PBDW: A non-intrusive Reduced Basis Data Assimilation method and its application to an urban dispersion modeling framework

Janelle Katherine Hammond, Rachida Chakir, Frédéric Bourquin, Yvon Maday

▶ **To cite this version:**

Janelle Katherine Hammond, Rachida Chakir, Frédéric Bourquin, Yvon Maday. PBDW: A non-intrusive Reduced Basis Data Assimilation method and its application to an urban dispersion modeling framework. Biotechnology, 2019, 76, pp 1-25. 10.1016/j.apm.2019.05.012 . hal-02186277

HAL Id: hal-02186277
https://hal.science/hal-02186277

Submitted on 26 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PBDW: a non-intrusive Reduced Basis Data Assimilation Method and its application to outdoor Air Quality Models

J.K. Hammond[a,*], R. Chakir[a], F. Bourquin[a], Y. Maday[b,c]

[a] *Université Paris Est, IFSTTAR, 10-14 Bd Newton, Cité Descartes, 77447 Marne La Vallée Cedex, France*
[b] *Sorbonne Universités, UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005 Paris, France*
[c] *Institut Universitaire de France and Division of Applied Mathematics, Brown University, Providence, RI, USA*

Abstract

The challenges of understanding the impacts of air pollution require detailed information on the state of air quality. While many modeling approaches attempt to treat this problem, physically-based deterministic methods are often overlooked due to their costly computational requirements and complicated implementation. In this work we extend a non-intrusive reduced basis data assimilation method (known as PBDW state estimation) to large pollutant dispersion case studies relying on equations involved in chemical transport models for air quality modeling. This, with the goal of rendering methods based on parameterized partial differential equations (PDE) feasible in air quality modeling applications requiring quasi-real-time approximation and correction of model error in imperfect models. Reduced basis methods (RBM) aim to compute a cheap and accurate approximation of a physical state using approximation spaces made of a suitable sample of solutions to the model. One of the keys of these techniques is the decomposition of the computational work into an expensive one-time offline stage and a low-cost parameter-dependent online stage. Traditional RBMs require modifying the assembly routines of the computational code, an intrusive procedure which may be impossible in cases of operational model codes. We propose a less intrusive reduced order method using data assimilation for measured pollution concentrations, adapted for consideration of the scale and specific application to exterior pollutant dispersion as can be found in urban air quality studies. Common statistical techniques of data assimilation in use in these applications require large historical data sets, or time-consuming iterative methods. The method proposed here avoids both disadvantages. In a case study presented in this work, the method allows to correct for unmodeled physics and treat cases of unknown parameter values, all while significantly reducing online computational time.

Keywords: Reduced Basis method, Model order reduction, Parameterized partial differential equations, dispersion modeling, Variational data assimilation.

*Corresponding author
Email address: hammond.janelle@gmail.com (J.K. Hammond)

1. Introduction

With the urbanization of world populations and estimations of millions of deaths caused yearly by air pollution [1], air quality modeling is of increasing interest. The need for improved approximation and model reduction is particularly pertinent in these applications, modeling complex and not-fully-known physics. Our focus here is on a new technique combining Model Order Reduction (MOR) and variational data assimilation, which is generally still developmental in the context of air quality modeling, but can provide insight into these complex flux by allowing more practical and feasible study, in terms of computational costs and from imperfect input data and models. In this paper we propose an extension of a new Reduced Basis Data Assimilation technique, previously applied to small-scale experimental problems, applied to a problem of pollutant dispersion in order to treat the practical problems associated to MOR and data assimilation of these flux involved in many sophisticated methods of urban air quality modeling.

Many air quality modeling techniques exist, from statistical and empirical, to deterministic methods [2]. Within the category of deterministic models, approaches vary in sophistication from simple box models [3], to Gaussian plume models, to physically-based Lagrangian methods [4] and Eulerian CFD models [5, 6, 7]. The more sophisticated models, when applied with precise information on the environment and emissions, and if correctly calibrated, can provide very detailed information on spatial and time-varying pollutant concentrations, as well as the physical phenomena affecting air quality. These models commonly rely on PDE dispersion modeling, which will be our focus here. However, these models can be computationally expensive to solve. Additionally, given the complexity of real-world applications, we cannot assume that even a highly informed and sophisticated deterministic (or non-deterministic for that matter) model can exactly represent all the physical phenomena at play. Therefore, the combination of model order reduction methods and data assimilation methods is of great interest to these complicated and pertinent applications.

In various data assimilation methods, the goal is to use the *a priori* information encoded in the best model possible, and available data, to find the most precise approximation of the physical system. A common concept in meteorological forecasting, data assimilation requires a set of observations of the state, a mathematical model, and a data assimilation scheme. Many data assimilation methods involve the minimization of a cost function, such as least-squares type, designed to compute the mismatch between the model approximation and the observations. Common techniques include Bayesian techniques [8], in which the cost function minimizes the expected value of the mismatch and depends on the conditional probability of the random variable u (which corresponds to the state estimate in deterministic models) given the observation data. This probability is commonly estimated using an iterative gradient method, and requires more significant historical data sets. The Best Linear Unbiased Estimator [9] method also minimizes the expected value of the mismatch, relying on knowledge of error covariance matrices, which also demands large quantities of historical data and iterative implementation. The Kalman Filter [10] recursively computes corrected state estimation using data and the model estimate from a previous time step, for real-time data assimilation, however also relies on error covariance from historical data, and each step

requires the construction of a gain filter matrix, which could limit the possibility of real-time calculation. Kriging [11] is a group of interpolation methods relying strongly on the contribution of the data set, in which weights of the interpolation depend on a formula derived by minimizing the variance of the prediction error requiring the covariance matrix based on a large data set. A common drawback of these statistical methods is the necessity of a historical dataset and a strong dependence of the state estimate on observational data. Variational data assimilation techniques rely more heavily on the model, and impose it as a constraint on the cost function. For example, the adjoint method [12, 13] is a typical method to treat the reconstruction of a physical state involving the minimization of a cost function to optimize the parameters of the model with respect to the measurement data. A sensitivity analysis of the adjoint problem for air quality models can be found in [14]. The standard adjoint method, however, relies solely on model precision and uses data to correct the inputs, not the state. 4D-Var can include a term which corrects model error in the state. Both methods have the disadvantage of requiring the iterative computation of the forward and adjoint problems, and in the case of a non-linear wind field as a parameter the computation of the adjoint solution is non-trivial. In [15], a Proper Orthogonal Decomposition (POD) representation of an ensemble of forward problem state estimations is computed around an initial guess input parameter vector, and used to approximate the model outputs in the cost function. This renders the cost function quadratic and solvable by non-iterative means. This method avoids the adjoint problem, however still requires the computation of numerous model estimates, and relies solely on model precision, using data to correct the inputs in the context of source identification.

These methods require the forward resolution of the model for many parameter values, which can prove costly; MOR methods can offer highly advantageous reduction of computational effort without significant loss of precision. A common approach to rapidly compute reliable approximations of solutions to complex parameter-dependent problems is by projection-based reduction methods, such as reduced basis methods (RBM) [16]. These methods aim to reduce the complexity of the model using the information given by a well-chosen set of particular solutions to the problem. A basis (called the reduced basis) of a low-dimensional subspace of the space representing all the solutions to the parametrized problem, is constructed from these particular solutions. The equations of the full model are projected onto the reduced basis space by a Galerkin method. Examples of reduced basis methods used in the adjoint problem framework can be found in [17, 18, 19], and specifically in the case of air quality modeling in [20, 21]. RBMs used for 4D-Var data assimilation on an advection-diffusion model are presented in [22].

One of the drawbacks of standard variational data assimilation methods is that it is intrusive from a computational point of view, requiring the development of an adjoint calculation code, despite efforts to automatically differentiate a given software. In some cases this could mean relatively small modification of the original calculation code, while in others more significant modifications could be required. For example, when the wind field is a varying parameter in the model, the implementation of the adjoint method would require the reconstruction of the wind field at each iteration during the approximation of the optimal parameter (i.e. for each approximation of the adjoint solution). For these reasons, less intrusive options can be valuable. The method used in [15] is less

intrusive, however has not been applied with reduced order models and is designed with the goal of source and parameter identification rather than improving the representation of pollutant concentrations.

The Parameterized-Background Data-Weak (PBDW) state estimation method [23, 24] can represent the physics of the state using a sophisticated model, and applies non-intrusive and non-iterative real-time variational data assimilation employing RBMs, with correction of model error, and not requiring a large database of observations. The PBDW relies on the knowledge of some particular solutions to the parameterized model, and some measurements over the physical state to be approximated. The weak formulation of the PBDW method is based on least-squares approximation, as is the case of the adjoint inverse method and many variational data assimilation methods. In this paper we will apply this non-intrusive reduced basis method of data assimilation for parameterized PDEs modeling particulate matter dispersion. Given a parameterized model for a physical system, which we will refer to as the "best-knowledge" (bk) model, and a number of measurements of the state we wish to approximate, we employ the PBDW method to achieve the best possible approximation by a formulation actionable in real-time. Our decision to treat the wind field as a parameter of a pollutant transport model is of particular advantage in the context of dispersion modeling, by which we avoid online solution of the wind field. In order to extend the PBDW to air quality problems modeling complex dispersion phenomena involved in air quality modeling problems, we propose placement of observatational sensors by a technique adapted to RBM-based data assimilation, using a double-Greedy algorithm derivied from the Generalized Empirical Interpolation Method (GEIM) [25, 26]. The GEIM is another non-intrusive and non-iterative method combining MOR and data assimilation, in which an empirical interpolation is constructed from knowledge of particular solutions and measurement data. We will also introduce a modified H^1-norm in the PBDW formulation in order to address dimensionality problems induced by large-scale calculation domains inherent to urban dispersion modeling and small pollutant sensor sizes.

In section 2 we will present the application in particulate dispersion modeling, in section 3 the mathematical formulation of the PBDW method, and in section 4 we will discuss important factors in the numerical implementation of the PBDW method. In section 5 we will show through numerical application that the PBDW method succeeds in the reconstruction of a concentration field on the case study considered for well-chosen sensor locations. We will also show a comparison of the PBDW state estimation to the GEIM method, demonstrating that the PBDW method outperforms the GEIM method when model error is present. We finally give computational times required for state estimation on this case study, showing the significant advantages of the RB technique in the PBDW method, and compare the process to that of the adjoint method for the current case study and a more complicated urban domain in section 5.3

2. A Case study in Dispersion modeling

The application studied in this work represents a simplified outdoor urban scenario of particulate air pollution. In this section we will first explain the geometry of the test domain considered for this case study, then describe our best-knowledge mathematical model, and finally set the reduced basis framework to this model.

2.1. Physical problem formulation

Let us consider a physical system described by a PDE, and denote $\mathbf{p}$ the parameter configuration of the physical system, encoding information such as operation conditions (e.g. emissions or frequency), environmental factors (e.g. temperature), or physical components. Let $\mathbf{p} \in \mathcal{D}$, where $\mathcal{D}$ is the set of all parameters of interest, and a bounded domain $\Omega \subset \mathbb{R}^d$. We will assume a solution space $\mathcal{X}$, a Hilbert space, such that $H_0^1(\Omega) \subset \mathcal{X} \subset H^1(\Omega)$, and associated inner product $(\cdot, \cdot)_{\mathcal{X}}$. We will denote $\mathcal{X}'$ its dual space.

We study here a simple two-dimensional domain of dimensions $75m \times 120m$, seen in Figure 1. The domain represents a neighborhood with a house, a building, and pollution source of a street. These choices were made to give a simplified case study representing a residential area with particulate pollution as would be produced by road traffic.

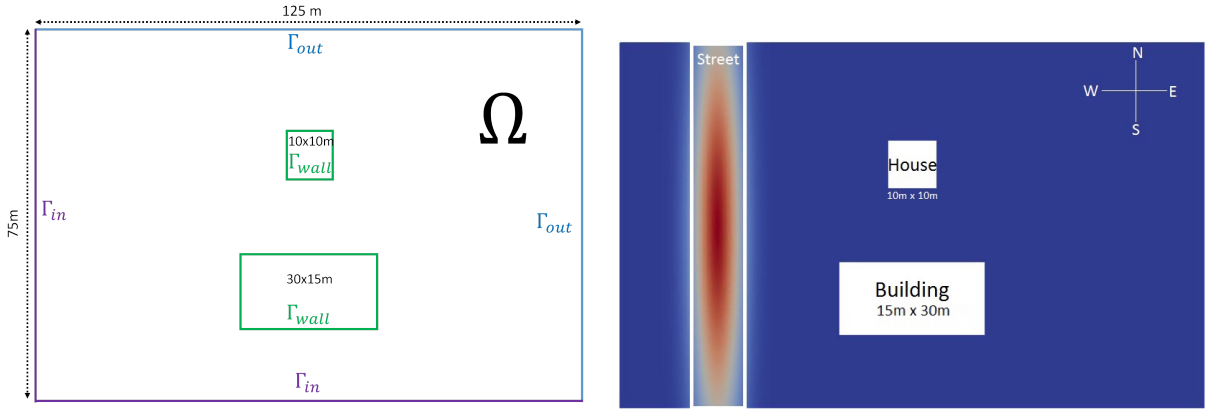

Figure 1: Two-dimensional test domain with boundaries corresponding to the velocity field (left) and particulate pollution source representing a street (right), residential character represented by a house and a building.

We chose a particulate pollutant $PM_{2.5}$ (particulate matter of diameter $d \leq 2.5\mu m$) in this study, which on the short term can be considered to have negligible reaction. We set wind velocities (in a fixed direction $(1,1)^T$) up to force 1 as the varying parameter in the best-knowledge parameter space $\mathcal{D}^{bk} \subset \mathcal{D}$, and set source intensity representing varying traffic of 1×10^{-3} and 1×10^{-2} $\frac{mg}{m^3 \cdot s}$.

In order to compromise between accuracy, numerical stability, and computational time in the wind field provided to the pollutant transport model, we use pseudo-steady-state CFD wind fields, solutions to Reynolds-Averaged Navier-Stokes with $k-\epsilon$ turbulence by *Code_Saturne* [27] (a general purpose finite-volume CFD software). While other turbulence models such as RNG $k-\epsilon$ may show improved modeling of flow around buildings in recirculation zones [28, 29], we chose the standard $k-\epsilon$ for it's universal applicability, good numeric stability, and availability in general purpose CFD software. This study could, of course, be done with different turbulence models and CFD software (e.g. LES or DNS software [30]) in future work, as the versatility of the non-intrusive method is among its strongest advantages. The grid resolution near obstacles was fixed with consideration of y^+ values. Time stepping was done with a reference time step value of $0.1s$ in *Code_Saturne*, which was found in practice to be a good compromise between numerical stability and computation time. Boundary conditions on explicitly outflow boundaries are set to homogeneous Neumann conditions, and on all other boundaries we impose an inflow in direction $(1,1)^T$. The CFD model can be coupled with transport equations, or precalculated for a decoupled procedure. In our study we chose to decouple the computation of the wind fields, and then used the velocity and turbulent viscosity fields in the dispersion model. This allowed us to use a larger domain for wind field calculation with a *buffer* zone between the obstacles and outflow boundaries of $15L$, where L is some characteristic length of the obstacle.

For our case study, we consider a simple stationary advection-diffusion PDE as our best-knowledge parametrized transport model $\mathcal{P}^{bk}$: Find $c^{bk}(\mathbf{p}) \in \mathcal{X}$ such that

$$\begin{cases} \rho\,\vec{v}(\mathbf{p})\cdot\nabla c^{bk}(\mathbf{p}) - \operatorname{div}\big(\epsilon_{tot}(x)\,\nabla c^{bk}(\mathbf{p})\big) &= \rho\,F_{src}(\mathbf{p}) & \text{in } \Omega, \\ c^{bk}(\mathbf{p}) &= 0 & \text{on } \Gamma_D = \{x\in\partial\Omega \,|\, \vec{v}(x)\cdot\vec{n} < 0\}, \\ \epsilon_{tot}\nabla c^{bk}(\mathbf{p})\cdot\vec{n} &= 0 & \text{on } \Gamma_N = \partial\Omega\setminus\Gamma_D, \end{cases} \tag{1}$$

where $\rho = 1.225\frac{kg}{m^3}$ is the density of the air, $\vec{v}$ is the wind field, F_{src} the pollutant source term. Considering turbulent (or eddy) diffusion $\epsilon_{turb} = \frac{\nu_F}{s_c}$, where ν_F is the turbulent viscosity and $s_c = 0.7$ the dimensionless Schmidt number, the total diffusion is thus $\epsilon_{tot} = \epsilon_{mol} + \epsilon_{turb}$, with $\epsilon_{mol} = 1.72\times10^{-5}\frac{m^2}{s}$ the molecular diffusion in air. The (strict) inflow boundary is denoted by $\Gamma_D = \Gamma_{in}$ and $\Gamma_N = \Gamma_{wall}\cup\Gamma out$ represents non-inflow boundaries.

$$\begin{cases} c = c_0 & \text{on } \Gamma_D = \{x\in\partial\Omega | \vec{v}(x)\cdot\vec{n} < 0\} \\ -\epsilon_{turb}(x)\frac{\partial c}{\partial z} = -\epsilon_{turb}(x)\nabla c\cdot\vec{n} = 0 & \text{on } \Gamma_N = \partial\Omega\setminus\Gamma_D \end{cases} \tag{2}$$

Problem (1) is solved in FreeFem++ [31] by the finite element method over $\mathcal{N}_h$ degrees of freedom, combined with a SUPG stabilization method [32, 33] to avoid numerical instabilities known to affect transport problems solved by finite element methods. The resolution $\mathcal{N}_h$ of the finite element problem is sufficiently fine to assume that the concentration field $c^{bk}(\mathbf{p}) = c_h^{bk}(\mathbf{p})$ is assumed to commit minimal discretization error (with respect to the errors we will see by model reduction).

The use of this simple dispersion model and our own calculation code has the advantage of treating the predominant physical effects involved in AQMs, while allowing us full control over the experiment to study the results of the method with knowledge of the true solution. This case study aims to provide a proof of concept, and is not fully compatible with real-world pollution data. We will thus work with synthetic data (as seen in e.g. [15]) , taken as the sensor function outputs over a *trial* state estimate, for both the model in equation 1 and in section 5.1. Using synthetic data allows us to better study the results of the mathematical method by comparing to fully-resolved true solutions, as opposed to sparse data measurements.

2.2. Reduced basis background

Reduced basis methods exploit the parametrized structure of our problem and construct a low-dimensional approximation space representing the manifold of solutions, $\mathcal{M}^{bk} = \{c^{bk}(\mathbf{p}) \in \mathcal{X} \mid \mathbf{p} \in \mathcal{D}^{bk}\}$, to the parameterized model $\mathcal{P}^{bk}$ in equation (1). A key factor of the reduced basis methods is the small Kolmogorov n-width [34]. The n-width measures to what extent the manifold $\mathcal{M}^{bk}$, the set of solutions to problem (1), can be approximated by an n-dimensional subspace of $\mathcal{X}$ [35]. If the manifold $\mathcal{M}^{bk}$ can be sufficiently approximated by a low-dimensional space, we can identify parameter values $S_N = (\mathbf{p}_1, \ldots, \mathbf{p}_N) \in \mathcal{D}^{bk}$ such that the particular solutions $\left(c^{bk}(\mathbf{p}_1), \ldots, c^{bk}(\mathbf{p}_N)\right)$ will generate a RB approximation space. We find our state approximations in this low-dimensional space, essentially replacing a large-dimensional finite element space of dimension $\mathcal{N}_h$, with a RB space generated by $N << \mathcal{N}_h$ particular solutions to $\mathcal{P}^{bk}$. Thus for any parameter value $\mathbf{p} \in \mathcal{D}^{bk}$, the solution can be approximated by a linear combination of these particular solutions:

$$c_N^{bk}(\mathbf{p}) \simeq \sum_{i=1}^{N} \alpha_i(\mathbf{p}) c^{bk}(\mathbf{p}_i). \tag{3}$$

The parameters generating reduced basis spaces can be chosen by multiple methods, and we chose to focus on Greedy algorithms. We present a weak-Greedy algorithm (Algorithm 1 in appendix) employed in the construction of reduced basis spaces from the best-knowledge model $\mathcal{P}^{bk}$ over the bk parameter space $\mathcal{D}^{bk}$. We refer to [36] for a justification of this construction where quasi optimality of the procedure is proven.
This RB approximation space will be henceforth referred to as the *Background* space $\mathcal{Z}^N$, representing solutions to the best-knowledge model $\mathcal{P}^{bk}$ in the PBDW method, and we will construct our Background spaces as a sequence of nested RB spaces

$$\mathcal{Z}^1 \subset \cdots \subset \mathcal{Z}^N \subset \cdots \subset \mathcal{X}.$$

In order to achieve stable implementation of RBMs, it is common practice to improve the basis of the RB space by a Gram-Schmidt orthonormalization method. We introduce new orthonormal basis functions $\{\zeta_i\}_{i=1}^N$ and denote our background RB space as

$$\mathcal{Z}^N = \text{span}\{\zeta_i\}_{i=1}^N = \text{span}\{c^{bk}(\mathbf{p}_i)\}_{i=1}^N \subset \mathcal{X}. \tag{4}$$

To minimize the appoximation error associated to discretization error (on the reduced N-dimensional space), we need to construct a suitably precise RB space $\mathcal{Z}^N$ such that, for a tolerance ϵ_Z,

$$\forall\, \mathbf{p} \in \mathcal{D}^{bk}, \qquad \inf_{w \in \mathcal{Z}^N} \|c^{bk}(\mathbf{p}) - w\|_{\mathcal{X}} \leq \epsilon_Z. \tag{5}$$

This RB space representing the solution manifold to $\mathcal{P}^{bk}$ described by equation (1) could be used in the implementation of RBMs in the framework of an inverse problem. Here we wish to take advantage of the simple and non-intrusive character of the PBDW method as an alternative to this integration of MOR into a classical inverse technique.

3. PBDW Formulation

The goal of the Parameterized-Background Data-Weak formulation (PBDW) is to estimate the true state $c^{true}(\mathbf{p}) \in \mathcal{X}$ (or desired output quantity $\ell^{out}(c^{true}(\mathbf{p})) \in \mathbb{R}$, where we assume ℓ^{out} linear and continuous, for example the average value over a domain of interest.) using the best-knowledge model $\mathcal{P}^{bk}$ and M observations associated to the parameter configuration $\mathbf{p}$.
The RB Background space is built from $\mathcal{P}^{bk}$, as in section 2.2. Information on the sensors is then used to build an *Update* space of low dimension representing the information gathered by the sensors.

A recent PhD thesis [37] gives detailed analysis of PBDW error and stability, as well as discussion of treatment in the case of noisy data. The case of noisy data, which was first studied in the PBDW formulation in [24], is treated with a probabilistic distribution, for example independent normal distributions, with an added regularization term over the observations (similarly to the 3D-var formulation), dependent on the variance of the distribution, in the minimization statement. In this study we will not treat the case of noisy data, as a proposed extension for this case has been well documented in [37]. In addition, we could consider that concentration sensors are not just noisy: relative errors may be large, but are small on a log scale, which is more pertinent to concentration measurements involved in dispersion or air quality modeling.

3.1. Data-informed Update

We assume that we have M sensors, which we will mathematically represent as follows (for example):

$$\varphi_m = \exp\left(\frac{-(x - x_m)^2}{2r^2}\right) \qquad \text{such that } \int_\Omega \varphi_m(x)\, d\Omega = 1, \qquad 1 \leq m \leq M \tag{6}$$

where $x_m \in \mathbb{R}^d$ is the center of the m^{th} sensor, of radius r. The underlying idea of such sensor modeling is that a sensor, especially a gas sensor (as well as PM sensors), is a complex system with spatial extension. Such a sensor does not sense pointwise, but rather performs some averaging around the sensor location. To evaluate the information these sensors can gather from a physical state $v \in \mathcal{X}$, we define the following linear functionals

$\ell_m \in \mathcal{X}'$

$$\ell_m(v) = \int_\Omega \varphi_m(x)\, v(x) d\Omega \qquad 1 \leq m \leq M. \tag{7}$$

We want to use these sensors to construct an additional approximation space $\mathcal{U}^M \subset \mathcal{X}$ of low dimension, the *Update* space. We consider that $\mathcal{U}^M$ represents the information which the sensors can provide, and its basis functions, denoted q_m, $1 \leq m \leq M$, represent the functionals ℓ_m. Let us thus define the Riesz operator $\mathcal{R}_\mathcal{X} : \mathcal{X}' \to \mathcal{X}$ such that

$$(v, \mathcal{R}_\mathcal{X} \ell)_\mathcal{X} = \ell(v) \;\; \forall v \in \mathcal{X}. \tag{8}$$

We then introduce the Update basis functions $q_m = \mathcal{R}_\mathcal{X} \ell_m \in \mathcal{X}$ such that

$$(v, q_m)_\mathcal{X} = \ell_m(v) \qquad \forall v \in \mathcal{X}. \tag{9}$$

The construction of this space takes place *offline*, as it can be relatively computationally expensive, although often less than the construction of the background space.

3.2. PBDW problem statement

The PBDW aims at approximating the true physical state $c^{true}(\mathbf{p})$ for some configuration $\mathbf{p}$ by

$$c_{N,M} = z_N + \eta_M. \tag{10}$$

where the first right-hand-side term z_N is in $\mathcal{Z}^N$ and corresponds to some RB approximation of the best-knowledge solution $c^{bk}(\mathbf{p})$, and the second right hand side term η_M is in $\mathcal{U}^M$ and is a correction term associated with the M observations. We pose the PBDW approximation as the solution to the following minimization problem. Find $(c_{N,M} \in \mathcal{X}, z_N \in \mathcal{Z}^N, \eta_M \in \mathcal{U}^M)$ such that

$$(c_{N,M}, z_N, \eta_M)_\mathcal{X} = \underset{\substack{\tilde{c}_{N,M} \in \mathcal{X} \\ \tilde{z}_N \in \mathcal{Z}^N \\ \tilde{\eta}_M \in \mathcal{U}^M}}{\operatorname{arginf}} \left\{ \|\tilde{\eta}_M\|_\mathcal{X}^2 \;\middle|\; \begin{array}{l} \tilde{c}_{N,M} = \tilde{z}_N + \tilde{\eta}_M \\ (\tilde{c}_{N,M}, \phi)_\mathcal{X} = (c^{true}, \phi)_\mathcal{X}, \forall \phi \in \mathcal{U}^M \end{array} \right\}. \tag{11}$$

The minimization over the Update term $\eta_M \in \mathcal{U}^M$ (proven to be equivalent to minimizing over $\eta_M \in \mathcal{X}$ in [23]) translates to requiring the PBDW approximation to remain close to the manifold $\mathcal{M}^{bk}$ represented by $\mathcal{Z}^N$, ensuring that the approximation maintains a physical sense with respect to the physics of the model $\mathcal{P}^{bk}$. The constraints on the minimization impose the two-part Background-Update PBDW solution, and the measured values at sensor locations. This minimization problem can be expressed by a Lagrangian and the derivation of Euler-Lagrange equations. Simplifying the Euler-Lagrange equations, the PBDW estimation statement can be written, for a given parameter configuration $\mathbf{p} \in \mathcal{D}$, as the following saddle problem [23, 24]. Find $(\eta_M \in \mathcal{U}^M, z_N \in \mathcal{Z}^N)$ such that:

$$\begin{cases} (\eta_M, q)_\mathcal{X} + (z_N, q)_\mathcal{X} = (c^{true}(\mathbf{p}), q)_\mathcal{X} \qquad \forall q \in \mathcal{U}^M, \\ (\eta_M, p)_\mathcal{X} = 0 \qquad \forall p \in \mathcal{Z}^N. \end{cases} \tag{12}$$

We recall here that given the definition of the Update basis functions $q_m \in \mathcal{X}$ in equation (9), the right-hand-side of this formulation is assumed to be $(c^{true}(\mathbf{p}), q_m)_{\mathcal{X}} = y_m^{obs}(\mathbf{p})$, with $y_m^{obs}(\mathbf{p}) = \ell_m(c^{true}(\mathbf{p}))_{\mathcal{X}}$, $1 \leq m \leq M$.
The corresponding algebraic formulation to problem (12) is : find $(\vec{\eta_M} \in \mathbb{R}^M, \vec{z_N} \in \mathbb{R}^N)$ such that

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \vec{\eta}_M \\ \vec{z}_N \end{pmatrix} = \begin{pmatrix} \vec{y}^{obs} \\ \mathbf{0} \end{pmatrix} \tag{13}$$

where $(\vec{y}^{obs})_m = y_m^{obs}$, $\mathbf{A}_{m,m'} = (q_m, q_{m'})$ and $\mathbf{B}_{m,n} = (\zeta_n, q_m)$ for $1 \leq m, m' \leq M$ and $1 \leq n \leq N$. The PBDW approximation can then be rewritten as

$$c_{N,M} = \sum_{m=1}^{M} (\vec{\eta_M})_m q_m + \sum_{n=1}^{N} (\vec{z_N})_n \, \zeta_n.$$

RBMs are often considered particularly well-suited to problems in which the quantity of interest is not the full reconstruction of the solution, but the evaluation of an output functional over the solution, allowing for complete independence from the calculation mesh in the online stage. The desired output functional can be evaluated without reconstructing the full solution:

$$\ell^{out}(c_{N,M}) = \sum_{m=1}^{M} (\vec{\eta_M})_m \ell^{out}(q_m) + \sum_{n=1}^{N} (\vec{z_N})_n \, \ell^{out}(\zeta_n).$$

This saddle problem (12) is not a function of the original PDE, making the method non-intrusive. Once the background RB space has been constructed from particular solutions to the $\mathcal{P}^{bk}$ model, the procedure is independent of the $\mathcal{P}^{bk}$ computational code provided the mesh information is available.

The key to most model reduction methods is a decomposition of the computational effort into *offline* and *online* stages. The majority of the workload is computed only once in advance, *offline*, while only parameter-dependent computations are completed during the *online* stage, which is much more efficient. The construction of the background space $\mathcal{Z}^N$, Update space $\mathcal{U}^M$, as well as the matrices A and B, also takes place during the *offline* stage — as computation time of these procedures depends on the mesh with $\mathcal{N}_h$ degrees of freedom — allowing for an efficient *online* phase. Thus, when observation data is collected, the linear system can generally be solved online in at most $\mathcal{O}((N+M)^3)$ operations. The output quantity over the basis functions of the two approximation spaces can be precalculated, allowing for evaluation of the output of the PBDW approximation in $\mathcal{O}(N+M)$ operations, without fully reconstructing the PBDW approximation from the basis functions $\{\zeta_n\}_{n=1}^N$ and $\{q_m\}_{m=1}^M$, a procedure in $\mathcal{O}(\mathcal{N}_h)$ operations. However depending on the visualization method, reconstruction of full solutions can be very efficient, making RBMs equally suitable for the general case.

3.3. PBDW error and stability considerations

The well-posedness of the PBDW problem depends on the construction of the Background and Update spaces. In fact we can define the inf-sup stability constant depending on the two approximation spaces.

$$\beta_{N,M} = \inf_{w \in \mathcal{Z}^N} \sup_{v \in \mathcal{U}^M} \frac{< w, v >_{\mathcal{X}}}{\|w\|_{\mathcal{X}} \|v\|_{\mathcal{X}}}. \tag{14}$$

$\beta_{N,M}$ is a non-increasing function of N and a non-decreasing function of M, with $\beta_{N,M} = 0$ for $N > M$.

In [23] an *a priori* error estimation is derived for the formulation as a function of the stability constant and the best-fit of the approximation spaces.

$$\|c^{true} - c_{N,M}\|_{\mathcal{X}} \leq \left(1 + \frac{1}{\beta_{N,M}}\right) \inf_{q \in \mathcal{U}^M} \inf_{z \in \mathcal{Z}^N} \|c^{true} - z - q\|_{\mathcal{X}}. \tag{15}$$

Given the strong dependence of the PBDW approximation error on the stability constant, we need to build the approximation spaces in a manner to maximize the stability of the formulation.

If we have the option of choosing the M best measurements, we want to:

(a) Maximize the stability constant $\beta_{N,M}$ for each M with respect to the Background Space $\mathcal{Z}^N$

(b) Minimize the best-fit error in the secondary approximation by the Update space $\mathcal{U}^M$:

$$\inf_{q \in \mathcal{U}^M \cap \mathcal{Z}^{N\perp}} \|\Pi_{\mathcal{Z}^{N\perp}} c^{true} - q\|_{\mathcal{X}} \tag{16}$$

If we consider that the $\mathcal{P}^{bk}$ model provides most of the information about the solution, the primary approximation will be taken from the Background space $\mathcal{Z}^N$, as imposed by equation (11). The Update term η will be taken from outside the Background space, as stated in equation (12). The best-fit error in the Update space is thus given by the projection of the portion of the true state not approximated by the Background space onto the Update space orthogonal to the Background space.

This can be attempted through optimal construction of the Update space employing a Greedy-type selection of sensor functions (among a set of possible locations) to improve the space with respect to **(a)** or **(b)**. The former can be done for example using an algorithm to maximize $\beta_{N,M}$ under a certain tolerance, reverting otherwise to minimization of the best-fit error, as in [37]. The latter can be done using for example a double-greedy procedure in order to minimize the GEIM [25, 26] interpolation error, which selects Background RB basis functions and Update sensor basis functions simultaneously. The sensor placement optimization found in [15] is based on the sensitivity of the flow to varying input parameters, designed specifically for parameter identification. To the contrary, here the sensor placement optimization is designed to better represent the concentration field based on our best knowledge of the states, and combined with the choice of Riesz representation, maximizes the ability of the Update space to correct error in the state estimates.

4. Numerical Implementation of the PBDW method

In this section we will discuss problem-specific details of the implementation of the PBDW method.

The goal of this application is to test the feasibility of the PBDW method to represent the complex physical phenomena involved in air quality modeling. In fact RBMs are notoriously ill-suited to problems of transport

by convection or to problems with too many varying parameters. For this reason we focus this study on the dispersion of pollutants in an outdoor urban setting, and with the addition of a reaction term represent the most predominant terms in physically-based AQMs. We aim to demonstrate that the modeling of dispersion-reaction by an imperfect model is feasible with the PBDW method, which will suggest that an extension to operational CFD-based AQMs may be possible thants to the strategic treatment of the velocity field as a parameter in the bk problem and the non-intrusive data assimilation allowing to correct for unmodeled physics.

In real-world air quality applications, sensors are often limited in number; we want to respect this constraint in the methodology considered here, and consider a relatively small number of sensors over the domain (we'll consider up to 20) testing various sensor locations. We will consider PBDW results in the (academic) case of a perfect $\mathcal{P}^{bk}$ model, and in the case of unmodeled physics such as a reaction term or a true solution calculated with a different computational model.

4.1. Background RB space

The construction of a RB Background space $\mathcal{Z}^N$ for our 2D case study was done using the weak Greedy algorithm 1 on a training set of particular solutions for varying parameters of wind velocity $\mathbf{p}_v$ and source intensity $\mathbf{p}_s$ in the parameter set $\mathcal{D}^{bk} = \{(\mathbf{p}_v, \mathbf{p}_s) \in [0.1; 1.3\frac{m}{s}] \times [1 \times 10^{-3}; 1 \times 10^{-2}\frac{mg}{m^3}]\}$.
A sign of a good reduced basis is the estimation of a small Kolmogorov n-width by rapid decay of projection errors of these training solutions onto the N-dimensional RB space. In figure 2 we see the mean and maximal relative projection errors in H^1 norm as a function of N

$$Err_{mean}^{Greedy} = \frac{1}{Nb_{trial}} \sum_{i=1}^{Nb_{trial}} \frac{\|c^{bk}(p_i) - \Pi_{\mathcal{Z}^N} c^{bk}(p_i)\|_{H^1}}{\|c^{bk}(p_i)\|_{H^1}}, \tag{17}$$

as well as mean relative projection errors over the calculation domain, corresponding to a pointwise mean on the calculation mesh over the following error formula.

$$Err_{\Omega}^{Greedy}(p_i) = \frac{|c^{bk}(p_i) - \Pi_{\mathcal{Z}^N} c^{bk}(p_i)|}{\|c^{bk}(p_i)\|_{L^\infty}} \in \mathcal{X} \tag{18}$$

This serves as a representation of the approximation quality of the reduced basis space $\mathcal{Z}^N$ for the solution space $\mathcal{M}^{bk}$.

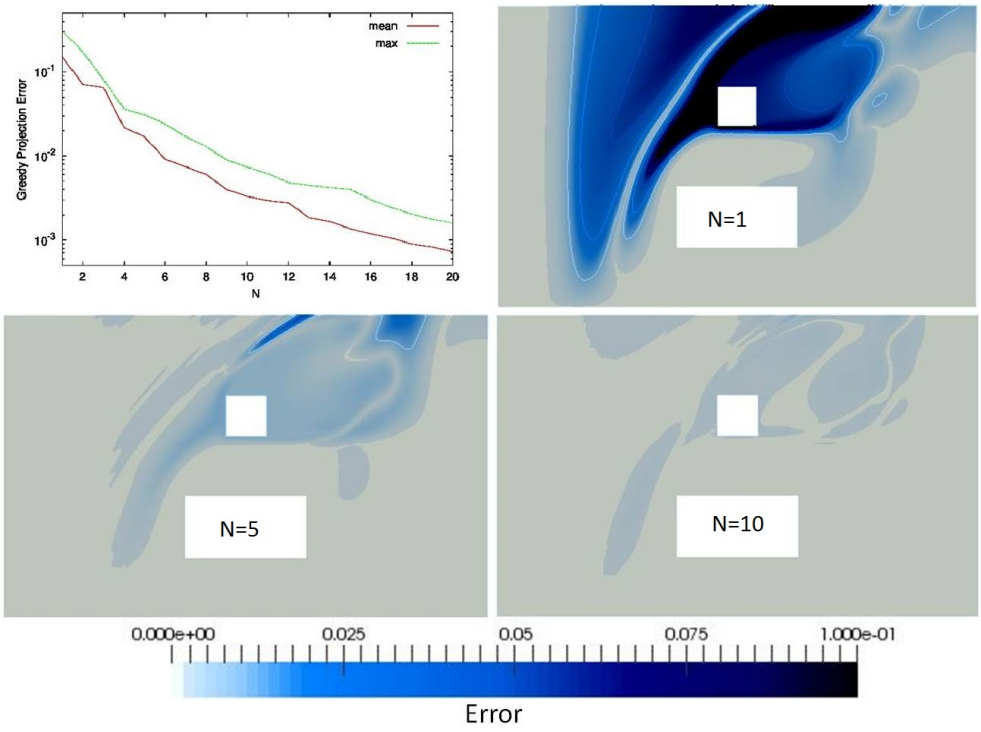

Figure 2: Relative mean and maximal projection errors in H^1 norm of the training solutions during the greedy construction of the RB space, as a function of N following equation (17) (top left) ; Relative mean projection error of the training solutions over the Greedy RB space, pointwise over domain Ω from equation (18) for RB dimensions $N = 1$ (top right), $N = 5$ (bottom left), and $N = 10$ (bottom right). The lowest contour curve represents 1% error.

We can see that the discretization error of the RB Background space rapidly converges to under 1%. Given the complexity of reducing convection-dominated problems and the uncertainty involved in real-world modeling of dispersion, we consider this wholly satisfactory. An additional 1% error (with respect to the $\mathcal{P}^{bk}$ model) from the dimensional reduction of the approximation space from a finite element space to a RB space would thus be considered negligible. We will note from the RB discretization error maps over the domain that for RB dimension $N = 10$, we have nearly eliminated the error, excepting small but unavoidable "shocks" from varying convection fields. We can thus hope to fix our online basis size at $N \sim 5$, which we will consider further in section 5.1.

4.2. Sensor locations and Update Space

We will compare two cases of sensor locations in this case study: the case of sensor locations chosen randomly, and the case of sensor locations chosen by a weak Greedy method as in the GEIM.

The GEIM simultaneously defines the set of so-called generating functions (e.g. the Background basis functions) $\xi_i \in \mathcal{M}^{bk}$ and the associated linear forms (i.e. the sensor functions). The first chosen generating function

ξ_1 is the "largest" bk solution by $\mathcal{X}$-norm, and the associated sensor function ℓ_1 (chosen among the set of available sensor locations Σ) is the sensor which gives the most "information" on $c^{bk}(\mathbf{p}_1)$. We then define the interpolation operator

$$\mathcal{I}_M(c^{bk}) = \sum_{j=1}^{M} \beta_j \xi_j \text{ such that } \ell_i\big(\mathcal{I}_M(c^{bk})\big) = \ell(c^{bk}) \; \forall 1 \leq i \leq M \tag{19}$$

Ideally we want to choose the linear forms ℓ_i and basis functions $\xi_i \in \mathcal{M}^{bk}$ in an optimal manner. We can consider a Greedy algorithm similar to algorithm 1, selecting each new generating function to maximize the interpolation error. We defined a double-Greedy algorithm based on this interpolation error in order to select sensor placements specifically adapted to RBM-based data assimilation.

In figure 3 we can see a set of sensor locations chosen randomly, as well as the set Σ of possible sensor locations chosen for this application and those selected by the GEIM-based double-Greedy algorithm.

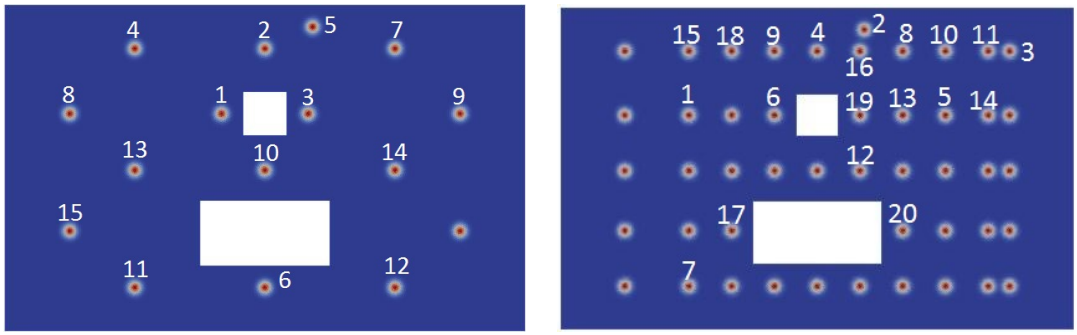

Figure 3: Sensors locations chosen randomly (left) and chosen by a Greedy algorithm (right).

In figure 4 we see the values of the stability constant $\beta_{N,M}$ from equation (14), with $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_{H^1}$, for various N-values as a function of M, for each sensor set. This figure represents the stability of the PBDW system induced by choice of sensor locations.

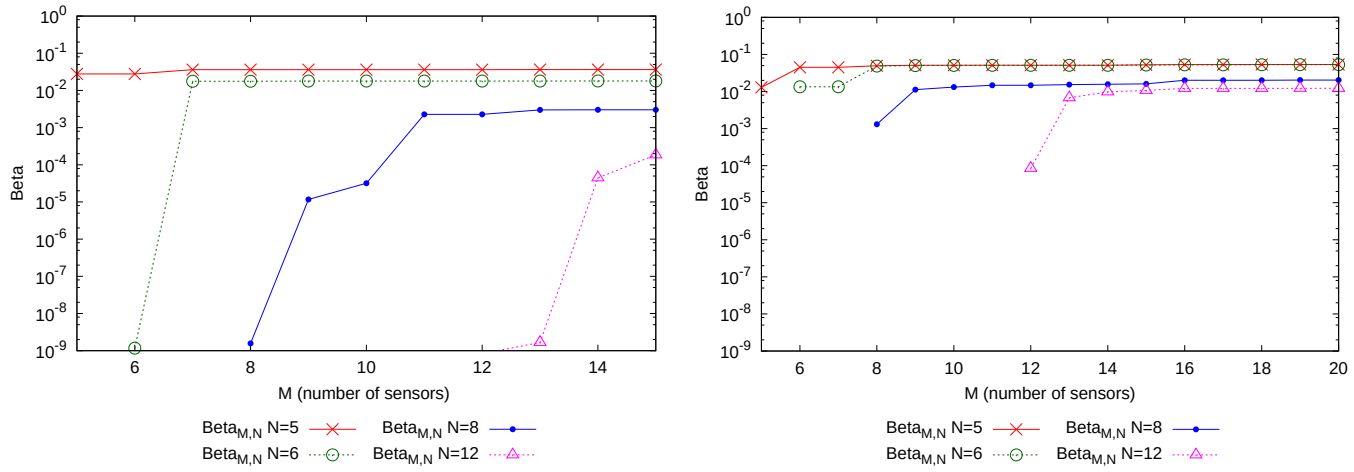

Figure 4: PBDW inf-sup stability constant $\beta_{N,M}$ in H^1 norm, equation (14), for the associated PBDW linear systems (13) as a function of the number of data points M, for various Background RB dimensions N. Sensors chosen randomly (left) and chosen by a Greedy algorithm (right).

The PBDW systems were constructed from equation (13) using the RB Background space discussed in section 4.1 and an Update spaces built from these respective sensor locations (placed randomly or by the Greedy algorithm). As $\beta_{N,M}$ is a non-decreasing function of M, we see improvement in the stability constants for larger numbers of data points, for each fixed Background RB dimension N. We note that in general for $N \simeq M$ the formulation is less stable, as evidenced by very low values of $\beta_{N,M}$ and discussed in [26]. Given this knowledge, we make the choice to disregard PBDW results for $N \simeq M$ (as we will see in section 5.1).

If we compare the stability constants for randomly chosen sensor locations to those for sensor locations chosen via Greedy, we can see that in our case study we've improved by multiple orders for some M and N values, and at least by a factor of 2 for smaller Background dimensions.

Given the relatively small size of standard concentration sensors which provide observational data in real-world applications, with respect to the large domain of study, we aim to respect this constraint in the development of the methodology and testing on pollution dispersion studies. In order to extend the PBDW under this constraint while maintaining a mathematically sound definition and realistically smooth output concentration field, wechose to modify the norm used in the definition of the Update basis functions by Riesz representation in equation (9). We introduce the following $\tilde{H}^1$ scalar product for $u, v \in H^1$.

$$\langle u, v \rangle_{\tilde{H}^1(\Omega)} = \langle u, v \rangle_{L^2} + L_g^2 \langle \nabla u, \nabla v \rangle_{L^2}, \tag{20}$$

where $L_g = 75$ is a characteristic length of the domain. This scalar product serves to enlarge the support of the Update basis functions and to smooth the Update contribution, in order to provide improved approximation properties to the Update approximation space (see (16)). The induced $\tilde{H}^1$ norm is used in the variational formulation (12) for equivalence.

5. State Estimation Results

In this section we will present the numerical results of the PBDW method on the 2D case study presented in section 2. We will present the PBDW state estimation results over the full domain and over a domain of interest, considering the variations in sensor choice discussed in paragraph 4.2. Above we presented analysis of stability of the system, and in this section we will present the state estimation results of the associated PBDW systems, along with error bounds for parametric variation only (the case of a perfect $\mathcal{P}^{bk}$ model), and for little to significant model error. We will also compare the results of the PBDW method to those obtained by the GEIM, both non-intrusive reduced order data assimilation methods, in precision and computational time.

For purposes of analyzing results and numerically calculating the error bound in equation (15), we will consider the following relative *best-fit* error onto what we will refer to as the PBDW approximation space $\mathcal{Z}^N \oplus (\mathcal{U}^M \cap \mathcal{Z}^{N\perp})$:

$$\frac{\|c^{true} - \Pi_{\mathcal{Z}^N \oplus (\mathcal{U}^M \cap \mathcal{Z}^{N\perp})} c^{true}\|_X}{\|c^{true}\|_X}. \tag{21}$$

5.1. PBDW applied to a case study in exterior dispersion modeling

The two-dimensional case study on the domain represented in figure 1 was considered for varying parameters in $\mathcal{D}^{bk}$ introduced in section 4. In figure 5 we can see concentration fields for lowest and highest wind velocity and emission rates.

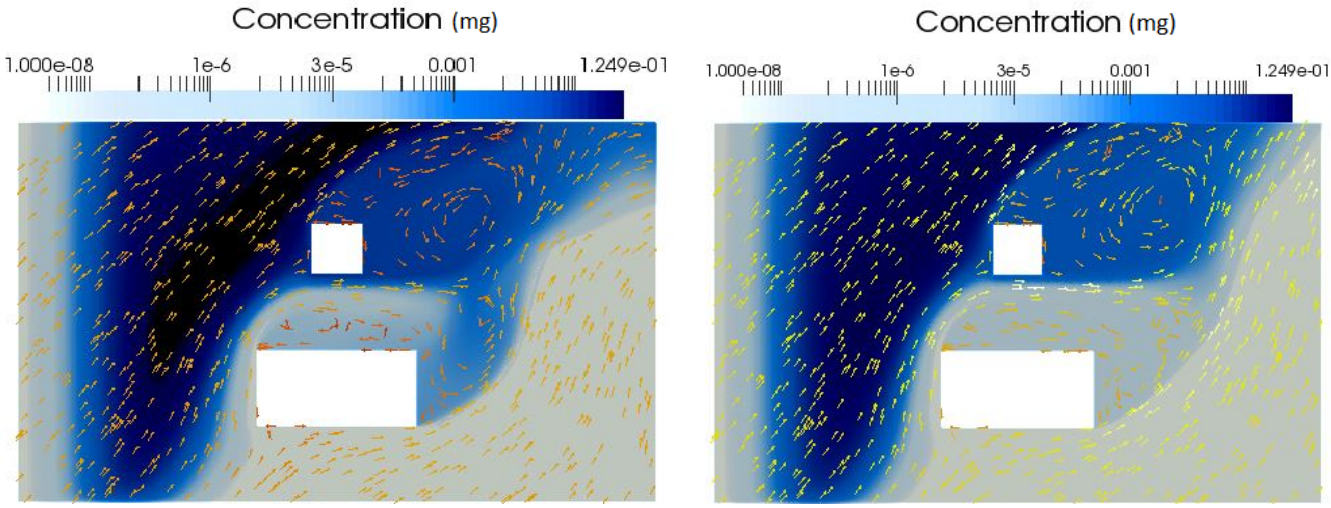

Figure 5: Concentration fields (logarithmic scale) from the $\mathcal{P}^{bk}$ model (1) over velocity fields and different pollutant source intensities. $(\mathbf{p}_v, \mathbf{p}_s) = (0.1\frac{m}{s}, 1 \times 10^{-3}\frac{mg}{m^3})$ (left), and $(\mathbf{p}_v, \mathbf{p}_s) = (1.3\frac{m}{s}, 1 \times 10^{-2}\frac{mg}{m^3})$ (right).

In the following we will consider three sets of 6 trial solutions to test the method. Each of the trials corresponds to velocity parameters $\mathbf{p}_v$, and to varying intensity of the pollutant sources $\mathbf{p}_s$. The values of the trial parameters lie within $\mathcal{D}^{bk}$ but are different from the values used in the training set for the RB space: $\mathcal{D}^{trial} = \{(\mathbf{p}_v, \mathbf{p}_s) \in \{0.15, 0.6, 1.28\}\frac{m}{s} \times \{3 \times 10^{-3}, 7 \times 10^{-3}\}\frac{mg}{m^3}\} \subset \mathcal{D}^{bk} \setminus \mathcal{D}^{training}$. One set consists of solutions to equation (1) representing the (unrealistic) case of a perfect $\mathcal{P}^{bk}$ model, with the goal of demonstrating the error inherent to the MOR approach of the PBDW method. The remaining trial sets consist of solutions to an advection-diffusion-reaction problem:

$$\rho\vec{v} \cdot \nabla c - div((\epsilon_{mol} + \epsilon_{turb})\nabla c) + \rho R c = \rho F_{src}, \tag{22}$$

with linear reaction terms of coefficients $R = 0.001$ and $R = 0.0001$. These sets are used to demonstrate how the method handles two levels of model error, with an average error over 8% (and up to 17%) and 1%, respectively.

In figure 6 we compare the FEM solution to PBDW state estimates for trial solutions with significant model error: we can see the trial solution corresponding to maximal error, $c^{trial}(\mathbf{p}_{max})$, with

$$\mathbf{p}_{max} = \underset{\mathbf{p} \in \mathcal{D}^{trial}}{\operatorname{argmax}} \frac{\|c^{trial}(\mathbf{p}) - c_{N,M}(\mathbf{p})\|_{H^1}}{\|c^{trial}(\mathbf{p})\|_{H^1}} \tag{23}$$

compared with the PBDW approximations from randomly-chosen sensor locations and Greedy sensors.

We see reasonable reconstruction of the physical state with both sensor sets. While the Greedy sensors add a very small phantom concentration in some regions, this error is negligible. The Greedy system has more accurately reconstructed the concentration peak near the source, however both PBDW approximations underestimate the peak. The under-representation of the concentration remains relatively small.

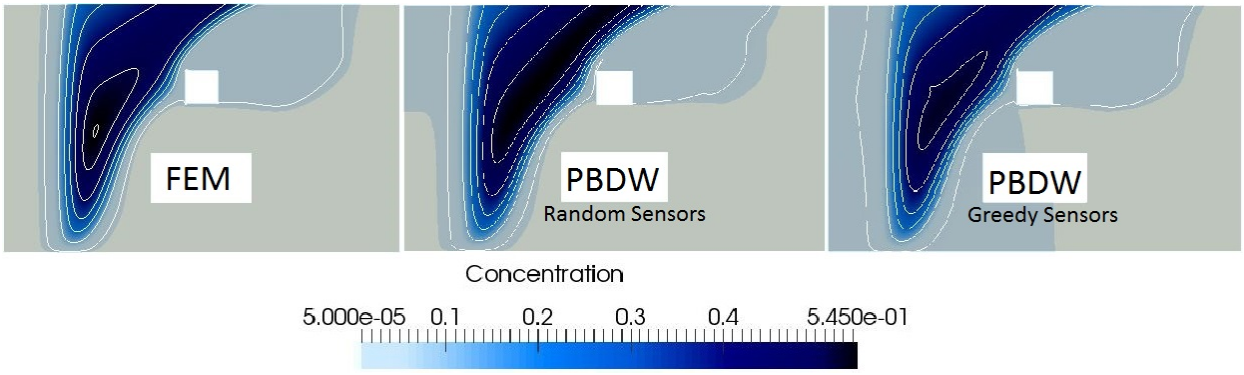

Figure 6: Approximation of the concentration for $\mathbf{p} = \mathbf{p}_{max}$. Trial solution with model error simulated by a reaction term of $R = 0.001$. FEM solution c^{true} (left), PBDW approximation using synthetic data, with random sensors (middle), PBDW approximation with greedy selected sensors (right). We set $M = 13$ and $N = 6$ here.

In figure 7 we can see relative mean best-fit errors from equation (21), measure in the H^1 norm, over our set of trial solutions with significant model error. We notice that in the case of a perfect model, for each N-value the relative best-fit error is nearly constant with respect to M. This implies that our Update basis functions

q_m do not provide new information outside the span of the background approximation space $\mathcal{Z}^N$. This effect is to be expected, as the trial solutions were computed with the same model as the reduced basis, which is meant to approximate the associated solution space. However, we see improvement of the best-fit error in the case of an imperfect model. The added Update basis functions enlarge the span of the PBDW approximation space $\mathcal{Z}^N \oplus (\mathcal{U}^M \cap \mathcal{Z}^{N\perp})$ to capture information on the trial solutions from the shifted model not spanned by the background space. We also note that additional background basis functions do not greatly improve the approximation, as the trial solutions do not lie on the same solution manifold.

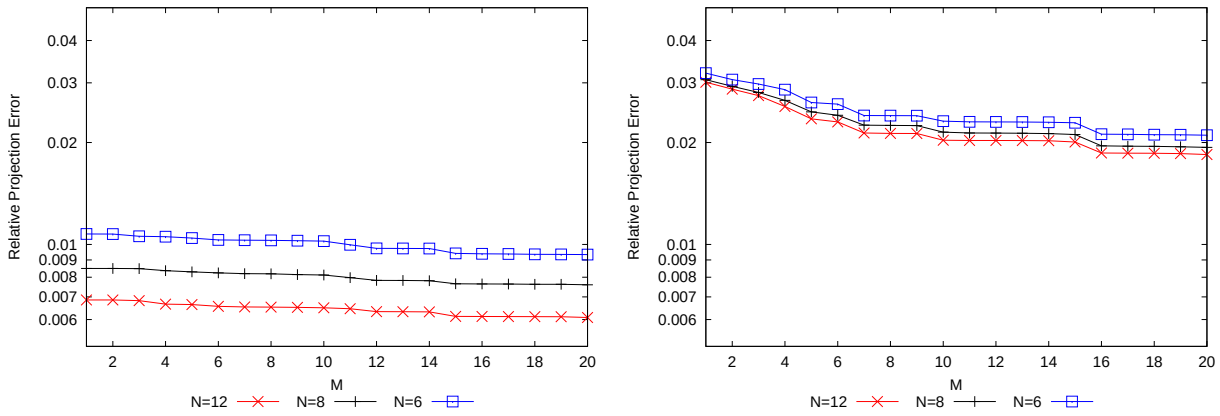

Figure 7: Relative mean best-fit error, equation (21), for the set of trial solutions over $\mathbf{p} \in \mathcal{D}^{trial}$, as a function of M in H^1-norm. No model error (left), and model error with an added reaction term of $R = 0.001$ (right). Sensors chosen by a Greedy algorithm.

In figure 8 we see relative mean PBDW approximation errors mapped over the domain for the case of significant model error given by.

$$Err_{\Omega}^{PBDW}(p_i) = \frac{|c^{trial}(p_i) - c_{N,M}(p_i)|}{\|c^{trial}(p_i)\|_{L^\infty}} \in \mathcal{X}. \tag{24}$$

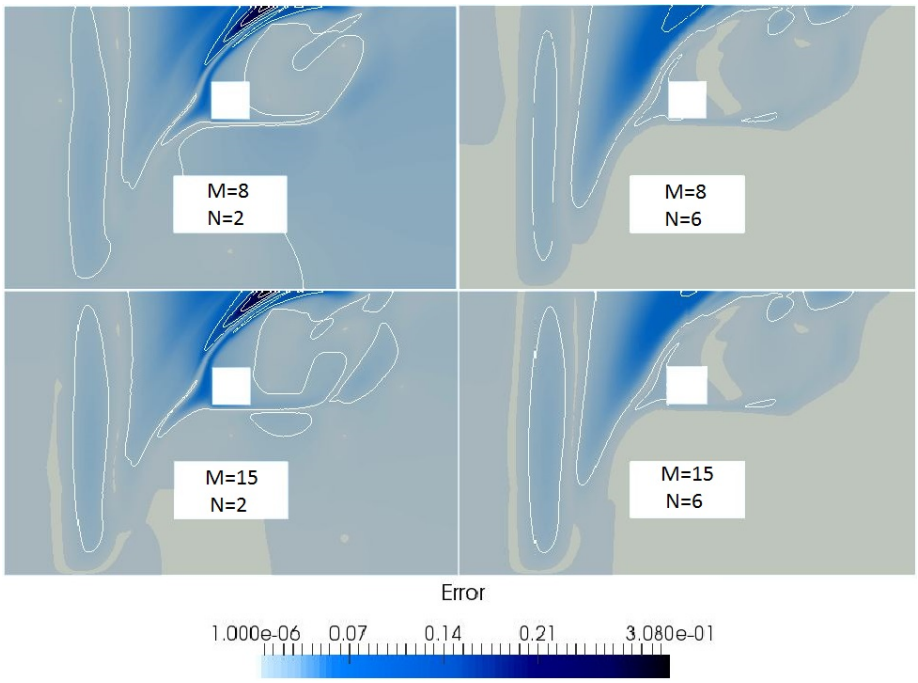

Figure 8: Relative mean pointwise PBDW approximation error maps, equation (24) over trial set $\mathbf{p} \in \mathcal{D}^{trial}$ with model error by an added reaction term of $R = 0.001$, for $N = 2$ (left), $N = 6$ (right), and for $M = 8$ (top) and $M = 15$ (bottom). Randomly-chosen sensor locations. The lowest contour line shows 1% error.

We see significant improvement between $N = 2$ and $N = 6$, but smaller improvements when adding more data points. In this simple test, $M = 8$ is sufficient data for the PBDW system to approximate the state over the $N = 6$ Background functions, and adding more Update basis functions does not greatly improve the approximation, which we attribute to sensor placement.

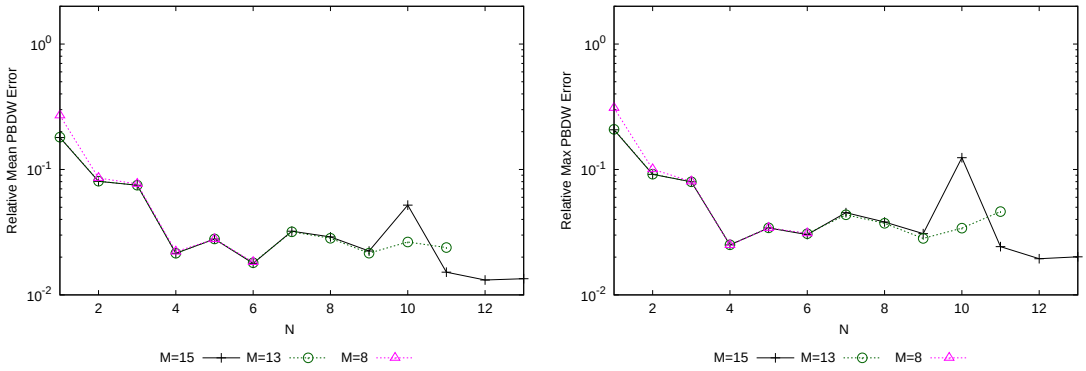

Figure 9: Relative mean (equation (25), left) and maximal (equation (26), right) PBDW approximation error in H^1-norm as a function of Background RB dimension N, for various numbers of data points M, over $\mathbf{p} \in \mathcal{D}^{trial}$ with no model error. Randomly-chosen sensor locations.

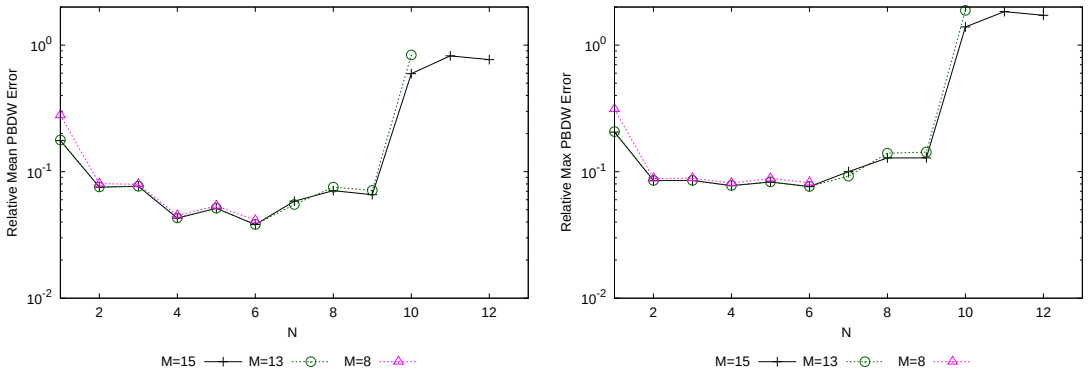

Figure 10: Relative mean (equation (25), left column) and maximal (equation (26), right column) PBDW approximation error in H^1-norm as a function of Background RB dimension N, for various numbers of data points M, over $\mathbf{p} \in \mathcal{D}^{trial}$, model error with an added reaction term of $R = 0.001$. Randomly-chosen sensor locations.

We define mean and maximal PBDW approximation errors in the H^1-norm:

$$Err_{mean}^{PBDW} = \frac{1}{Nb_{trial}} \sum_{i=1}^{Nb_{trial}} \frac{\|c^{trial}(\mathbf{p}_i) - c_{N,M}(\mathbf{p}_i)\|_{H^1}}{\|c^{trial}(\mathbf{p}_i)\|_{H^1}} \tag{25}$$

$$Err_{max}^{PBDW} = \underset{\mathbf{p} \in \mathcal{D}^{trial}}{\text{MAX}} \frac{\|c^{trial}(\mathbf{p}) - c_{N,M}(\mathbf{p})\|_{H^1}}{\|c^{trial}(\mathbf{p})\|_{H^1}} \tag{26}$$

In figures 9 and 10 we see relative mean and maximal error curves for the PBDW approximation with randomly sensor locations for two trial sets, showing of the quality of the PBDW state estimation in the H^1 norm, using randomly-chosen sensor locations. We can see that with no model error with $N = 6$ Background functions we achieve $\sim 2\%$ mean error (and $\sim 3\%$ maximal error on the worst trial solution), and $\sim 4\%$ (and under 8% maximal error on the worst trial solution) error with significant model error. In applications of air quality modeling input errors are commonly much larger, in the range of $30-70\%$ if not higher, much of which stemming from the factors represented here (transport, diffusion, reaction, and source representation). This study finds smaller errors but of similar order and the possibility of correction a portion of this error would prove advantageous. We note that the non-monotone error curves are to be expected: there is no mathematical argument for strictly decreasing error, as the error depends not only on the best-fit of the PBDW approximation space, but also on the stability and conditioning of the system. We can observe that the instability for N approaching M (seen in the stability coefficient $\beta_{M,N}$ of equation (14)) has an amplified effect on the error in the case of more significant model error. This is consistent with equation (15).

In figures 11 and 12 we see relative mean and maximal error curves for the PBDW approximation with Greedy sensor locations for each of two trial sets. We can see that with no model error with $N = 6$ Background functions we achieve $\sim 1\%$ mean error (and under 3% maximal error on the worst trial solution), and $\sim 3\%$ error (and 6% maximal error on the worst trial solution) with significant model error. We note that we see more consistent error results for varying N-values, with fewer peaks in the error, as compared to sensors chosen randomly. We

can attribute this to the increased stability and conditioning of the PBDW linear system. We also note that while we see only small improvement of the approximation error in the best case (of N- and M-values), we see global improvement with the Greedy sensors. We could thus draw the preliminary conclusion that the Greedy-placed sensors is no guarantee of improved precision in the PBDW approximation (here it depends on N- and M-values), but seems to improve the stability of the system and consistency of the results, which would be a non-negligible advantage in the online stage when precise *a posteriori* error analysis is not feasible.

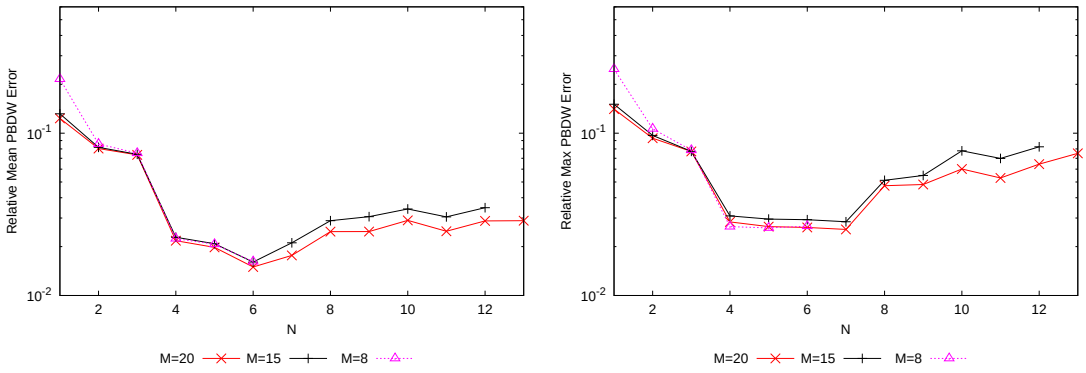

Figure 11: Relative mean (equation (25), left) and maximal (equation (26), right) PBDW approximation error in H^1-norm as a function of Background RB dimension N for various numbers of data points M, over $\mathbf{p} \in \mathcal{D}^{trial}$ with no model error. Sensor locations chosen by a greedy procedure.

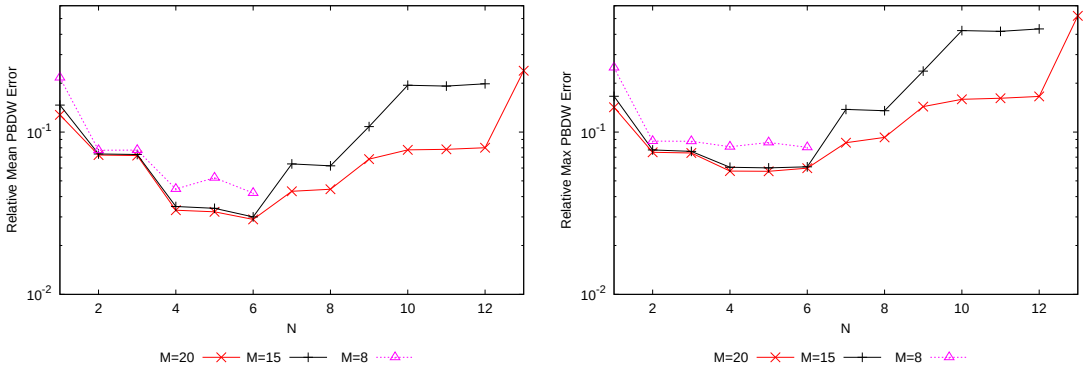

Figure 12: Relative mean (equation (25), left) and maximal (equation (26), right) PBDW approximation error in H^1-norm as a function of Background RB dimension N for various numbers of data points M, over $\mathbf{p} \in \mathcal{D}^{trial}$, model error with an added reaction term of $R = 0.001$. Sensor locations chosen by a greedy procedure.

In figure 13 for Greedy sensors we see relative mean errors mapped over the domain in the case of no model error. Here we see a bit more improvement between $M = 8$ and $M = 15$, which can be attributed to better-placed sensors. However, the background space alone can represent these trial solutions, so as expected the most improvement is provided by N.

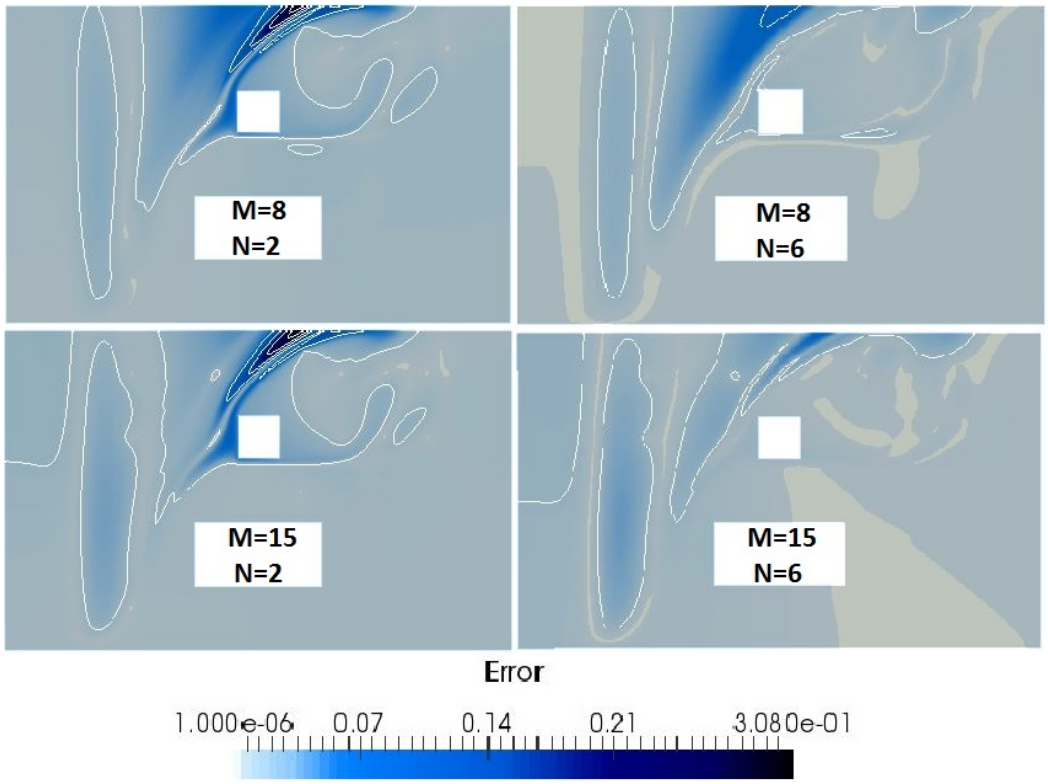

Figure 13: Relative mean pointwise PBDW approximation error maps, equation (24), for $N = 2$ (left), $N = 6$ (right), and for $M = 8$ (top) and $M = 13$ (bottom), over $\mathbf{p} \in \mathcal{D}^{trial}$ with no model error. The lowest contour line shows 1% error. Sensor locations chosen by a greedy procedure.

In figure 14 we consider Greedy sensors for the case of significant model error. Here we see more significant improvement with added data points. We again note that the correction by the Update basis functions can add non-physical error to the approximation, however this is generally of negligible order. Again we see significant improvement between $N = 2$ and $N = 6$. We see that with $N = 6$ and $M = 15$ the error is under 7% everywhere, and often under 1%. Compare to the corresponding case with randomly placed sensors, where the approaches and error surpasses 7% in a some areas.

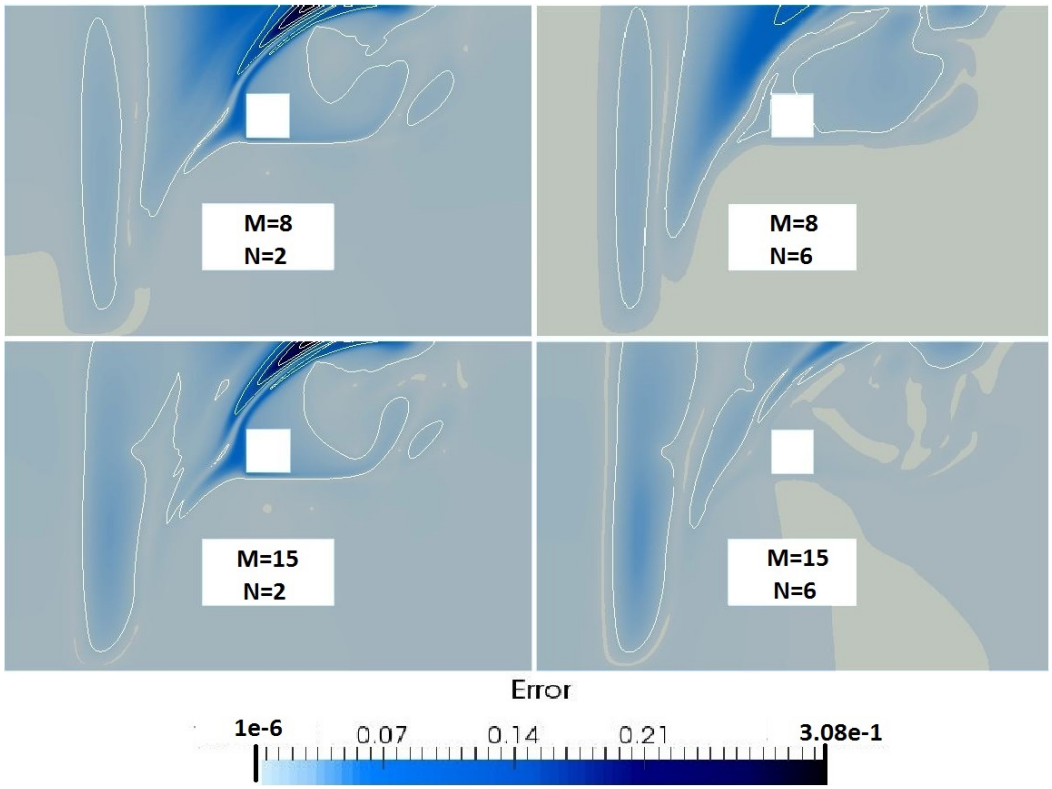

Figure 14: Relative mean pointwise PBDW approximation error maps, equation (24), for $N = 2$ (left), $N = 6$ (right), and for $M = 8$ (top) and $M = 13$ (bottom), over $\mathbf{p} \in \mathcal{D}^{trial}$ with model error by an added reaction term of $R = 0.001$. The lowest contour line shows 1% error. Sensor locations chosen by a greedy procedure.

In RBM applications it is often unnecessary to reconstruct the approximated solution over the full domain Ω; instead the solution of some output value on the solution over a smaller domain of interest $\Omega_{out} \subset \Omega$ is approximated. This is highly compatible with air quality studies, as often the physical *quantity of interest* (QoI) is a concentration peak in an area or the average concentration over a period of time in an area, such as a playground or a school area. This renders RBMs much more advantageous (no online complexity is dependent on the mesh dimension $\mathcal{N}_h$). In this case study we considered the quantity of interest to be the average concentration over a subdomain of interest, and achieved greatly reduced computational times (seen in table 2) for equivalent precision.

In figure 15 we can see relative mean PBDW approximation errors and bounds over $\mathbf{p} \in \mathcal{D}^{trial}$, comparing a set without model error and a set with model error (an added reaction term of $R = 0.0001$). Plots show best-fit error from equation (21), PBDW approximation error (i.e. the left-hand-side of equation (15)), and an a priori error bound given by (the right-hand-side of) equation (15), all in relative mean with respect to $\|c^{trial}(\mathbf{p}_i)\|_{H^1(\Omega_{out})}$ over the trial set. We choose to fix the Background basis size at $N = 6$, as would be chosen in the online implementation of this study. We notice that in this case with N chosen well after offline study of results, the improvement by Greedy-placed sensors is less important, however we attribute this to the simplified case study.

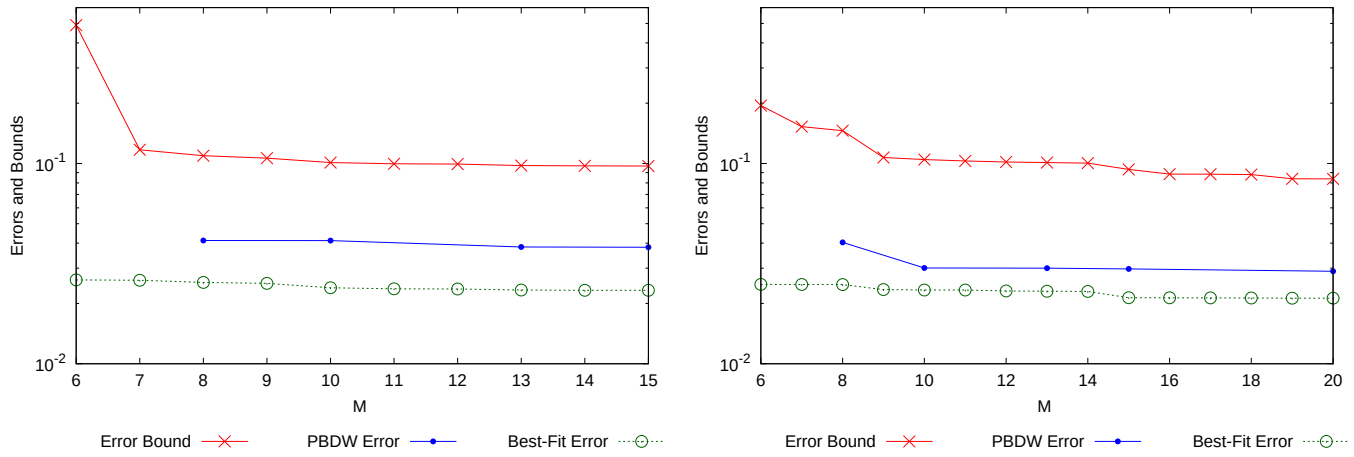

Figure 15: Relative mean PBDW results in H^1-norm as a function of number of data points M for Background basis dimension $N = 6$. Error bound from equation (15), PBDW approximation error, and best fit error from equation (21), over $\mathbf{p} \in \mathcal{D}^{trial}$ with model error of $R = 0.001$. Randomly chosen sensors (left), and sensors chosen by Greedy (right).

5.2. Comparison of non-intrusive methods: PBDW or GEIM?

In this section we want to compare the results of the PBDW state estimation on this two-dimensional case study to those optained by the GEIM interpolation method discussed in previous sections. The GEIM method is implemented with $M = N$, equal number of basis functions and data points. Below we can see the results of the two methods, both of which we implemented offline from the same set of training solutions and selection from the same sensor grid, and applied to the same set of 6 trial solutions of varying parameters and with added model error, described in section 5.1.

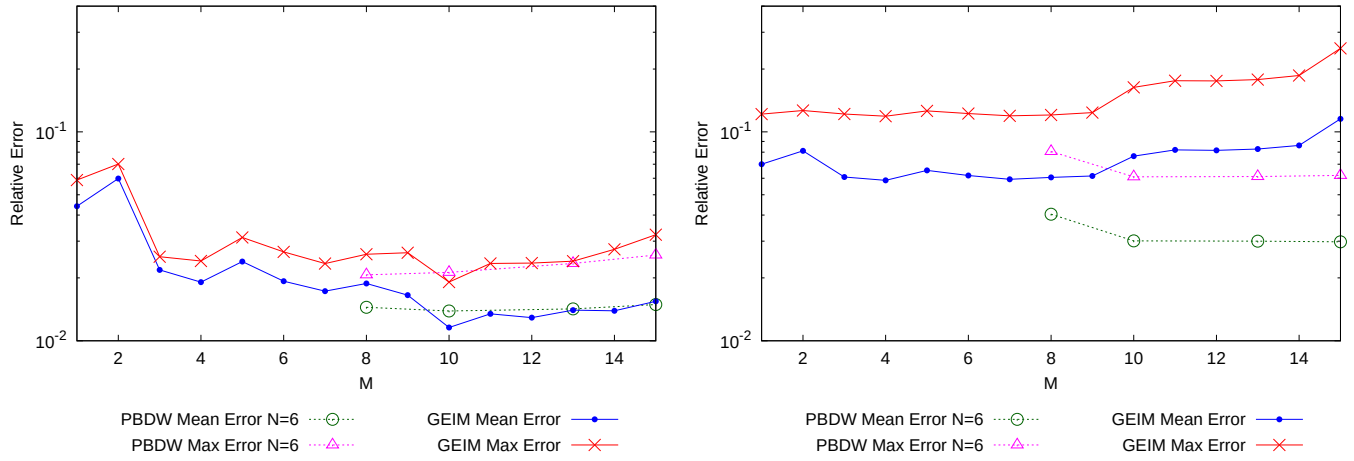

Figure 16: Relative mean and maximal PBDW H^1-errors as a function of number of data points M for PBDW Background basis dimension $N = 6$, and GEIM H^1 interpolation errors as a function of $M = N$, over $\mathbf{p} \in \mathcal{D}^{trial}$. Model error by an added reaction term $R = 0.0001$(left), and an added reaction term $R = 0.001$(right). Greedy sensor set used in both methods.

We can see that the GEIM method performs similarly, and even surpasses for $M = 10$, to the PBDW method in the case of little model error. However in the case of significant model error and $M > 10$, the PBDW method provides a significantly better estimation. In this particular case study, we seem to have more consistent error results for varying M-values, an aspect that could be valuable in online studies without feasible *a posteriori* error analysis.

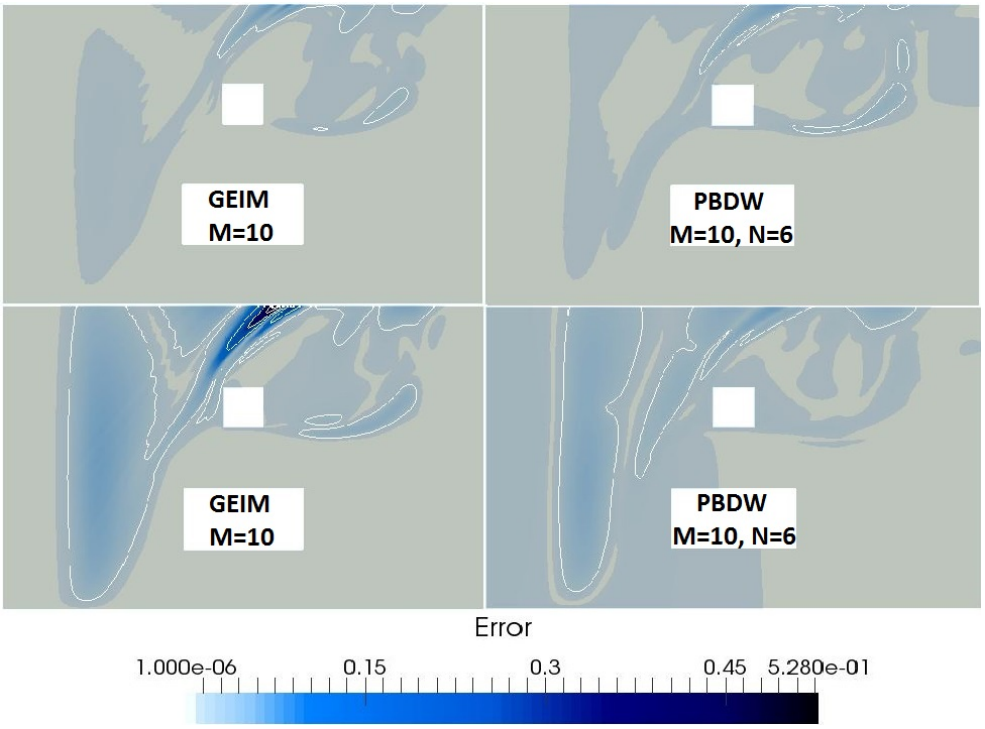

Figure 17: Relative mean pointwise GEIM (left) and PBDW (right) approximation error maps for $M = 10$ and $N = 6$-$M = 10$, respectively, over $\mathbf{p} \in \mathcal{D}^{trial}$. Model error of $R = 0.0001$ (top) and $R = 0.001$ (bottom). Mapping of the errors is truncated at 1×10^{-6}, and the lowest contour line shows 1% error.

In figure 18 we compare relative mean error maps for the GEIM and PBDW approximations over trial sets with little or significant model error. We consider the case of $M = 10$, the best case of the GEIM approximation according to figure 16. We can see similar results for little model error, with only a small region over 1% error in both approximations, while the GEIM approximation reduces a region of error with respect to the PBDW estimation. In the case of significant model error, however, we see a clear advantage in the PBDW estimation, with no peak near or above 15% and only a small misrepresentation of the source intensity.

In table 1 we see computational times for the classical FEM approximation of equation (1), with no data assimilation or model error correction.

[1]In *Code_Saturne*, in order to treat the nonlinearity of the fluid problem, the steady-state solution is compute as the limit of a transient one, leading to an iterative procedure requiring sufficient solutions to reach a stabilized velocity field.

CPU Times	Best-knowledge State Estimatation $\Omega : 125m \times 75m$
FEM-SUPG $c^{bk}(\mathbf{p})$ $\mathcal{N}_h \sim 323,000$	$7.4h^1$+$61s$ (fluid) (dispersion)

Table 1: Computational times of the standard FEM approximation of (the imperfect) equation (1), before applying any model order reduction or data assimilation techniques. Average over the set of trial solutions considered here.

In table 2 we compare computation times of the PBDW state estimation and GEIM approximation. Both of these methods rely on a training set of solutions to the best-knowledge problem, for which we set $N_{train} = 40$, requiring approximately $296.6h$ of calculations. After calculating the training set, the offline stage of the PBDW method, with $M = 10$ and $N = 6$, requires another 10.26 minutes, whereas the GEIM with $M = 10$ requires 42.7 minutes. Once the one-time offline stage has been completed, in the case of full reconstruction of the physical state the PBDW method requires a computational time of 10 times less than that needed to approximate a single direct best-knowledge dispersion solution, and even nearly 5000 times less if we recomputed a wind field. The GEIM method saves even a few more seconds, given the smaller linear system size. This is for the reconstruction of the concentration over the full domain, thus a finite element vector of dimension $\mathcal{N}_h$. We also compare computational times for the PBDW estimation and the GEIM approximation of an output quantity, considering the average pollution concentration over a $10m \times 20m$ subdomain Ω_{out}. In the case of a QoI, rendering full reconstruction of the physical state unnecessary, we see a reduction by nearly 30 times with respect to the already inexpensive full state estimate for the PBDW method. The GEIM method requires equivalent time to compute the QoI, leaving nearly negligible calculation times. These differences could be taken into consideration in the case of full reconstruction of the pollution field, along with the precision and peaks in error results when determining which MOR data assimilation method is most pertinent and advantageous to the application. However the improved model error correction provided by the PBDW method for relatively equivalent calculation times gives a clear advantage to PBDW state estimation.

CPU Times: Non-intrusive reduced order data assimilation	Online Stage (average CPU times)	
	State Estimate $c(\mathbf{p})$ $\Omega : 125m \times 75m$	Quantity of Interest $\ell_{out}(c(\mathbf{p}))$ $\Omega_{out} : 20m \times 10m$
PBDW ($M = 10$, $N = 6$)	$5.35s$	$0.18s$
GEIM ($M = 10$)	$3.32s$	$0.17s$

Table 2: Computational times of the two MOR-data assimilation methods for state estimation over the full calculation domain and estimation of a quantity of interest (average concentration over a subdomain) during the online stage. Average over the set of trial solutions considered here.

5.3. *Computational savings compared to adjoint-based methods*

We would also like to reference at this point the computational savings found over a larger three-dimensional urban domain studied in [38] (chapter 7), of approximate size $800m \times 800m \times 30m$. In order to demonstrate the advantage of the PBDW method, we will consider a sort of best-case scenario in favor of adjoint-based methods: the wind field will not need to be recalculated, and the associated parameter is a multiplicative coefficient for a provided wind field. This leaves only the adjoint of the dispersion problem to be solved at each iteration of the adjoint method, and ignores the question of intrusivity and non-linearity in the adjoint to a CFD wind field model.

If we compare these variational methods to the common statistical interpolation methods of kriging, in the latter method we'd find after precalculation an interpolation system of size M for each set of measurements, which corresponds to the complexity of the GEIM method in table 2 at minimum. However these methods still require a large training set of measurements (not to be confused with the PBDW training set of simulations), do not account for physical phenomena represented by a sophisticated model, and often require numerous data points to obtain acceptable results.

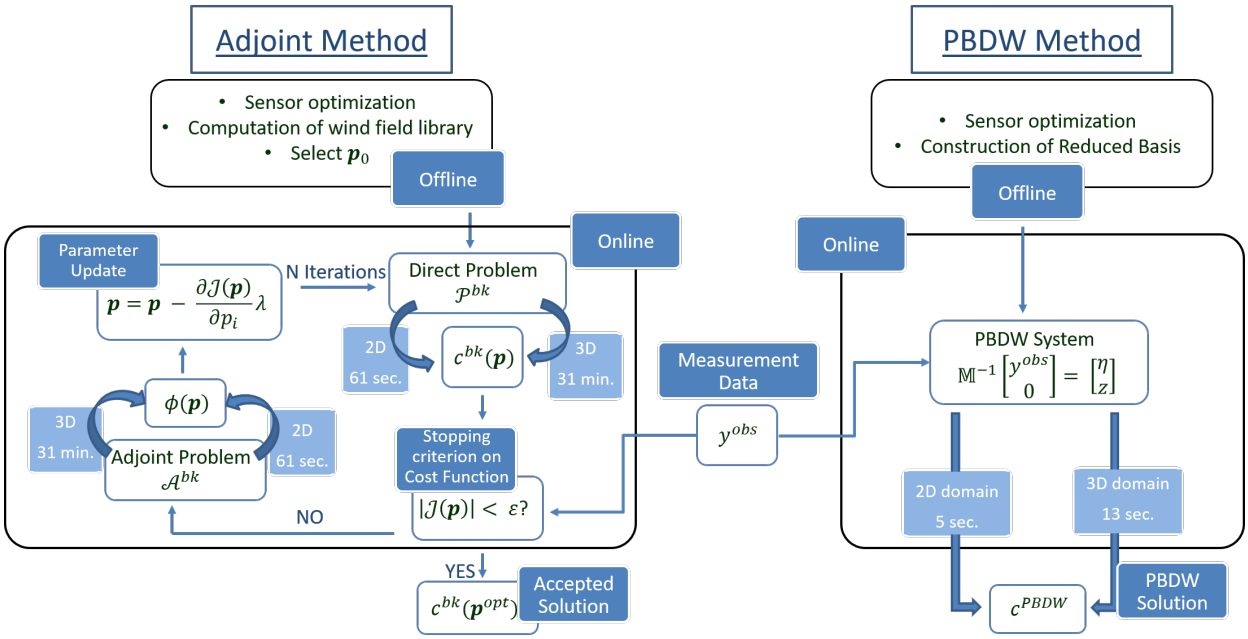

Figure 18: Schema comparing the PBDW method to adjoint-based inverse methods. 2D domain refers to the case study described in section 2. 3D domain refers to the three-dimensional urban domain studied in [38] (chapter 7).

6. Conclusions

In this paper we presented the PBDW state estimation method for non-intrusive real-time data assimilation, and give an exploratory application of its extension to dispersion modeling over a large outdoor domain, representing the predominant terms present in CFD-based AQMs by an imperfect model. This method shows great promise for extension to more complicated case studies in the AQM context. We discussed the advantages of the PBDW method with respect to other data assimilation methods, such as inverse methods, and we discussed the importance of sensor placement, giving a possible method of improving data points based on the physical quantity to measure. We also provided a modified norm in the variational formulation in order to improve the quality of the Update contribution given the dimensionality gap between calculation domain scale and sensor size. We then presented the results of the PBDW state estimation in the case of a perfect $\mathcal{P}^{bk}$ model (and thus only parametric variation), as well as the cases of an imperfect model. We found that in the case of significant model error the PBDW method was able to approximate the physical state with an overall error of $\sim 3\%$ and no more than 15% peaks.

When compared to the GEIM approximation, results were similar between the two methods with little model error, but the PBDW method proves advantageous in the case of significant model error. Computational times of the two reduction methods are similar, however, the GEIM does have the slight advantage of a smaller linear system. This advantage is outweighed however by the PBDW's improved ability to correct model error. An important conclusion of this paper is that the definition using (20) of the properly scaled Riesz representation in (9) greatly affects the ability of the PBDW to correct model error.

We aimed in this study to demonstrate the feasibility of RBMs to represent dispersion phenomena in the context of air quality data assimilation and modeling, and the ability of the PBDW to contribute to the use of parameterized PDE models by reducing computational costs and accounting for unmodeled physics. The results presented above are encouraging, and show that this method may prove very useful in operational air quality studies relying on physically-based deterministic models, if adapted and implemented properly for the case of study. The implementation on a more operational model and comparison to real measurements would be the next step in the validation of this method for AQM applications.

Acknowledgements

This research was done at and financed by the French Institute of Science and Technology for Transport, Development, and Networks.

AppendixA. Greedy Algorithm

Algorithm 1 : Weak Greedy algorithm to construct $\mathcal{Z}^N$

Initialization: GIVEN

$\Xi_{test} = (\mathbf{p}_1, \ldots, \mathbf{p}_{n_{train}}) \in \mathcal{D}^{n_{train}}$, $n_{train} >> 1$

CHOOSE RANDOMLY $\mathbf{p}_1 \in \mathcal{D}$

SET $S_1 = \{\mathbf{p}_1\}$ and $\mathcal{X}_h^1 = \text{span}(c_h^{bk}(\mathbf{p}_1))$.

for $N = 2$ to N_{max} **do**

$\mathbf{p}_N = \underset{\mathbf{p}\in\Xi_{test}}{argmax} \frac{\|c_h^{bk}(\mathbf{p}) - P_{N-1}c_h^{bk}(\mathbf{p})\|_{H^1}}{\|c_h^{bk}(\mathbf{p})\|_{H^1}}$

(where P_{N-1} is the H^1-orthogonal projection operator from $\mathcal{X}_h$ into $\mathcal{X}_h^{N-1}$)

$S_N = S_{N-1} \cup \mathbf{p}_N$

$\mathcal{X}_h^N = \mathcal{X}_h^{N-1} + \text{span}(c_h^{bk}(\mathbf{p}_N))$

end for

[1] W. H. Organization, Ambient air pollution: A global assessment of exposure and burden of disease, Tech. rep. (2016).

[2] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, A. Baklanov, Real-time air quality forecasting, part I: History, techniques, and current status, Atmospheric Environment 60 (2012) 632–655.

[3] B. Sportisse, Fundamentals in Air Pollution, Springer Netherlands, Dordrecht, 2010.

[4] J. Hu, H. Zhang, S.-H. Chen, C. Wiedinmyer, F. Vandenberghe, Q. Ying, M. J. Kleeman, Predicting Primary PM $_{2.5}$ and PM $_{0.1}$ Trace Composition for Epidemiological Studies in California, Environmental Science & Technology 48 (9) (2014) 4971–4979.

[5] F. Deutsch, C. Mensink, J. Vankerkom, L. Janssen, Application and validation of a comprehensive model for PM10 and PM2.5 concentrations in Belgium and Europe, Applied Mathematical Modelling 32 (8) (2008) 1501–1510.

[6] C. Yuan, E. Ng, L. K. Norford, Improving air quality in high-density cities by understanding the relationship between air pollutant dispersion and urban morphologies, Building and Environment 71 (2014) 245–258.

[7] P. Gousseau, B. Blocken, T. Stathopoulos, G. 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 (2) (2011) 428–438.

[8] C. K. Wikle, L. M. Berliner, A Bayesian tutorial for data assimilation, Physica D: Nonlinear Phenomena 230 (1-2) (2007) 1–16. doi:10.1016/j.physd.2006.09.017.
URL http://linkinghub.elsevier.com/retrieve/pii/S016727890600354X

[9] F. Bouttier, P. Courtier, Data assimilation concepts and methods, ECMWF Meteorological Training Course Lecture Series, 14, ECMWF, Reading, UK (1999) 59.

[10] R. E. Kalman, others, A new approach to linear filtering and prediction problems, Journal of basic Engineering 82 (1) (1960) 35–45.
URL https://fluidsengineering.asmedigitalcollection.asme.org/pdfaccess.ashx?resourceid=4506257&pdfsource=13

[11] O. Roustant, D. Ginsbourger, Y. Deville, DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization, Journal of Statistical Software 51 (1) (2012) 54p.
URL https://hal-emse.ccsd.cnrs.fr/emse-00741762

[12] A. Nassiopoulos, F. Bourquin, Fast three-dimensional temperature reconstruction, Computer Methods in Applied Mechanics and Engineering 199 (49) (2010) 3169–3178.

[13] J. Waeytens, P. Chatellier, F. Bourquin, Inverse computational fluid dynamics: Influence of discretization and model errors on flows in water network including junctions, Journal of Fluids Engineering 139 (5) (2017) 051402–051402–10.

[14] A. Sandu, D. N. Daescu, G. R. Carmichael, T. Chai, Adjoint sensitivity analysis of regional air quality models, Journal of Computational Physics 204 (1) (2005) 222–252.

[15] V. Mons, L. Margheri, J.-C. Chassaing, P. Sagaut, Data assimilation-based reconstruction of urban pollutant release characteristics, Journal of Wind Engineering and Industrial Aerodynamics 169 (2017) 232–250. doi:10.1016/j.jweia.2017.07.007.
URL https://linkinghub.elsevier.com/retrieve/pii/S016761051630633X

[16] C. Prud'homme, D. V. Rovas, K. Veroy, L. Machiels, Y. Maday, A. T. Patera, G. Turinici, Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods, Journal of Fluids Engineering 124 (1) (2002) 70–80.

[17] T. Lassila, A. Manzoni, A. Quarteroni, G. Rozza, A reduced computational and geometrical framework for inverse problems in hemodynamics, International journal for numerical methods in biomedical engineering 29 (7) (2013) 741–776.

[18] F. Negri, G. Rozza, A. Manzoni, A. Quarteroni, Reduced basis method for parametrized elliptic optimal control problems, SIAM Journal on Scientific Computing 35 (5) (2013) A2316–A2340.

[19] E. Bader, M. Karcher, M. A. Grepl, K. Veroy, Certified Reduced Basis Methods for Parametrized Distributed Elliptic Optimal Control Problems with Control Constraints, SIAM Journal on Scientific Computing 38 (6) (2016) A3921–A3946.

[20] A. Quarteroni, G. Rozza, A. Quaini, Reduced basis methods for optimal control of advection-diffusion problems, in: Advances in Numerical Mathematics, RAS and University of Houston, 2007.

[21] L. Dedè, Reduced basis method and error estimation for parametrized optimal control problems with control constraints, Journal of Scientific Computing 50 (2) (2012) 287–305.

[22] M. Karcher, S. Boyaval, M. Grepl, K. Veroy, Reduced basis approximation and a posteriori error bounds for 4D-Var data assimilation (2017).

[23] Y. Maday, A. T. Patera, J. D. Penn, M. Yano, A parameterized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics, Int. J. Numer. Meth. Engng 102 (5) (2015) 933–965.

[24] Y. Maday, A. Patera, J. Penn, M. Yano, PBDW state estimation: Noisy observations; configuration-adaptive background spaces; physical interpretations., Proceedings SMAI CANUM 2014, Carry-le-Rouet, France, in ESAIM: Proceedings and Surveys.

[25] Y. Maday, O. Mula, A. T. Patera, M. Yano, The Generalized Empirical Interpolation Method: Stability theory on Hilbert spaces with an application to the Stokes equation, Computer Methods in Applied Mechanics and Engineering 287 (2015) 310–334.

[26] Y. Maday, O. Mula, A generalized empirical interpolation method: application of reduced basis techniques to data assimilation, in: Analysis and numerics of partial differential equations, Springer, 2013, pp. 221–235.

[27] F. Archambeau, N. Mehitoua, M. Sakiz, Code Saturne: a Finite Volume Code for the Computation of Turbulent Incompressible flows Int, J. Finite Volumes, Electronical edition: http://averoes. math. univ-paris13. fr/html.

[28] H. G. Kim, V. C. Patel, Test Of Turbulence Models For Wind Flow Over Terrain With Separation And Recirculation, Boundary-Layer Meteorology 94 (1) (2000) 5–21. doi:10.1023/A:1002450414410.
URL http://link.springer.com/10.1023/A:1002450414410

[29] C. Argyropoulos, N. Markatos, Recent advances on the numerical modelling of turbulent flows, Applied Mathematical Modelling 39 (2) (2015) 693–732. doi:10.1016/j.apm.2014.07.001.
URL https://linkinghub.elsevier.com/retrieve/pii/S0307904X14003448

[30] L. Fick, Y. Maday, A. T. Patera, T. Taddei, A stabilized pod model for turbulent flows over a range of reynolds numbers: optimal parameter sampling and constrained projection, Journal of Computational Physics.

[31] F. Hecht, New development in FreeFem++, Journal of numerical mathematics 20 (3-4) (2012) 251–266.

[32] T. J. Hughes, A. Brooks, A multidimensional upwind scheme with no crosswind diffusion, Finite element methods for convection dominated flows 34 (1979) 19–35.

[33] A. N. Brooks, T. J. R. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, Computer Methods in Applied Mechanics and Engineering 32 (1) (1982) 199–259.

[34] A. Kolmogoroff, Uber die beste Annaherung von Funktionen einer gegebenen Funktionenklasse, Annals of Mathematics (1936) 107–110.

[35] A. Cohen, R. Devore, Kolmogorov widths under holomorphic mappings, IMA Journal of Numerical Analysis 36 (1) (2016) 1–12.

[36] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, P. Wojtaszczyk, Convergence Rates for Greedy Algorithms in Reduced Basis Methods, SIAM J. Math. Anal. 43 (3) (2011) 1457–1472. doi:10.1137/100795772.

[37] T. Taddei, Model order reduction methods for data assimilation; state estimation and structural health monitoring, Ph.D. thesis, Massachusetts Institute of Technology (2016).

[38] J.K. Hammond, Reduced Basis Methods for Urban Air Quality Modeling, Ph.D. thesis, Université Paris Est (Nov. 2017).