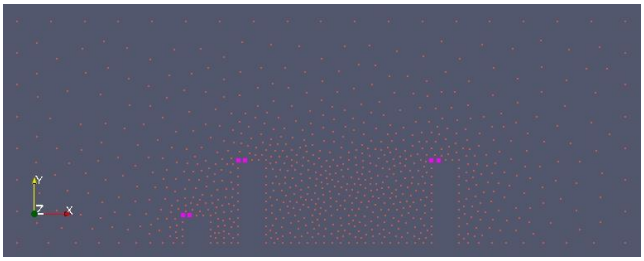

Fig. 12 Number of observations *nobs* = 6.

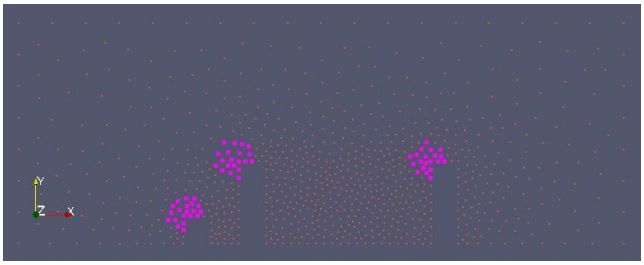

Fig. 13 Number of observations *nobs*=60.

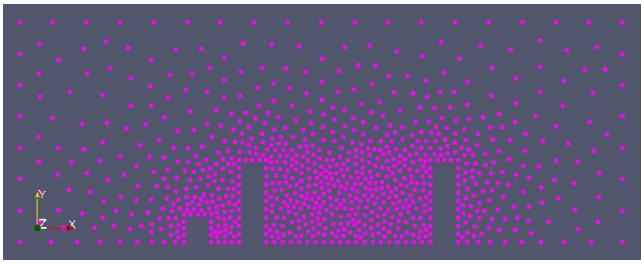

Fig. 14 Number of observations *nobs*=852.

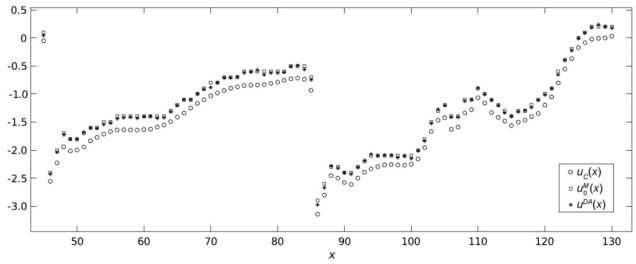

Fig. 15 Results comparison with a control variable u_C for *nobs* = 6.

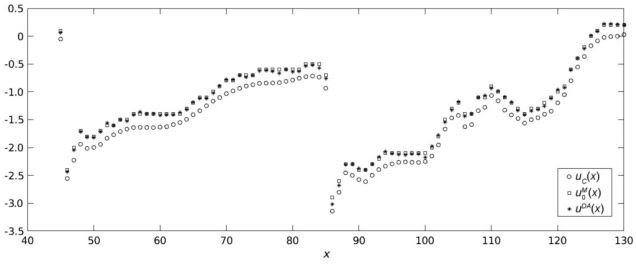

Fig. 16 Results comparison with a control variable u_C for *nobs*=60.

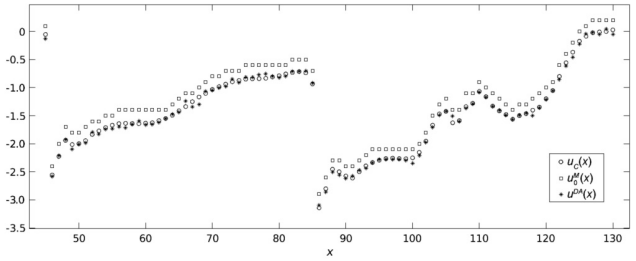

Fig. 17 Results comparison with a control variable u_C for *nobs*=852.

show that the error (i.e., the distance of the DA solution by the control variable) strongly decreases as the number of observed data increase.

7 Conclusion

Numerical issues faced in developing a VarDA algorithm include the ill-conditioning of the background covariance matrix and the choice of the regularization parameter. The EOFs method has been here used in order to reduce the ill-conditioning and remove statistically less significant modes that could add noise to the data assimilation estimate. EOFs strongly reduce the dimension, alleviating the computational cost as they make the running less expensive, but a consequence is that important information can be missed. This can be a severe drawback in the reliability of the EOFs truncation if the regularization parameter τ is not properly chosen.

We proved that an optimal regularization parameter is the mean value of the maximum and minimum singular values of the background error covariance matrix. Results provided show this choice allows minimization of the running time without significant loss in the solution accuracy. The forecast data were produced by Fluidity and the state variable consists of values of pressure and velocities. Observed values of the state variable from sensors located on the top of the three buildings were assimilated. We have seen that for small value of τ, the numerical error propagates into the solution with impact of the observed data and that the choice of τ as the mean value allows the observed data to be assimilated, even if the problem is solved in a reduced dimension space, i.e., by alleviating the computational cost.

Acknowledgment

This work was supported by the EPSRC Grand Challenge grant "Managing Air for Green Inner Cities" (MAGIC) EP/N010221/1.

References

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

[2] 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, *Mon. Wea. Rev*., vol. 132, pp. 897–914, 2004.

[3] E. Andersson, J. Haseler, P. Undén, P. Courtier, G. Kelly, D. Vasiljevic, C. Brancovic, C. Cardinali, C. Gaffard, A.Hollingsworth, et al., The ECMWF implementation of three dimensional variational assimilation (3DVar), Part III: Experimental results, *Quarterly Journal Royal Met. Society*, vol. 124, pp. 1831–1860, 1998.

[4] R. Arcucci, L. D'Amore, L. Carracciuolo, G. Scotti, and G. Laccetti, A decomposition of the tikhonov regularization functional oriented to exploit hybrid multilevel parallelism, *International Journal of Parallel Programming*, vol. 45, no. 5, pp. 1214–1235, 2017.

[5] J. N. Nichols, Mathematical concepts in data assimilation, in *Data Assimilation,* W. Lahoz, B. Khattatov, and R. Menard, eds. Springer, 2010.

[6] J. S. A. Haben, Conditioning and preconditioning of the minimisation problem in variational data assimilation, PhD dissertation, University of Reading, UK, 2011.

[7] A. M. Moore, H. G. Arango, G. Broquet, B. S. Powell, A. T. Weaver, and J. Zavala-Garay, The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I – System overview and formulation, *J. Progress in Oceanography*, vol. 91, pp. 34–49, 2011.

[8] C. 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*, vol. 190, pp. 3771–3796, 2001.

[9] D. R. Davies, C. R. Wilson, and S. C. Kramer, Fluidity: A fully unstructured anisotropic adaptive mesh computational modeling framework for geodynamics, *Geochemistry, Geophysics, Geosystems*, vol. 12, no. 6, 2011.

[10] D. C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, vol. 45, pp. 503–528, 1989.

[11] J. Nocedal, R. H. Byrd, P. Lu, and C. Zhu, L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, *ACM Transactions on Mathematical Software*, vol. 23, pp. 550–560, 1997

[12] 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*, vol. 335, pp. 311–326, 2017.

[13] E. N. Lorenz, Empirical orthogonal functions and statistical weather prediction, Statistical Forecasting Project, MIT, Sci. Rep. No. 1, 1956.

[14] D. G. Cacuci, I. M. Navon, and M. Ionescu-Bujor, *Computational Methods for Data Evaluation and Assimilation*. CRC Press, 2013.

[15] C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems, Numerical Aspects of Linear Inversion*. SIAM, 1998.

[16] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*. SIAM, 2006.

[17] J. P. Courtier, A strategy for operational implementation of 4D-VAR, using an incremental approach, *Q. J. R. Meteorol Soc.*, vol. 120, 1367-1387, 1994.

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

[19] A. C. Lorenc, Development of an operational variational assimilation scheme, *Journal of the Meteorological Society of Japan*, vol. 75, pp. 339-346, 1997.

[20] E. Aristodemou, T. Bentham, C. Pain, and A. Robins, A comparison of mesh-adaptive LES with wind tunnel data for flow past buildings: Mean flows and velocity fluctuations, *Atmospheric Environment Journal*, vol. 43, pp. 6238–6253, 2009.

[21] R. Ford, C. C. Pain, A. J. H. Goddard, C. R. E. De Oliveira, and A. P. Umpleby, A nonhydrostatic finite-element model for three-dimensional stratified oceanic flows. Part I: Model formulation, *Monthly Weather Review*, vol. 132, pp. 2816–2831, 2004.

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

Rossella Arcucci received a Master Degree (cum laude) in Mathematics in 2008 from the University of Naples Federico II, Italy, and received the PhD degree in Computational and Computer Science from the same university in 2012. She is a Research Associate of the Data Science Institute (DSI) at Imperial College London (ICL) in UK. Her area of expertise is in Numerical Analysis, Scientific Computing and development of methods, algorithms and software for scientific applications on high performance architectures including parallel and distributed computing. She works on numerical and parallel techniques for accurate and efficient Data Assimilation by exploiting the power of machine learning models. Efficiency is achieved by virtue of designing models specifically to take full advantage of massively parallel computers and general purpose graphics processing units. During her post doc at University of Naples Federico II in Italy, she coordinated the H2020-RISE-2015-NASDAC project as PI until September 2017, when she joined the DSI. She received the acknowledgement of Marie Sklodowska-Curie fellow from European Commission Research Executive Agency in Brussels on the 27th of November 2017.

Christopher Pain received a B.Sc. degree in Mathematics in 1988, from University of Reading, UK, an M.Sc. degree in Mathematical Techniques for CAD in 1989 from Cranfield University, UK, and received a Ph.D. degree in Computational Fluid Dynamics in 1991 from University of Exeter, UK. He leads the Applied Modelling and Computation Group (AMCG) at ICL — the largest research group at Imperial College, comprises of about 60 scientists and recipient of the ICL Research Excellence award in 2011. Award in recognition of its high academic achievement and significant future potential. He is visiting Prof at BU and USC, the director of the data assimilation (DA) lab. in the Data Science Institute (DSI) at ICL, leads the modelling for the MEMPHIS and MAGIC consortia and is PI of SmartGeoWells Newton consortium. He developed new balanced mixed finite elements for multi-phase flow and geophysical fluid dynamics. $>$ 180 journal papers, graduated 42 PhD students, completed 42 industry and research council grants, and won £23.5M in funding over 15 years.

Yi-ke Guo received a first-class honours degree in Computing Science from Tsinghua University, China, in 1985 and received the PhD degree in Computational Logic from Imperial College in 1993 under the supervision of Professor John Darlington. He founded InforSense, a software company for life science and health care data analysis, and served as CEO for several years before the company's merger with IDBS, a global advanced R&D software provider, in 2009. He is founding director of DSI at ICL as well as leading the Discovery Science Group in the department. Professor Guo also holds the position of CTO of the tranSMART Foundation, a global open source community using and developing data sharing and analytics technology for translational medicine. His research area is in data analysis and e-science. He has published over 200 papers on data analysis/applications, educated 80 PhD students & won £120M over 15 years. The projects he has contributed to have been internationally recognised, including winning the "Most Innovative Data Intensive Application Award" at the Supercomputing 2002 conferencefor Discovery Net, and the Bio-IT World "Best Practices Award" for U-BIOPRED in 2014. He is a Senior Member of the IEEE and is a Fellow of the British Computer Society.