Advanced Modeling and Simulation in Engineering Sciences

## RESEARCH ARTICLE

**Open Access**

# Meta-modeling of a simulation chain for urban air quality

J. K. Hammond[1*] ⬤, R. Chen[2] and V. Mallet[1]

*Correspondence:
janelle.hammond@inria.fr
[1] INRIA, 2 Rue Simone IFF, 75012
Paris, France
Full list of author information is
available at the end of the article

## Abstract

Urban air quality simulation is an important tool to understand the impacts of air pollution. However, the simulations are often computationally expensive, and require extensive data on pollutant sources. Data on road traffic pollution, often the predominant source, can be obtained through sparse measurements, or through simulation of traffic and emissions. Modeling chains combine the simulations of multiple models to provide the most accurate representation possible, however the need to solve multiple models for each simulation increases computational costs even more. In this paper we construct a meta-modeling chain for urban atmospheric pollution, from dynamic traffic modeling to air pollution modeling. 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 non-intrusive reduced order scheme, and study its application to a full chain of operational models. Reduced basis are constructed using principal component analysis (PCA), and the concentration fields are approximated as projections onto this reduced space. We use statistical emulation to approximate projection coefficients in a non-intrusive manner. We apply a multi-level meta-modeling technique to a chain using the dynamic traffic assignment model LADTA, the emissions database COPERT IV, and the urban dispersion-reaction air quality model SIRANE to a case study on the city of Clermont-Ferrand with over $45,000$ daily traffic observations, a $47,000$-link road network, a simulation domain covering $180\,\mathrm{km}^2$. We assess the results using hourly $NO_2$ concentration observations measured at stations in the agglomeration. Computational times are reduced from nearly 3 h per simulation to under 0.1 s, while maintaining accuracy comparable to the original models. The low cost of the meta-model chain and its non-intrusive character demonstrate the versatility of the method, and the utility for long-term or many-query air quality studies such as epidemiological inquiry or uncertainty quantification.

**Keywords:** Model order reduction, Surrogate model, meta-model, Reduced basis methods, urban dispersion model, Air quality simulation, Modeling chain

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 2 of 22

## Introduction

Air quality simulations at urban scale are a key tool for the evaluation of population exposure to particulate matter and gaseous air pollutants. The simulations are however subject to costly computational requirements and complicated implementation. Studies in exposure estimation or uncertainty quantification, for example, require many solutions to the model. The 2016 study by the World Health Organization [1] on the global disease burden of air pollution excluded many pollutant species and health outcomes from the study due to lack of robust evidence. The use of advanced modeling methods in air pollution studies can provide precise estimations, however lower-cost but less precise models are often used in these scenarios due to high computational costs. Advanced models can be rendered feasible in this context if we can reduce the computational cost without significant loss of accuracy.

Let us consider a generic stationary model over a physical domain $\Omega \subset \mathbb{R}^d$, with $d = 2$ or 3, and parameter domain $\mathcal{D} \subset \mathbb{R}^{N_p}$

$$\mathcal{M} : \mathcal{D} \to \mathbb{R}^{\mathcal{N}}$$
$$\mathbf{p} \mapsto c(\mathbf{p})$$

The model output for a given parameter vector $\mathbf{p} \in \mathcal{D}$, $c(\mathbf{p}) \in \mathbb{R}^{\mathcal{N}}$, will be a large-dimension vector representing the solution over a grid covering $\Omega$. $\mathcal{M}$ can represent various types of atmospheric pollution models, from highly complex formulations based on partial differential equations and fluid dynamics [2,3] to simpler, and more commonly operational, formulations such as Gaussian dispersion models. Even in the case of the (comparatively) simpler models, the computational time necessary for the solution of $\mathcal{M}$ in practical applications over large domains with many parameters (e.g., emissions sources) can be high. This would make numerous solutions to the model too costly in practice. Methods of model order reduction (MOR) can reduce computational costs without introducing significantly increased model error, and for a range of varying parameters $\mathbf{p} \in \mathcal{D}$.

Various MOR techniques have been studied in the context of air quality models (AQMs). In [4] the meta-modeling technique using statistical emulation by radial basis functions (RBF) was tested on pollutant concentration fields over Clermont-Ferrand approximated by the ADMS-Urban model [5] using daily profiles for traffic emissions. In [6], statistical emulation was used to evaluate the sensitivity of some input parameters on a global aerosol model. A Gaussian process emulation was used for the study of model uncertainty in [7] for accidental release scenarios. Gaussian process emulation was also used in [8] for the Sobol' sensitivity analysis of a dispersion model representing the Fukushima event.

In this paper, we will consider a modeling chain for air quality modeling over the agglomeration of Clermont-Ferrand and surrounding area in France. Air quality models are known to commit significant errors [2,9–11], however these errors are strongly dependent on the calibration and inputs to the model. Providing more precise input data, such as data on pollutant emissions from road traffic, can greatly improve the accuracy of the modeled concentration field. The advantage of a modeling chain is the use of the best (most precise) information available on various inputs by using traffic and emissions models. In [9], the authors provide a review of modeling chain techniques for traffic pollutant emissions, atmospheric dispersion, and effects on water quality.

The modeling chain studied here consists of the dynamic traffic assignment model LADTA [12,13], an emissions model Pollemission [14] based on COPERT-IV emissions database [15], and an urban AQM, Sirane [16]. The computation of a pollutant concentration field over the agglomeration for any given time requires the solution of each model in the chain, which proves costly for long time periods.This brings us back to MOR techniques. However in this case, we have a chain of multiple models to reduce, which leads us to questions on the implementation of MOR techniques: whether to build a single reduction over the full chain, or a chain of meta-models? How can we treat the large parameter dimension of the chain? The use of modeled traffic emissions here presents additional difficulty in the construction of an air quality meta-model, due to the increased spatial and temporal variation of pollution emissions (compared to daily profiles or averaged emissions).

We resort to projection-based MOR techniques based on reduced basis (RB) [17] to construct cheap and accurate meta-models. A projection-based meta-model for the dynamic traffic model was built in [18]. Here we will complete the model chain with the conversion from traffic assignment and emissions model outputs on a coarse traffic network to pollutant dispersion model inputs on a fine traffic network. We then construct a meta-model for the AQM using statistical emulation by RBF interpolation with a weighted distance on the parameter domain to build a low-cost meta-model chain for the entire system. The motivation for this choice will be discussed in detail in "Case study on Clermont-Ferrand" section. An important aspect of the selected MOR method is its non-intrusive character. Among non-intrusive methods, various techniques are used to approximate the coefficients of a projection onto the reduced basis without relying on the equations of the original model. Refs. [19] and [20] present a two-grid non-intrusive method using a rapid low-fidelity numerical simulation followed by a post-processing step to aproximate the reduced basis solution from high-fidelity numerical simulation. This was applied to computational fluid dynamics and to a geotechnics problem with non-linear behavior, respectively. In [21], a non-intrusive reduced order data assimilation method was applied to particle dispersion in the case of sufficiently numerous measurement data, using a reduced basis of the model solution manifold and a second basis representing the available measurement data to correct model error. In [22], a regression mapping training inputs to the coefficients of the projected model output is approximated using an artificial neural network, and is tested on a one-dimensional unsteady combustion problem. In [23], a non-intrusive method is applied to stress tensor field reconstruction of a parametrized beam and pressure field reconstruction in computational fluid mechanics. This method also employs POD interpolation [24] to reconstruct reduced basis projection coefficients, and treats parameter domain reduction based on sensitivity analysis using a coupling with active subspaces, which can be useful in the case of problems presenting a low-dimension active subspace.

In "Meta-modeling methods" section, we will describe the meta-modeling technique based on RB methods. In "Case study on Clermont-Ferrand" section, we will describe the case study over Clermont-Ferrand: input and measurement data, computational domain, and selected models. In "Results" section, we will summarize the results of the meta-model on the AQM chain, studying accuracy, precision, and computational savings. The full meta-model chain will reduce computational costs to under 0.1 s per simulation while

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 4 of 22

maintaining comparible accuracy, which will allow us to use the chain for high numbers of simulations in future work.

## Meta-modeling methods

Computation times for large problems are commonly on the order of hours, making many-query contexts, such as sensitivity analysis and optimization, hardly feasible. Model reduction methods are of great interest to applications of parametrized problems involving many-query or real-time study. We will begin here by detailing the MOR method as applied to the AQM part of the chain, and we will discuss the details of the full meta-model chain in "Case study on Clermont-Ferrand" section.

### Reduced basis method

We will rely on a projection-based method of model order reduction using a reduced basis. Let us consider a model, or model chain, $\mathcal{M}$ which takes input parameter vector $\mathbf{p} \in \mathcal{D} \subset \mathbb{R}^{N_p}$ and computes an output vector $c(\mathbf{p})$ over a grid of $\mathcal{N}$ points. We will define the output solution set to the model $\mathcal{X}_\mathcal{N} = \{c(\mathbf{p}) | \mathbf{p} \in \mathcal{D}\} \subset \mathbb{R}^\mathcal{N}$, where the parameter dimension is $N_p$. Reduced basis methods exploit the parametrized structure of the model and construct a low-dimensional space approximating the solution set $\mathcal{X}_\mathcal{N}$ [25,26]. While the discrete model output is of high dimension $\mathcal{N}$, the reduced order solution will be of dimension $N \ll \mathcal{N}$. A key factor of the reduced basis methods is the small Kolmogorov n-width [27]. The n-width measures to what extent $\mathcal{X}_\mathcal{N}$ can be approximated by an n-dimensional subspace, and can be studied during the sampling of the solution space.

Our objective is to construct a reduced basis $\{\Psi_n^{AQ}\}_{1 \leq n \leq N}$ of $N$ basis functions such that the projection of any simulated state, $\Pi_N c(\mathbf{p})$, onto the reduced basis is sufficiently precise. The basis representing atmospheric concentration fields will be denoted by AQ (air quality). To construct a RB, we first need to sample a large number of solutions in $\mathcal{X}_\mathcal{N}$. This so-called *training* set should represent the variability in the solution states. We will sample the solution space by Latin Hypercube Sampling (LHS). Sampling by LHS was chosen for this study because many of the parameters are independent in practice. In addition this allows more flexibility when using the meta-model in the (quite realistic) case of uncertain parameters, or in rare scenarios such as pollution peaks, where a reliable meta-model is necessary but not guaranteed if it is trained over the most likely input values. Next we will construct the RB by principal component analysis (PCA).

We use LHS to select $N_{train}$ sample points $(\mathbf{p}_1, \ldots, \mathbf{p}_{N_{train}})$ in the parameter domain $\mathcal{D}$, and compute model simulations from each point to build the training ensemble $\mathbf{Y}^{AQ} = [c(\mathbf{p}_1), \ldots, c(\mathbf{p}_{N_{train}})]$ to train the model reduction. As is common practice in PCA applications, we will first compute the ensemble mean $\bar{c} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} c(\mathbf{p}_i)$ of the training ensemble. PCA is computed on the centered ensemble $\bar{\mathbf{Y}}^{AQ} = [c(\mathbf{p}_1) - \bar{c}, \ldots, c(\mathbf{p}_{N_{train}}) - \bar{c}]$. The eigenvalues $\{\lambda_k\}_{1 \leq k \leq \mathcal{N}}$ and eigenvectors $\{\Psi_k^{AQ}\}_{1 \leq k \leq \mathcal{N}}$ of the covariance matrix $\bar{\mathbf{C}}^{AQ} = (\bar{\mathbf{Y}}^{AQ})^T (\bar{\mathbf{Y}}^{AQ})$ of the training ensemble are such that

$$\sum_{i=1}^{N_{train}} \left\| c(\mathbf{p}_i) - \bar{c} - \sum_{n=1}^{N} \Psi_n^{AQ} \Psi_n^{AQ^T} (c(\mathbf{p}_i) - \bar{c}) \right\|_2^2 = \sum_{k=N+1}^{\mathcal{N}} \lambda_k, \tag{1}$$

for eigenvalues $\lambda$ arranged in decreasing order. $N = 5$ principle component basis functions $\psi_n^{AQ}$ are selected to represent $I_N = 98\%$ of the variability in the concentration state,

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 5 of 22

where the Relative Information Content is $I_N = \frac{\sum_{k=1}^{N} \lambda_k}{\sum_{k=1}^{N_{train}} \lambda_k}$. This means that the error of projecting any member of the training ensemble onto the basis, $Err_N$, will be bounded by the tolerance $Err_N \leq \epsilon_N = \sqrt{1 - I_N}$ [25]. The 98% tolerance cutoff is selected on a case-by-case basis: the goal is to keep $N$ small and $I_N$ as close to 1 as possible. Here the 98% precision is attained relatively quickly, then improvement slows as $N > 5$ increases. For any new parameter, we can thus represent the solution as

$$c(\mathbf{p}) \simeq \Pi_N c(\mathbf{p}) = \bar{c} + \sum_{n=1}^{N} \alpha_n^{AQ} \Psi_n^{AQ} \tag{2}$$

with projection coefficients $\alpha_n^{AQ} = \Psi_n^{AQ^T}(c(\mathbf{p}) - \bar{c})$.

**Statistical emulation**

Once we have constructed the reduced basis by PCA, we need a reduced order modeling scheme to approximated new solutions. Classical reduced basis methods which replace the approximation space with the reduced basis space are intrusive and require the modification of the computational code. We would like to use a non-intrusive method which can be applied to a black-box model or model chain, which is particularly pertinent in the context of operational models. The non-intrusive implementation allows the freedom of choice of the best available model. In particular, this allows the models to be updated with technological advances, and a model chain which is meta-modeled by linking non-intrusive meta-models maintains maximal versatility. It also makes for simpler implementation, as the calculation code does not need to be modified. While many MOR methods exist, the non-intrusive character of few of these methods is particularly advantageous in problems relying on operational models.

We consider meta-modeling by the emulation of projection coefficients $\alpha_n^{AQ}, 1 \leq n \leq N$. First we select a linear *trend*, which will be a least squares regression $\mathcal{R}_n(\mathbf{p}) = \sum_{k=1}^{N_p} \beta_{n,k} p_k$, calculated from the training simulations $\{c(\mathbf{p}_i)\}_{1 \leq i \leq N_{train}}$. To this we add an interpolation term on the residuals $\alpha_n(\mathbf{p}_i) - \mathcal{R}_n(\mathbf{p}_i), \mathcal{I}_N(\mathbf{p}) = \sum_{i=1}^{N_{train}} \omega_{n,i} \phi(d_\theta(\mathbf{p}, \mathbf{p}_i))$. We chose to compute this interpolation using RBF. We chose cubic RBFs $\phi$ and a weighted Euclidean distance $d_\theta(\cdot, \cdot)$ to represent the varying ranges of each input parameter.

$$d_\theta(\mathbf{p}_1, \mathbf{p}_2) = \sqrt{\sum_{i=1}^{N_p} \theta_i (\mathbf{p}_1^i - \mathbf{p}_2^i)^2}, \tag{3}$$

where $\theta_i = \dfrac{1}{\left(\max\limits_{\mathbf{p} \in \mathcal{D}} \mathbf{p}^i - \min\limits_{\mathbf{p} \in \mathcal{D}} \mathbf{p}^i\right)^2}$. We then define the emulated projection coefficients as follows.

$$\hat{\alpha}_n^{AQ}(\mathbf{p}) = \underbrace{\sum_{k=1}^{N_p} \beta_{n,k} p_k}_{\text{Least squares regression}} + \underbrace{\sum_{i=1}^{N_{train}} \omega_{n,i} \phi(d_\theta(\mathbf{p}, \mathbf{p}_i))}_{\text{Residual interpolation}}. \tag{4}$$

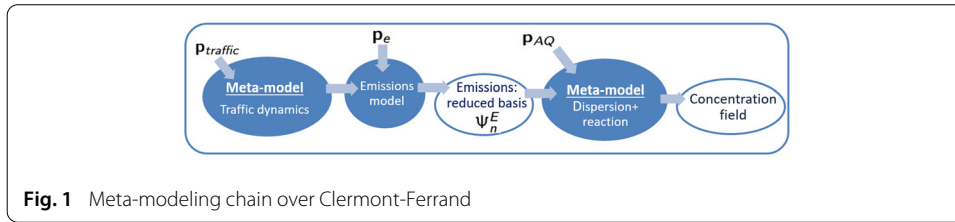Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 6 of 22



**Fig. 1** Meta-modeling chain over Clermont-Ferrand

The weights $\{\omega_{n,i}\}_{1\leq n\leq N;1\leq i\leq N_{train}}$ are chosen such that the interpolation is exact for the sample points $\{\mathbf{p}_i\}_{1\leq i\leq N_{train}}$,

$$\hat{\alpha}_n^{AQ}(\mathbf{p}_j) = \alpha_n^{AQ}(\mathbf{p}_j) = \sum_{k=1}^{N_p} \beta_{n,k}p_{j,k} + \sum_{i=1}^{N_{train}} \omega_{n,i}\phi\big(d_\theta(\mathbf{p}_j,\mathbf{p}_i)\big). \tag{5}$$

The emulated solution is finally

$$\hat{c}_N(\mathbf{p}) = \bar{c} + \sum_{n=1}^{N} \hat{\alpha}_n^{AQ}\Psi_n^{AQ}. \tag{6}$$

The regression represents the relation between the model parameters and the RB projection coefficients, and computed from the training set $(\mathbf{p}_i, \alpha(\mathbf{p}_i))_{1\leq i\leq N_{train}}$. This provides an initial trend to be corrected by the interpolation. In practice, the interpolation of the residual is the most important part of the emulation. The size of the training set $N_{train}$ plays an important role in the precision of this emulation, as the regression and interpolation are trained on this set. In [4], this method of approximating projection coefficients is compared to approximation by Kriging. The two meta-models showed similar results, and we chose RBF emulation for its simpler (and thus more accessible in operational applications) implementation and lower computational cost.

## Case study on Clermont-Ferrand

In this work we will apply the meta-modeling method described in "Meta-modeling methods" section to a modeling chain over the city of Clermont-Ferrand in France. We will build a meta-model chain representing road traffic emissions and the dispersion and reaction of pollutants over the urban agglomeration and surrounding area using data over a 2-year period form 2013 to 2015. The model chain is represented in Fig. 1.

### Traffic emissions modeling

Traffic emissions modeling is done using the dynamic traffic assignment model LADTA. A meta-model was constructed [28] to represent the traffic flow and speed simulations over a road network of 19, 628 oriented links, where nearly 45, 000 traffic flow observations are available each day. Emissions of $NO_x$ and PM are computed using Pollemission code [29] based on the COPERT-IV emissions database [30, 31]. A detailed description of this section of the modeling chain and its input parameters can be found in [28]. The varying input parameters consist of 23 traffic parameters and 6 emissions parameters. These parameters are time-dependent or considered sources of uncertainty. They include temporal traffic demand, computed using traffic observations, the capacity and speed limits of traffic network links, multiplicative coefficients on origin-destination matrices representing the spatial distribution of traffic demand, traffic direction (morning versus evening), engine

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 7 of 22

size, type, and emission standards of the vehicle fleet, and ratio of heavy-duty vehicles to personal cars.

The emissions model provides traffic emissions estimations for $NO_x$ and $PM_10$. However the atmospheric pollution model incorporates chemical reaction parametrizations which treat $NO_2$, $NO$, $PM_{2.5}$, and $PM_10$. In order to approximate emissions of $NO_2$, $NO$, $PM_{2.5}$, and $PM_10$, we would like to estimate what proportion of $NO_x$ consists of $NO$, and what proportion of $PM_10$ is $PM_{2.5}$. In the deterministic case, we set the ratio $\frac{NO_2}{NOx} = 0.15$ [32–34], and the ratio $\frac{PM_{2.5}}{PM_{10}} = 0.75$ [35,36]. In order to construct a meta-model which can account for varied or uncertain speciation ratios, we will draw LHS parameters for the training ensemble in the intervals $(p_{NO_2}, p_{PM_{2.5}}) \in [0.1, 0.25] \times [0.65, 0.8]$. The output of the traffic-emissions coupling is the emissions on each link of the traffic network in $g/15\,\mathrm{min}$.
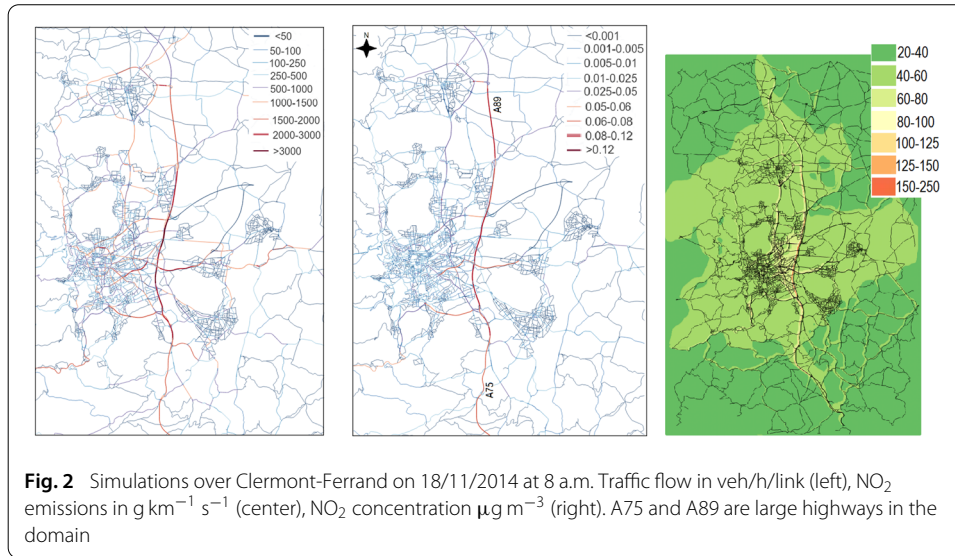
### Air quality modeling

Air quality modeling is done using the urban dispersion-reaction model Sirane [16,37] over a simulation domain of 180 km$^2$. Sirane is used as a static model which approximates the solution at a given time of the transport-reaction equations satisfied by the pollutant concentrations. The traffic emissions over a relatively coarse road network are converted to $g/s$/link on a finer network representing over $47,000$ line sources. For the calculation of $NO_2$ concentrations, we provide the so-called background concentrations of pollutant species involved directly or indirectly in the formation of $NO_2$. The background concentrations, provided for $NO_2$, $PM_10$, and $O_3$, represent the imported concentrations of pollutants, that is, concentrations transported from other locations to the city, and from the dispersion or reaction of previous emissions in the case of stationary solution. We will provide line emissions inputs on $NO_2$, $NO$, $PM_{2.5}$, and $PM_10$. Input data on meteorological conditions (wind velocity, cloud coverage, a precipitation parameter, and temperature) and surface emissions sources are also provided. The AQM output is the $NO_2$ concentration over a grid at ground level, at 20 m resolution. Hourly concentration observations are available over 2 years at 5 stations, or around $90,000$ $NO_2$ observations for analysis of model simulation outputs.

### Modeling chain

The modeling chain consists of these three steps—traffic modeling, emissions calculation, and dispersion-reaction modeling—and the conversions between outputs and inputs. In Fig. 2 we can see the traffic flow (veh/h/link) and associated emissions $(\mathrm{g\,km^{-1}\,s^{-1}})$, and $NO_2$ concentration $(\mathrm{\mu g\,m^3})$ simulations at 8 a.m. on a Tuesday in November 2014, provided by the traffic meta-model and full air quality model. The task remains to reduce the computational time required to obtain concentration fields by constructing a meta-model for the entire chain.

#### *Surrogate modeling chain construction*

As noted above, the traffic emissions on a geographically finer road network provided as input to the air quality model represent over $47,000$ line sources. In the context of model order reduction, this represents as many parameters, which in the practice of projection-based reduction methods makes the identification of the projection coefficients

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 8 of 22



**Fig. 2** Simulations over Clermont-Ferrand on 18/11/2014 at 8 a.m. Traffic flow in veh/h/link (left), $NO_2$ emissions in $\mathrm{g\,km^{-1}\,s^{-1}}$ (center), $NO_2$ concentration $\mathrm{\mu g\,m^{-3}}$ (right). A75 and A89 are large highways in the domain

$\alpha_n^{AQ}$ dependent on 47,000 parameters unfeasible (or impossible). We thus need to reduce the complexity of the problem by reducing the dimension of the input parameters. To do so we will construct a reduced basis of the traffic emissions, again using PCA.
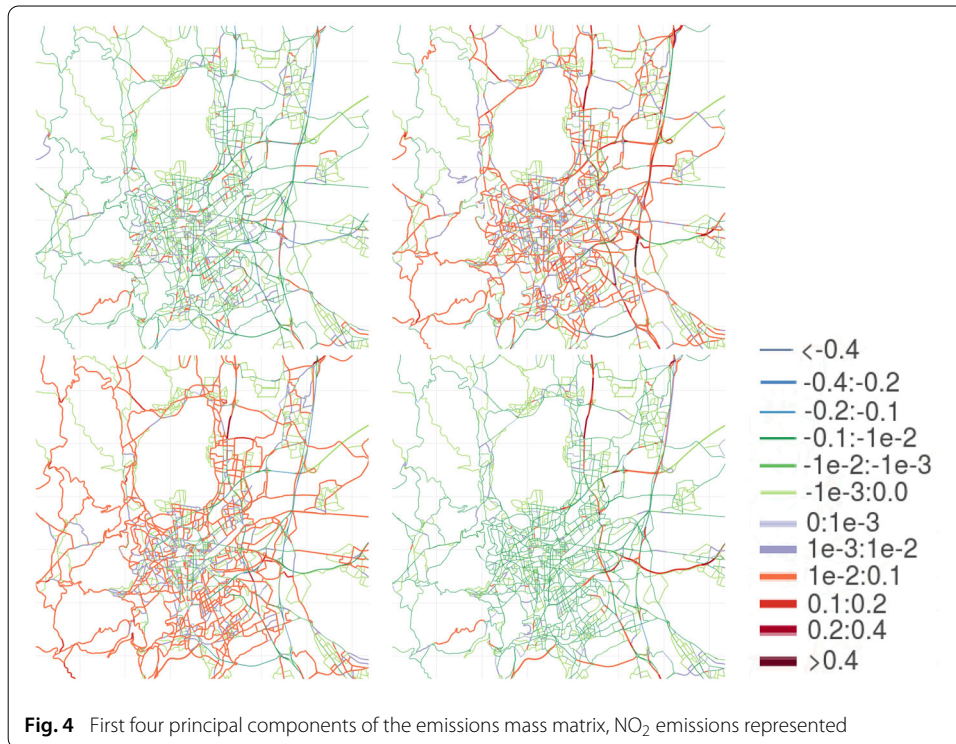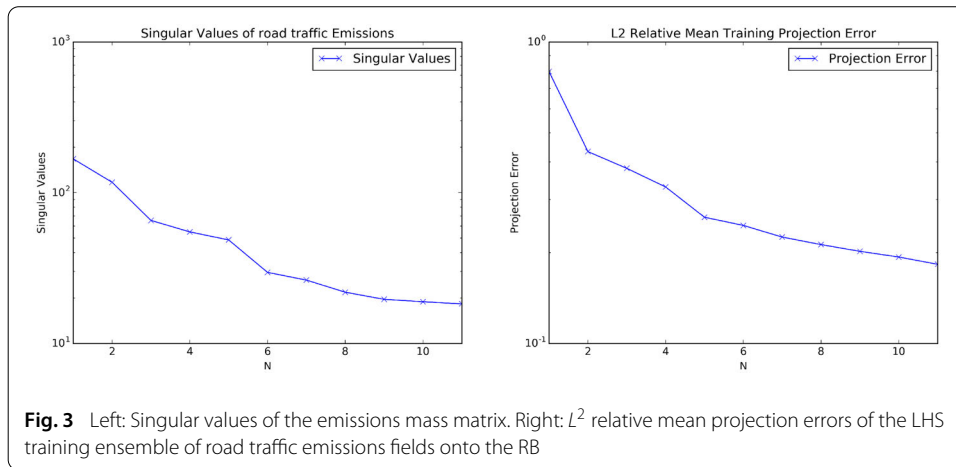
**Reduction of line emissions**  We currently have the full chain parameter vector $\mathbf{p}_{full}^T = (\mathbf{p}_{traffic}^T, \mathbf{p}_e^T, \mathbf{p}_{AQ}^T)$, where the outputs of the emissions model consist in a coefficient for each of the links in the road network. These coefficients are then treated as the (very large) input parameter vector for the air quality model. To reduce the dimension of this vector, we will use the same method as in "Reduced basis method" section. We first select a set of training parameters $(\mathbf{p}_{traffic}^T, \mathbf{p}_e^T)$ by LHS to represent the variations of these parameters in the admissible parameter space $\mathcal{D}$. We compute the emissions solutions $E(\mathbf{p}_{traffic}, \mathbf{p}_e)$ to construct a reduced basis $\{\Psi_n^E\}_{1 \leq n \leq N_{lin}}$ by PCA, representing the variations of the emissions fields centered around $\bar{E} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} E(\mathbf{p}_{traffic}^i, \mathbf{p}_e^i)$. We can compute the orthogonal projection of any emissions field onto the traffic emissions RB as follows.

$$E(\mathbf{p}_{traffic}, \mathbf{p}_e) \simeq \Pi_{N_{lin}} E(\mathbf{p}_{traffic}, \mathbf{p}_e) = \bar{E} + \sum_{n=1}^{N_{lin}} \left( (E(\mathbf{p}) - \bar{E})^T \Psi_n^E \right) \Psi_n^E = \bar{E} + \sum_{n=1}^{N_{lin}} \alpha_n^{lin} \Psi_n^E.$$

(7)

For our case study, we chose $N_{lin} = 11$ to represent 95% of the variability of the emissions solutions. This corresponds to a relative projection error tolerance over the training samples of $\epsilon_{lin}^2 = 0.05$. In the model chain, the over 47,000 line source parameters will henceforth be replaced by the $N_{lin} = 11$ projection coefficients $\{\alpha_n^{lin}\}_{n \leq N_{lin}}$, and the traffic emissions field for a given parameter approximated by its projection $\Pi_{N_{lin}} E(\mathbf{p}_{traffic}, \mathbf{p}_e)$ onto the traffic emissions RB. We perform the same reduction over the hourly surface emissions with $N_{surf} = 1$ and projection coefficient $\alpha_{surf}$. In Fig. 3, we can see the largest singular values of the PCA step, and the relative mean projection errors of the training traffic emissions simulations onto the RB $\{\Psi_n^E\}_{1 \leq n \leq N_{lin}}$, as defined by

$$Err_N = \frac{1}{N_t} \sum_{i=1}^{N_{train}} \frac{\|\Pi_N E(\mathbf{p}_i) - E(\mathbf{p}_i)\|_2}{\|E(\mathbf{p}_i)\|_2}.$$

(8)

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.(2020)7:37*

Page 9 of 22



**Fig. 3** Left: Singular values of the emissions mass matrix. Right: $L^2$ relative mean projection errors of the LHS training ensemble of road traffic emissions fields onto the RB



**Fig. 4** First four principal components of the emissions mass matrix, $NO_2$ emissions represented

In Fig. 4 we can see the first 4 principal components of the traffic emissions RB.

**Construction of the air quality meta-model** We now can write the reduced concentration model parameters $\mathbf{p}_c^T = (\alpha_{lin}^T, \alpha_{surf}^T, \mathbf{p}_{AQ}^T)$. We will construct a meta-model of the air quality model to complete the meta-modeling chain, with reduced full parameters as described in Table 1. The choice to build a separate air quality meta-model to complete the chain of meta-models (as opposed to a meta-model of the chain) was to allow multi-level assessment using traffic flow and air quality measurement data (a possibility particularly pertinent in a study of uncertainty quantcification), by a meta-modeling method which can be generalized in the case of additional models in the chain (such as an economic or

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 10 of 22

**Table 1 Summary of input parameters to the full meta-model chain**

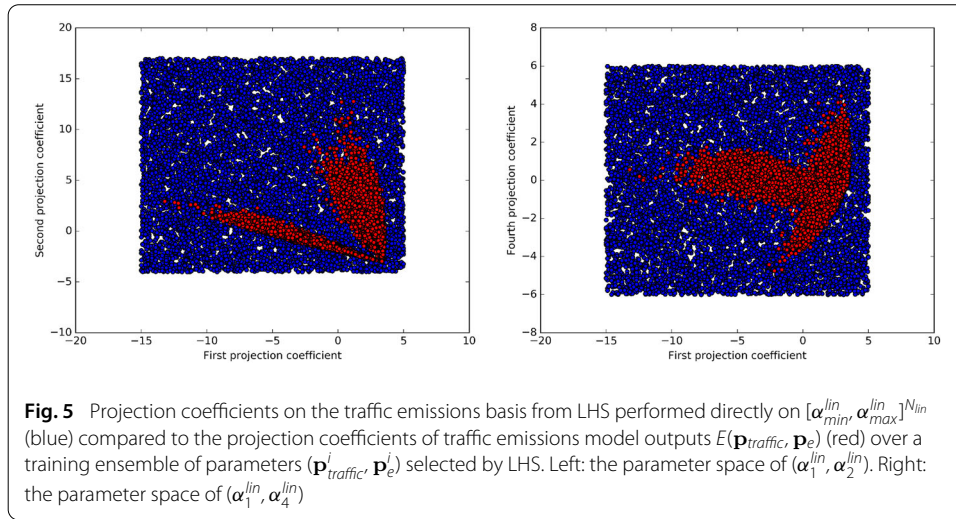| Dynamic traffic model | Emissions database | Air quality model |
|---|---|---|
| LADTA | COPERT IV | Sirane |
| $\mathbf{p}_{traffic} \in \mathbb{R}^{23}$ | $\mathbf{p}_e \in \mathbb{R}^{6}$ | $\mathbf{p}_{AQ} \in \mathbb{R}^{23}$ |
| Temporal traffic demand ; link capacity and speed limit ; distance traveled data ; direction coefficient | Vehicle fleet data | Meteorological data ; traffic, surface and railway emissions ; background pollution ; $NO_x$, PM composition parameters |

**Table 2 Model chain input parameter ranges**

| Input | Dimension | Range |
|---|---|---|
| Traffic meta-model parameters | | |
| Temporal traffic profile | 13 | [0.0, 2.0] |
| Traffic capacity | 2 | [0.5, 1.5] |
| Speed limit coefficient | 2 | [0.5, 1.5] |
| Traffic trip distance coefficient | 5 | [0, 1.8] |
| Traffic direction | 1 | [0, 1] |
| Traffic emissions parameters | | |
| Fleet engine type | 1 | [10, 100]% gasoline |
| Fleet engine capacity | 2 | [10, 100]% gasoline/diesel small-midsize |
| Emission standard | 2 | [10, 100]% gasoline/diesel after Euro3 |
| Heavy duty vehicles | 1 | [0, 30]% heave duty vehicles |
| Conversion $\frac{NO_2}{NOx}$, $\frac{PM_{2.5}}{PM_{10}}$ | 2 | $[0.1, 0.25] \times [0.65, 0.8]$ |
| Air quality model parameters | | |
| Wind velocity | 1 | $[0, 17]\frac{m}{s}$ |
| Wind direction | 1 | $[0, 360]°$ |
| Temperature | 1 | $[-10, 40]°C$ |
| Precipitation coefficient | 1 | [0, 26] |
| Cloud coefficient | 1 | [0, 8] |
| Background concentration $NO_2$, $O_3$ $PM_10$($\mu$g m$^{-3}$) | 3 | $[0, 200] \times [0, 180] \times [0, 120]$ |

epidemiological model). In addition, if a single meta-model represents the full chain, the training set must be at least as large as the largest training set in the chain, which could increase offline computational time if one model requires a larger training set than others. Here, $N_{train}^{traffic} = 3003$ and $N_{train}^{AQ} = 9347$.

When constructing the training ensemble for the air quality meta-model, we chose to draw LHS parameters for the full modeling chain $\mathbf{p}_{full}$. This choice lead to reduced variations in the emissions projection coefficients $\{\alpha_n^{lin}\}_{1 \leq n \leq N_{lin}}$ versus LHS selection over uniform distributions of the emissions projection coefficients $\alpha^{lin} \in [\alpha_{min}^{lin}, \alpha_{max}^{lin}]^{N_{lin}}$. The projection coefficients are in practice not independent; a strong first coefficient is often associated to a weaker second or third coefficient, as these principal components tend to represent different spatial distributions of the emissions. This means that the entire space $[\alpha_{min}^{lin}, \alpha_{max}^{lin}]^{N_{lin}}$ represents significantly more variation in the state $E(\mathbf{p}_{traffic}, \mathbf{p}_e)$ than the traffic-emissions model produces. By performing LHS over the full chain parameters $\mathbf{p}_{full} = (\mathbf{p}_{traffic}, \mathbf{p}_e, \mathbf{p}_{AQ}) \in \mathbb{R}^{41}$, the emissions projection coefficients are computed during the conversion of traffic meta-model outputs to concentration meta-model inputs.

In Fig. 5 we compare the parameters $\alpha_n^{lin}$ selected by these two methods by plotting the parameter spaces $(\alpha_1^{lin}, \alpha_2^{lin})$ and $(\alpha_1^{lin}, \alpha_4^{lin})$. We can see that the parameter spaces in red, which correspond to performing LHS on $\mathbf{p}_{full}$ and computing the projection coeffi-

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 11 of 22



**Fig. 5** Projection coefficients on the traffic emissions basis from LHS performed directly on $[\alpha^{lin}_{min}, \alpha^{lin}_{max}]^{N_{lin}}$ (blue) compared to the projection coefficients of traffic emissions model outputs $E(\mathbf{p}_{traffic}, \mathbf{p}_e)$ (red) over a training ensemble of parameters $(\mathbf{p}^i_{traffic}, \mathbf{p}^i_e)$ selected by LHS. Left: the parameter space of $(\alpha^{lin}_1, \alpha^{lin}_2)$. Right: the parameter space of $(\alpha^{lin}_1, \alpha^{lin}_4)$

cients $\alpha^{lin}_n$ of the traffic emissions model output $E(\mathbf{p}_{traffic}, \mathbf{p}_e)$ represents significantly less variation than LHS selection directly on the parameters $\alpha^{lin}_n$. This tactic avoids building a meta-model unnecessarily representing additional variation of the state by only considering realistic traffic emissions. In Table 2, we set the ranges of each input parameter which defines the parameter space $\mathcal{D}$.
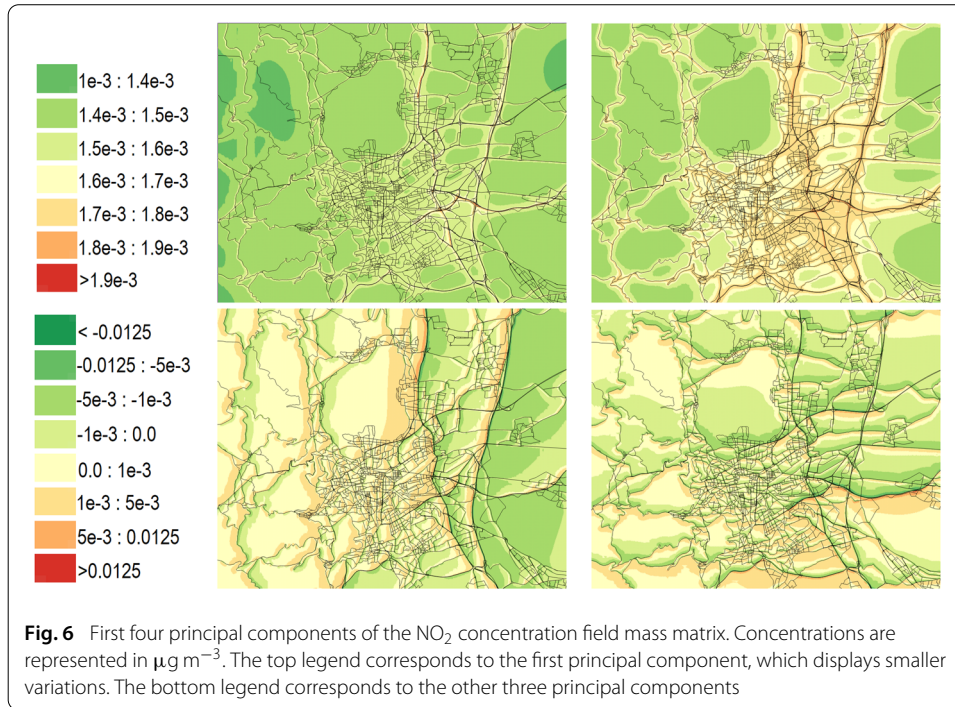
We use LHS to select a training set of $N_{train} = 9347$ concentration fields. Due to the large input parameter vector ($N_p = 41$), we used a LHS algorithm for 10, 000 training samples, and removed the concentration fields with numerical instability (this can be attributed to modeling error, which should not be confused with error in the meta-model). We use the $NO_2$ concentration fields $c(\mathbf{p}_{full})$ to construct a reduced basis $\{\Psi^{AQ}_n\}_{1 \leq n \leq N}$ by PCA, representing the variations of the concentration fields centered around the sample concentration mean $\bar{c}$. Due to the large size of the training set, we compute the largest 75 singular values and approximate $I_N \sim \frac{\sum^N_{k=1} \lambda_k}{\sum^{75}_{k=1} \lambda_k}$. We set the RB dimension $N = 5$ to represent 98% of this variability.

In Fig. 6 we see the first 4 principal components of the concentration RB. We can see that the first basis function represents urban background concentration in the denser urban areas. The second seems to represent additional pcollution from traffic. The third appears to represent situations with strong wind from the east, while the fourth shows the influence of wind from the north.

For any new parameter value, the concentration field can be approximated by the orthogonal projection onto the RB, for projection coefficients $\{\alpha^{AQ}_n\}_{1 \leq n \leq N}$,

$$c(\mathbf{p}_{full}) \simeq \Pi_N c(\mathbf{p}_{full}) = \bar{c} + \sum^N_{n=1} \alpha^{AQ}_n \Psi^{AQ}_n. \tag{9}$$

Finally we use the statistical emulation method described in "Statistical emulation" section to construct an emulator of the concentration projection coefficients $\alpha^{AQ}_n$. The full chain can be computed with a single code which applies the traffic-emissions meta-model, the calculation of emissions RB projection coefficients, and the atmospheric pollutant meta-model. This meta-model chain provides outputs on traffic flow, speed, and traffic emissions over the road network, and $NO_2$ concentrations over a 20 m-resolution grid.

**Fig. 6** First four principal components of the NO$_2$ concentration field mass matrix. Concentrations are represented in $\mu$g m$^{-3}$. The top legend corresponds to the first principal component, which displays smaller variations. The bottom legend corresponds to the other three principal components
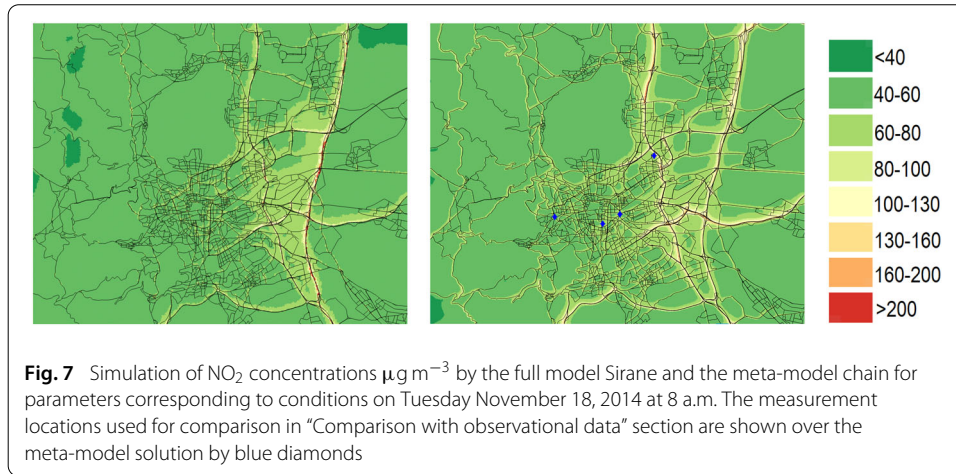
## Results

In this section, we will summarize the results of the method described in "Meta-modeling methods" section to the case study in "Case study on Clermont-Ferrand" section using data over the month of November 2014. Traffic flow measurement data serves as inputs to the model chain for deterministic simulation, and data on pollutant concentration serves to study model and meta-model performance. We will compare the meta-model output to simulations from the full model Sirane, as well as to concentration observation data, and we will assess computational savings.

### Meta-model performance

We introduce the following statistical scores commonly used for evaluation of models [4]: the normalized mean square error (NMSE), the normalized root mean square error (NRMSE), and the correlation. We define here the output functionals $\ell_o : \mathbb{R}^{\mathcal{N}} \to \mathbb{R}$ associated to each of the concentration sensors $o$, such that the observation data $y_o^{obs}(\mathbf{p}(t)) = \ell_o(c^{true}(t))$. We denote by $c^{true}(t)$ the unknown true concentration field at time $t$, and $p(t)$ the estimated parameters at time $t$.

For a data set of $M \leq N_{time}N_{obs}$ measurements (some measurements may be unavailable in practice) over $N_{time}$ times and $N_{obs}$ sensors, we use the index $m$, $1 \leq m \leq M$. $c_m = \ell_o(c(\mathbf{p}(t)))$ is the value of the output functional associated to sensor $o$ applied to the simulated state estimate at time $t$ indexed by $m$. We use the same notation here where the simulated state is the full model output $c(\mathbf{p})$ or the meta-model output $\hat{c}(\mathbf{p})$. $M$ is the total number of data available, and $y_m^{obs}$ is the $mth$ data point. $\bar{c}$ and $\bar{y}^{obs}$ are respectively the mean of $(c_m)_{1 \leq m \leq M}$ and $(y_m^{obs})_{1 \leq m \leq M}$.

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 13 of 22



**Fig. 7** Simulation of NO$_2$ concentrations $\mu$g m$^{-3}$ by the full model Sirane and the meta-model chain for parameters corresponding to conditions on Tuesday November 18, 2014 at 8 a.m. The measurement locations used for comparison in "Comparison with observational data" section are shown over the meta-model solution by blue diamonds

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (c_m - y_m^{obs})^2}. \tag{10}$$

$$\sqrt{\text{NMSE}} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \frac{(c_m - y_m^{obs})^2}{\bar{c}\bar{y}^{obs}}}. \tag{11}$$

$$\text{Correlation} = \frac{\sum_{m=1}^{M} (c_m - \bar{c})(y_m^{obs} - \bar{y}^{obs})}{\sqrt{\sum_{m=1}^{M} (c_m - \bar{c})^2} \sqrt{\sum_{m=1}^{M} (y_m^{obs} - \bar{y}^{obs})^2}}. \tag{12}$$
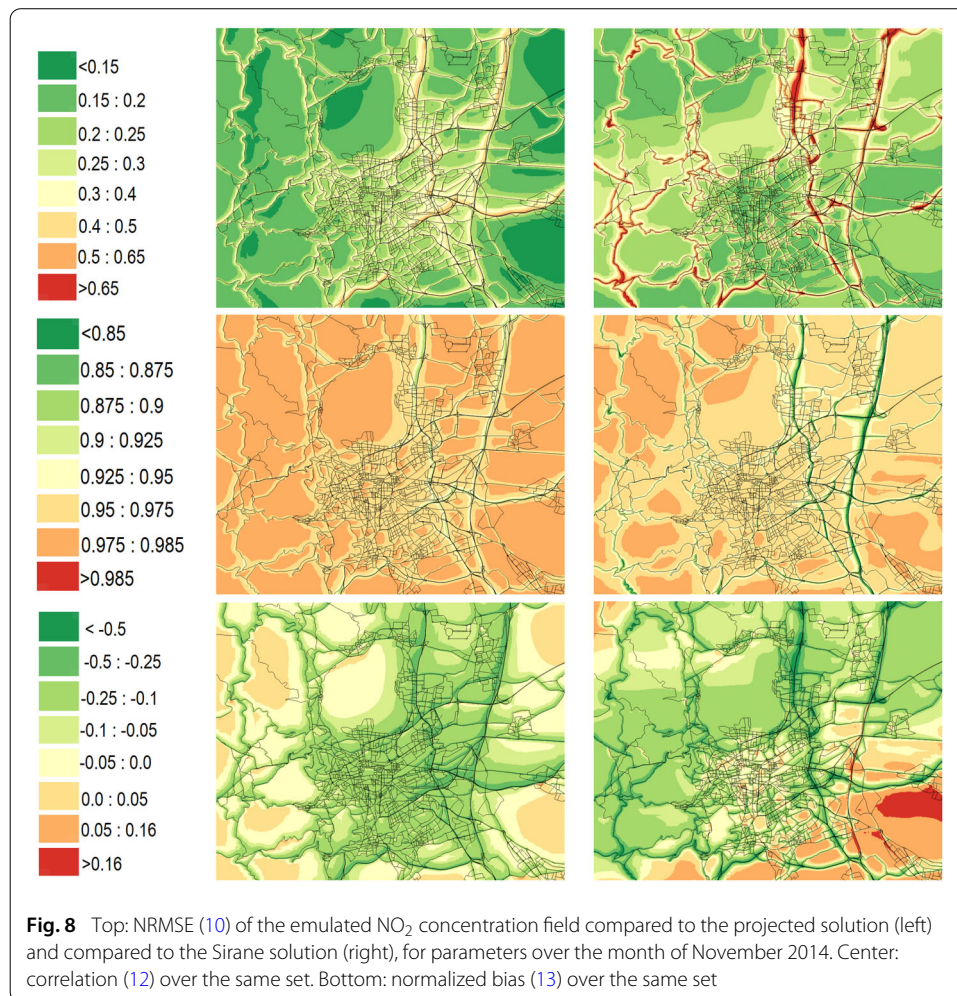
$$\text{Bias} = \frac{1}{M} \sum_{m=1}^{M} (y_m^{obs} - c_m) \tag{13}$$

Finally we define the NRMSE as $\frac{\text{RMSE}}{\bar{y}^{obs}}$, and the mean normalized root mean square error (MNRMSE) as the mean over all sensors (or grid points) of the NRMSE calculated over the concentration $c_i$ at each sensor (or grid point) over the month.

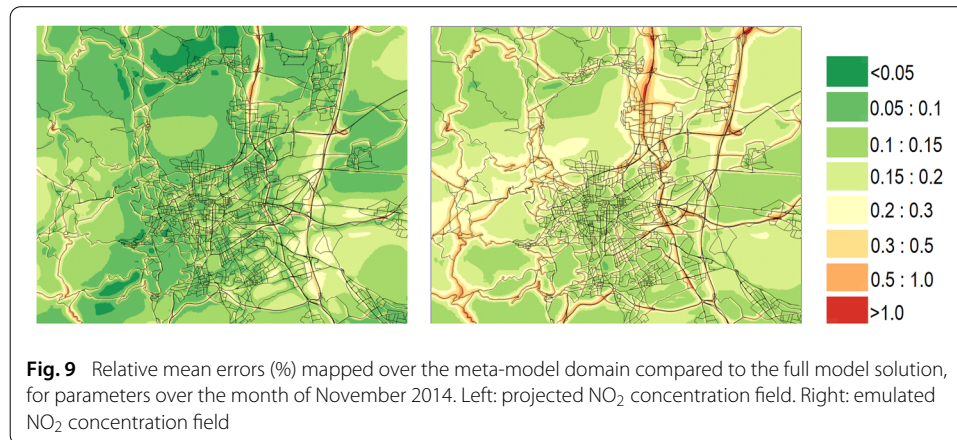$$\text{MNRMSE} = \frac{1}{N_{grid}} \sum_{i=1}^{N_{grid}} \text{NRMSE}(c_i) \tag{14}$$

### Comparison with the full model chain

We first analyze the precision of the meta-modeled concentration fields as compared to the full model Sirane. This will help us understand the ability of the meta-model to reproduce the concentration state and quantify the loss of precision caused the the dimensional reduction. In Fig. 7 we see the concentration fields of NO$_2$ simulated by the full model and the meta-model chains, as well as the sensor locations for concentration measurements. The parameters **p** correspond to conditions on Tuesday November 18, 2014 at 8 a.m. We see very similar approximations near the highways east of the city center, however the metamodel does not perfectly reproduce the variation between heavy-traffic areas and low-traffic areas. Overall the reduced order simulation is a good representation of the full model.

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 14 of 22



**Fig. 8** Top: NRMSE (10) of the emulated $NO_2$ concentration field compared to the projected solution (left) and compared to the Sirane solution (right), for parameters over the month of November 2014. Center: correlation (12) over the same set. Bottom: normalized bias (13) over the same set

In Fig. 8, we see statistical scores spatially mapped over the meta-model domain compared to both the projected solution and the full model solution. The scores of the reduced-order solution compared to the projected solution give insight into how well the RBF method of approximate the projection coefficients in order to reproduce the projected solution. The NRMSE shows that the emulated solutions perform well in approximating the urban background concentration levels, but do not capture the highest concentrations along the large highways, where we will see the highest bias levels. The correlation map also shows low correlation between the meta-model and full model only along the roadways, where the dimensional reduction has failed to capture the extent of the increased concentrations due to traffic emissions. Finally the bias map shows that the meta-model generally predicts higher concentrations in the denser urban areas when compared to the full model, again matching the trend of the dimensional reduction reducing the sensitivity of the meta-model to sharp spatial variations in concentrations. However, the areas with poor scores remain limited, and we will also consider the significant error that will inevitably be committed by the full model in the next section.

In Fig. 9 we see the relative errors of the full model concentration projected onto the reduced basis $\{\psi_n^{AQ}\}_{1 \le \le N}$, averaged over the set of deterministic simulations for the month of November 2014. We also see the emulated concentration relative error, averaged

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 15 of 22



**Fig. 9** Relative mean errors (%) mapped over the meta-model domain compared to the full model solution, for parameters over the month of November 2014. Left: projected $NO_2$ concentration field. Right: emulated $NO_2$ concentration field

**Table 3** Statistical scores of the meta-model approximation results compared to the scores of the model chain using the full air quality model

| State estimation | | | |
|---|---|---|---|
| $c(\mathbf{p})$ vs. $\hat{c}(\mathbf{p})$ | Statistical scores over 1-month simulations | | |
| Over the full grid | MNRMSE (14) | $\sqrt{\text{NMSE}}$ (11) | Correlation (12) |
| Meta-model vs. full model chain | 0.274 | 0.25 | 0.93 |
| At receptor locations | NRMSE (10) | $\sqrt{\text{NMSE}}$ (11) | Correlation (12) |
| Meta-model vs. full model chain | 0.22 | 0.18 | 0.96 |

over the same set of simulations. While the emulation of the projection coefficients is globally responsible for a significant portion of the error, we can see that the regions with the highest projection error correspond to high errors in the meta-model as well. This is expected, as the emulated solution can only perform as well as the projected solution. We see that larger errors are located on roads, mostly the large highway and outside the dense urban area. Meta-model error remains below 20% over a large portion of the domain, which shows that much of the spatial variation of the concentration is captured by the reduced order solution.
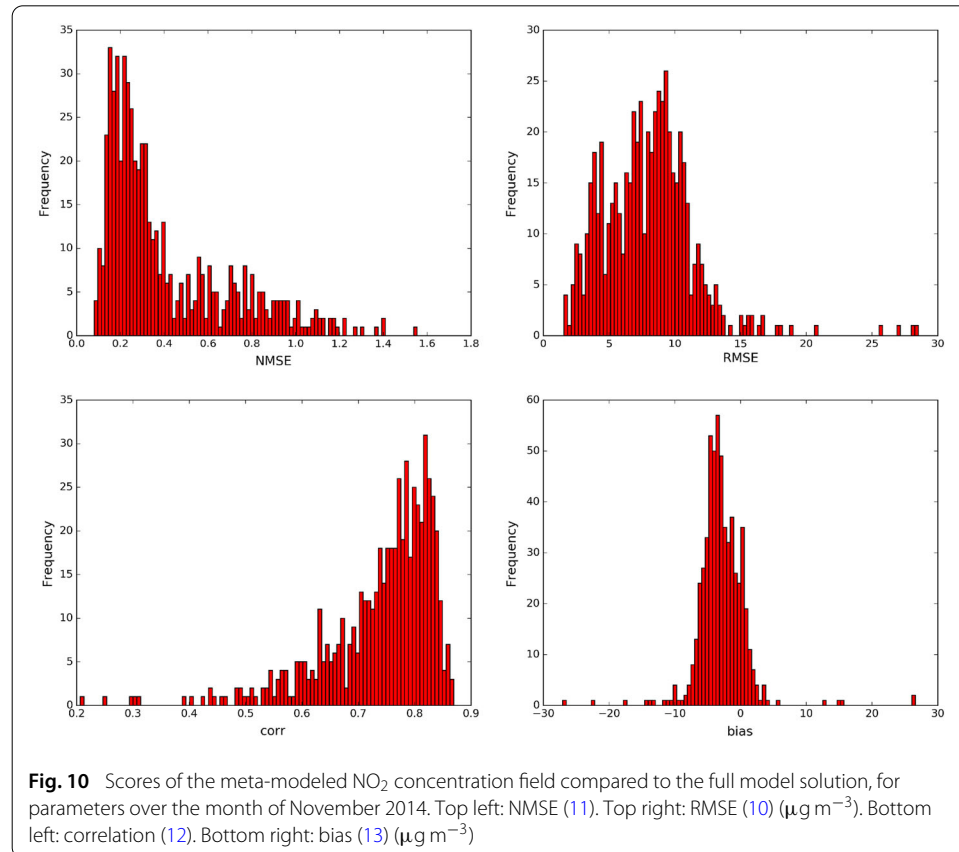
In Table 3 we can see statistical scores comparing the meta-modeled concentration to the full concentration model over all hours of November 2014. We compare both the entire grid (here $c_m$ is the concentration at a grid point and $M = N_{grid}$ is the total number of grid points) and at the $NO_2$ sensor locations. While the dimensional reduction means the meta-model does not fully capture spatial variations of the simulated concentration state, we can see that the relative RMSE errors are satisfactorily low, and the correlation between the two is very high.

In Table 4 we can see these scores when a reduced basis and metamodel are trained using a subset of 3000 members of the training set only. We can see the necessity of the larger training set for the air quality model. Here the difference in scores with respect to Table 3 is caused by the smaller training set of the emulation, rather than by a less precise reduced basis.

In Fig. 10, we see a visual representation of hourly scores of the meta-model solution compared to the full solution at each grid point for simulations corresponding to the month of November 2014. The NMSE (11) remains below 0.4 for most parameters, and the RMSE (10) often below $10\,\mu g\,m^{-3}$. Correlations scores are grouped above 0.75, and

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 16 of 22

**Table 4** Statistical scores of a meta-model approximation trained on $N_{train} = 3000$ compared to the scores of the model chain using the full air quality model

| State Estimation $c(\mathbf{p})$ vs. $\hat{c}(\mathbf{p})$ ($N_{train} = 3000$) | Statistical scores over 1-month simulations | | |
|---|---|---|---|
| Over the full grid | MNRMSE (14) | $\sqrt{\text{NMSE}}$ (11) | Correlation (12) |
| Meta-model vs. full model chain | 0.35 | 0.30 | 0.91 |
| At receptor locations | NRMSE (10) | $\sqrt{\text{NMSE}}$ (11) | Correlation (12) |
| Meta-model vs. full model chain | 0.30 | 0.24 | 0.94 |



**Fig. 10** Scores of the meta-modeled $NO_2$ concentration field compared to the full model solution, for parameters over the month of November 2014. Top left: NMSE (11). Top right: RMSE (10) ($\mu g\,m^{-3}$). Bottom left: correlation (12). Bottom right: bias (13) ($\mu g\,m^{-3}$)
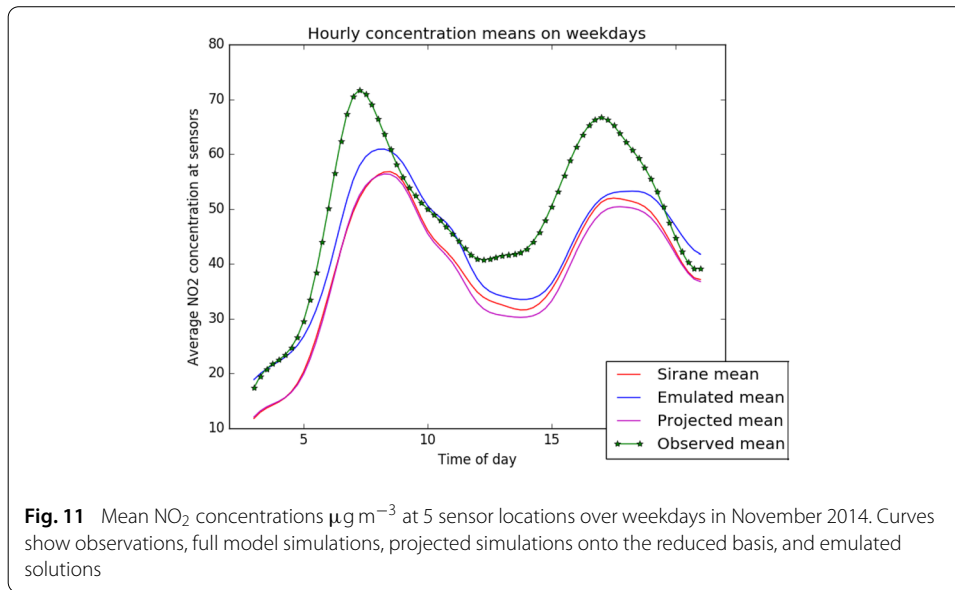
the bias distribution is nearly centered around $-2\,\mu g\,m^{-3}$, showing a slightly higher concentration approximation by the meta-model, when averaged over the grid.

### Comparison with observational data

We next analyze the accuracy of the full model and meta-model compared to observational data on $NO_2$ concentrations. Sensor locations can be seen in Fig. 7. In Fig. 11, we see the temporal profile of average $NO_2$ concentrations at $M = 4$ sensor locations: $\frac{1}{M}\sum_{m=1}^{M} c_m$.

We compare observed, emulated, projected and Sirane modeled concentrations of all weekdays in November 2014. We see that the bias in the modeled concentrations underestimating peak concentrations, notably during heavy traffic periods in the mornings and evenings. We also notice a seemingly delayed reaction of the model chain to the pollution increase during the evening peak hour. In [28], this delay was less evident, suggesting that factors such as the dispersion and reaction parametrizations in the AQ model or the aver-

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 17 of 22



**Fig. 11** Mean $NO_2$ concentrations $\mu g\, m^{-3}$ at 5 sensor locations over weekdays in November 2014. Curves show observations, full model simulations, projected simulations onto the reduced basis, and emulated solutions

**Table 5** Statistical scores of the meta-model approximation results compared to observation data

| Scores over 1-month simulations | Sirane solution $c(\mathbf{p})$ | | | | Meta-model solution $\hat{c}(\mathbf{p})$ | | | |
|---|---|---|---|---|---|---|---|---|
| | NRMSE | $\sqrt{\text{NMSE}}$ | Bias | Correlation | NRMSE | $\sqrt{\text{NMSE}}$ | Bias | Correlation |
| All 4 stations | 0.479 | 0.499 | 12.88 | 0.746 | 0.461 | 0.466 | 10.28 | 0.728 |
| Lecoq | 0.368 | 0.351 | 6.81 | 0.847 | 0.364 | 0.347 | 6.5 | 0.846 |
| Montferrand | 0.366 | 0.33 | 2.04 | 0.833 | 0.377 | 0.325 | −1.6 | 0.836 |
| Gare | 0.655 | 0.587 | 33.13 | 0.776 | 0.469 | 0.52 | 20.43 | 0.783 |
| Chamalières | 0.474 | 0.573 | 16.05 | 0.719 | 0.498 | 0.55 | 15.8 | 0.648 |

NRMSE (10), $\sqrt{\text{NMSE}}$ (11), Correlation (12), and bias (13) in $\mu g\, m^{-3}$

aging of time scales from 15 min to 1 h may have an effect. The exploration of this question will require more study of uncertainties in the model chain. We notice that the temporal trend representing morning and evening peak hours in traffic is reproduced by the model chain. We also note that the emulated concentrations are closer to the observations than the full model. This is likely due to the "smoothing" effect of the dimensional reduction causing less sharp concentration variations, as small parts of the modeled concentration fields are not reproduced by the reduced basis.

In Table 5, we compute statistical scores over the month of November 2014, comparing the full model simulations and the meta-modeled simulations to the observation data at $M = 4$ sensor locations. We again see that the emulated solutions are slightly more accurate than the full model. The stations at which both the model and meta-model perform best are those found in dense urban areas, excepting the station Gare, where heavy traffic induces high $NO_2$ concentrations, which the model fails to reproduce. We see the highest bias at this location. Finally, the station Chamalières is located outside the city center, where the model exhibits a higher level of bias. The performance of the meta-model with respect to observation data is highly satisfactory.

In Fig. 12, we see a visual representation of daily scores of the meta-model solution and the full solution compared to $NO_2$ observations over the month of November 2014. The meta-model shows similar score distributions to the full model, excepting the occasional

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 18 of 22



**Fig. 12** Scores of the Sirane NO$_2$ concentration field compared to the observation data over the month of November 2014. Top left: NMSE (11). Top right: RMSE (10). Bottom left: correlation (12). Bottom right: bias (13)
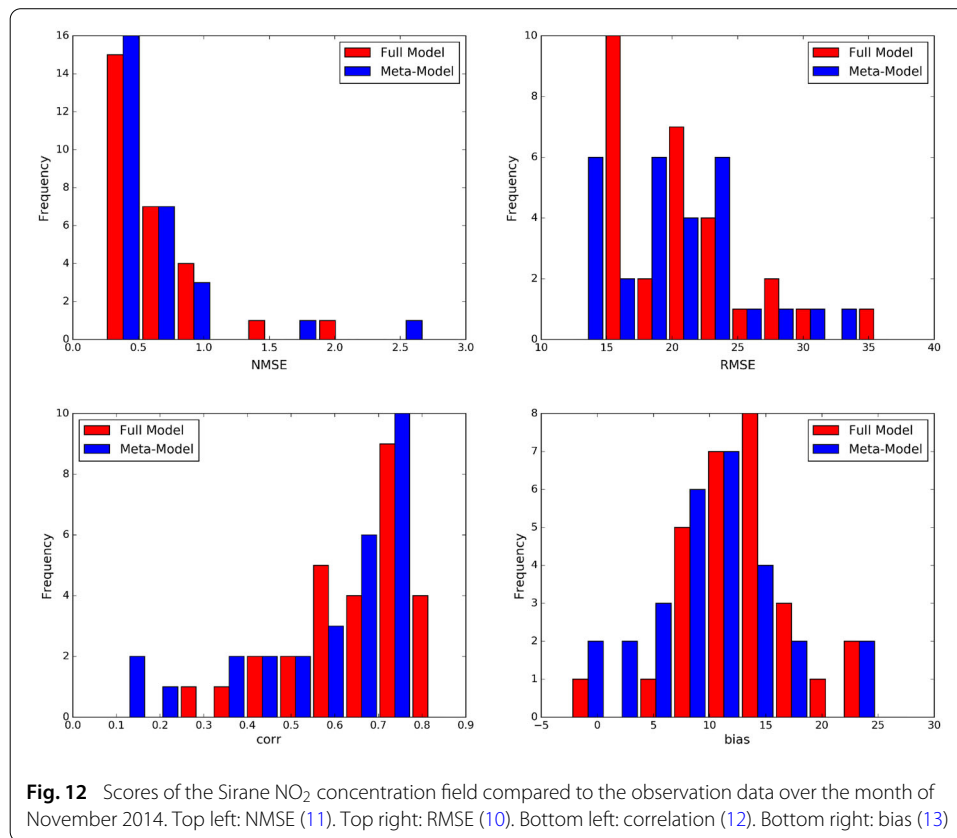
outlier. The RMSE (10) is below $25\,\mu\mathrm{g\,m^{-3}}$ on the majority of days for both the full and reduced simulations, with the bias distribution nearly centered around $10\text{--}15\,\mu\mathrm{g\,m^{-3}}$, showing an underestimation of concentrations by the simulations.

While we have seen that the model reduction by statistical emulation causes loss of precision, and the meta-model simulations contain error with respect to the full model, comparing to observation data suggests that this error is not significant with respect to the model error inherent to operational models for urban air quality, and does not reduce the accuracy of the predicted concentrations at sensor locations.

**Computational savings**

We have seen that the meta-model chain produces satisfactory results when compared to observational data, and determined that the loss of precision due to the dimensional reduction is not higher than the error committed by the full model. Now we will show the computational savings afforded by the meta-model chain. In Table 6, we can see the computational times required for a single simulation of the chain by the meta-models or the full models. The meta-models depend on three reduced bases, representing traffic for the traffic assignment meta-model, road emissions for the reduction of pollution model input dimension, and concentration fields for the pollution meta-model. The initialization of the meta-model chain requires loading these bases and building the RBF emulators. Once the chain is initialized, it can be run for any number of simulations at very low cost, under $0.1$ s for a simulation representing a 1-h period. In comparison, the full model chain requires nearly 3 h for a single simulation.

**Table 6 Computation times using the meta-model or full model chain**

| Computational times | | | | |
|---|---|---|---|---|
| Meta-model simulation | Traffic emissions | Compute $\hat{\alpha}^{lin}(\mathbf{p}_m)$ | Compute $\hat{\alpha}^{AQ}(\mathbf{p}_m)$ | Total CPU |
| Initialize meta-model chain | – | – | – | 24 min |
| Emulating $\alpha^{AQ}(\mathbf{p}_m)$ | 0.05 s | 0.006 s | 0.02 s | 0.076 s |
| | | | | |
| Full model simulation | Traffic emissions | – | Compute $c(\mathbf{p}_m)$ | Total CPU |
| Simulation of $c(\mathbf{p}_m)$ | 117 min | – | 23 min | 140 min |

The offline construction of the meta-models required 6000 traffic model simulations [28] and 10, 000 pollution model simulations, which represents a significant computational investment. However, these meta-models are trained over training points $\{\mathbf{p}_i\}_{1 \leq i \leq N_{train}} \in \mathcal{D}$ representing 2 years of data, and once constructed are useful for study over multiple years. In the absence of high performance computing machines or clusters, the simulations can be run using a pseudo-parallel technique running one simulation per core on desktop calculation machines. The Sirane simulations described in "Case study on Clermont-Ferrand" section took around one day using this method on multiple machines of 64 GB RAM or less. Once the meta-model chain is constructed, the online phase for the simulation given any parameter $\mathbf{p} \in \mathcal{D}$ is very cheap, which makes real-time or many-query contexts possible, for example for use in uncertainty quantification study.

## Conclusions

In this work we constructed a meta-model chain by statistical emulation of reduced basis projection coefficients for urban air quality modeling over the agglomeration of Clermont-Ferrand. We used the road traffic meta-model constructed in [28], built a reduced basis representing road traffic emissions, and constructed a second meta-model of $NO_2$ concentration fields over the agglomeration, substituting thus a low-cost chain of meta-models for a computationally costly modeling chain over a large urban area. This required the selection of a spatially finer road network for the AQM emissions inputs, the dimensional reduction of the inputs to the atmospheric pollution model, the treatment of traffic observation data to compute model input parameters, and the appropriate sampling of the parameter spaces to construct a reduced basis and reduced order modeling scheme. We chose the method of a meta-model chain, and restricted the variations in the AQM input parameters with respect to a standard LHS method without under-representing the solution space.

For each simulation of an hourly concentration field, we reduced computation time from over two computational hours to under 0.1 s. Results show good precision of the meta-model simulations with respect to the full model chain, and similar accuracy when compared to measurement data. We saw that a portion of the error between the meta-model and full-model chains can be attributed to model error, and the reduced order model does not show significantly increased error. The meta-model can be used in applications requiring numerous solutions to the model chain, rendering various otherwise impractical studies, for example exposure analysis, computationally feasible. This model reduction makes the model chain useful in a wide variety of applications.

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.* (2020)7:37

Page 20 of 22

Here we constructed a chain of meta-models as opposed to a single meta-model of the full modeling chain. This was done in order to make use of traffic and emissions simulations and data, and for the versatility of a chain of meta-models in our applications. A comparison of the precision, parameter sensitivity, as well as the stability of the meta-model formulation (inversion of matrices) of each method, would make for an interesting follow-up study. In future work, we will use this low-cost meta-modeling chain in the study of uncertainty quantification and the propagation of uncertainties throughout the model chain.

### Authors' contributions
RC constructed a meta-model of the dynamic traffic model and reduced database of COPERT IV emissions, and provided these programs. JKH implemented the air quality simulations, constructed reduced basis of traffic emissions and pollutant concentrations, constructed the chain of meta-models and analyzed the results using pollutant concentration data, and drafted the manuscripts. VM provided methodological guidance throughout the study and contributed to the revision of the manuscript. All authors read and approved the final manuscript.

### Availability of data and materials
The City of Clermont-Ferrand provided traffic flow measurement data. SMTC (Syndicat Mixte des Transports en Commun de l'agglomération clermontoise) provided the traffic network geometry for the agglomeration of Clermont-Ferrand and the static O-D matrix representing spatial traffic demand in the traffic assignment model. The SME NUMTECH provided data necessary for the model SIRANE, including geometrical features of the traffic network, meteorological data, emissions data, and background concentrations. Atmo Auvergne-Rhone-Alpes (https://www.atmo-auvergnerhonealpes.fr/) provided pollutant concentration measurements used in the calculation of statistical scores.

### Competing interests
The authors declare that they have no competing interests.

### Author details
[1]INRIA, 2 Rue Simone IFF, 75012 Paris, France, [2]CEREA, Joint Laboratory of Ecole des Ponts ParisTech - EDF R&D, Université Paris-Est, 6-8, Avenue Blaise-Pascal, Citéé Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée CEDEX 2, France.

### References
1. World-Health-Organization. Ambient air pollution: a global assessment of exposure and burden of disease. Tech rep. 2016. http://www.who.int/phe/publications/air-pollution-global-assessment/en/.
2. Milliez M, Carissimo B. Computational fluid dynamical modelling of concentration fluctuations in an idealized urban area. Boundary-Layer Meteorol. 2008;127(2):241–59. https://doi.org/10.1007/s10546-008-9266-1.
3. Tominaga Y, Stathopoulos T. CFD simulation of near-field pollutant dispersion in the urban environment: a review of current modeling techniques. Atmos Environ. 2013;79:716–30. https://doi.org/10.1016/j.atmosenv.2013.07.028.
4. Mallet V, Tilloy A, Poulet D, Girard S, Brocheton F. Meta-modeling of ADMS-Urban by dimension reduction and emulation. Atmos Environ. 2018;184:37–46. https://doi.org/10.1016/j.atmosenv.2018.04.009.
5. Carruthers DJ, Edmunds HA, McHugh CA, Singles RJ. Development of ADMS-urban and comparison with data for urban areas in the UK. In: Gryning SE, Chaumerliac N, editors. Air pollution modeling and its application XII. Berlin: Springer; 1998. p. 467–75.
6. Lee LA, Carslaw KS, Pringle KJ, Mann GW, Spracklen DV. Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters. Atmos Chem Phys. 2011;11(23):12253–73. https://doi.org/10.5194/acp-11-12253-2011.
7. Armand P, Brocheton F, Poulet D, Vendel F, Dubourg V, Yalamas T. Probabilistic safety analysis for urgent situations following the accidental release of a pollutant in the atmosphere. Atmos Environ. 2014;96:1–10. https://doi.org/10.1016/j.atmosenv.2014.07.022.

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 21 of 22

8. Girard S, Mallet V, Korsakissok I, Mathieu A. Emulation and Sobol' sensitivity analysis of an atmospheric dispersion model applied to the Fukushima nuclear accident. J Geophys Res Atmos. 2016;121(7):3484–96.

9. Fallah Shorshani M, André M, Bonhomme C, Seigneur C. Modelling chain for the effect of road traffic on air and water quality: techniques, current status and future prospects. Environ Modell Softw. 2015;64:102–23. https://doi.org/10.1016/j.envsoft.2014.11.020.

10. Russell A, Dennis R. NARSTO critical review of photochemical models and modeling. Atmos Environ. 2000;34(12):2283–324.

11. Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A. Real-time air quality forecasting, part II: state of the science, current research needs, and future prospects. Atmos Environ. 2012;60:656–76. https://doi.org/10.1016/j.atmosenv.2012.02.041.

12. Leurent F. On network assignment and demand-supply equilibrium: an analysis framework and a simple dynamic model. In: Proceedings of the European transport conference (ETC) 2003 held 8–10 October 2003, STRASBOURG, FRANCE. 2003.

13. Leurent F, Aguiléra V. Large problems of dynamic network assignment and traffic equilibrium: computational principles and application to Paris road network. Transp Res Rec. 2009;2132(1):122–32.

14. Chen R, Mallet V. Pollemission software computing traffic emissions of atmospheric pollutants with copert-iv formulations. https://github.com/pollemission.

15. Ntziachristos L, Gkatzoflias D, Kouridis C, Samaras Z. COPERT: a European road transport emission inventory model. In: Athanasiadis DIN, Rizzoli PAE, Mitkas PA, Gómez PD-IJM, editors. Information technologies in environmental engineering, environmental science and engineering. Berlin: Springer; 2009. p. 491–504. https://doi.org/10.1007/978-3-540-88351-7_37.

16. Soulhac L, Salizzoni P, Cierco F-X, Perkins R. The model SIRANE for atmospheric urban pollutant dispersion; part I, presentation of the model. Atmos Environ. 2011;45(39):7379–95. https://doi.org/10.1016/j.atmosenv.2011.07.008.

17. Prud'homme C, Rovas DV, Veroy K, Machiels L, Maday Y, Patera AT, Turinici G. Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. J Fluids Eng. 2002;124(1): 70–80. http://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1429475.

18. Chen R. Uncertainty quantification in the simulation of road traffic and associated atmospheric emissions in a metropolitan area. Thesis, Paris Est (May 2018). http://www.theses.fr/2018PESC1029.

19. CHAKIR R, Joly P, Maday Y, Parnaudeau P. A non intrusive reduced basis method: application to computational fluid dynamics. In: 2nd ECCOMAS young investigators conference (YIC 2013), Bordeaux, France. 2013. https://hal.archives-ouvertes.fr/hal-00855906.

20. Chakir R, Hammond JK. A non-intrusive reduced basis method for elastoplasticity problems in geotechnics. J Comput Appl Math. 2018;337:1–17. https://doi.org/10.1016/j.cam.2017.12.044.

21. Hammond J, Chakir R, Bourquin F, Maday Y. PBDW: a non-intrusive Reduced Basis Data Assimilation method and its application to an urban dispersion modeling framework. Appl Math Modell. 2019;76:1–25. https://doi.org/10.1016/j.apm.2019.05.012.

22. Wang Q, Hesthaven JS, Ray D. Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem. J Comput Phys. 2019;384:289–307. https://doi.org/10.1016/j.jcp.2019.01.031.

23. Demo N, Tezzele M, Rozza G. A non-intrusive approach for the reconstruction of POD modal coefficients through active subspaces. C R Mécanique. 2019;347(11):873–81. https://doi.org/10.1016/j.crme.2019.11.012.

24. Bui-Thanh T, Damodaran M, Willcox K. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. AIAA J. 2004;42(8):1505–16. https://doi.org/10.2514/1.2159.

25. Quarteroni A, Manzoni A, Negri F. Reduced basis methods for partial differential equations: an introduction. Vol. 92. Springer. 2015. https://books.google.fr/books?hl=en&lr=&id=e6FnCgAAQBAJ&oi=fnd&pg=PP1&dq=quarteroni+manzoni+negri+reduced+basis+methods+for+partial+differential+equations+springer&ots=jVBOqponFX&sig=D_VExRlkwaJbPLh08oVWs7tyquU.

26. Hesthaven JS, Rozza G, Stamm B. Certified reduced basis methods for parametrized partial differential equations., Springerbriefs in mathematicsBerlin: Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-22470-1.

27. Kolmogoroff A. Uber die beste Annaherung von Funktionen einer gegebenen Funktionenklasse, Ann Math. 1936;37:107–10. http://www.jstor.org/stable/1968691.

28. Chen R, Mallet V, Aguilera V, Cohn F, Poulet D. Metamodeling of a dynamic traffic assignment model at metropolitan scale, 43.

29. Ruiwei C, Vivien M. Pollemission software computing traffic emissions of atmospheric pollutants with COPERT-IV formulations, original-date: 2016-01-21T17:19:00Z. 2016. https://github.com/pollemission.

30. Gkatzoflias D, Kouridis C, Ntziachristos L, Samaras Z. COPERT 4: Computer programme to calculate emissions from road transport. Copenhagen: European Environment Agency; 2009.

31. EEA. EMEP/EEA air pollutant emission inventory guidebook—Part B.1.A.3.b.iiv Road transport. 2016. https://www.eea.europa.eu/publications/emep-eea-guidebook-2016.

32. Carslaw DC. Evidence of an increasing NO2/NOX emissions ratio from road traffic emissions. Atmos Environ. 2005;39(26):4793–802. https://doi.org/10.1016/j.atmosenv.2005.06.023.

33. Beevers SD, Westmoreland E, de Jong MC, Williams ML, Carslaw DC. Trends in NOx and NO2 emissions from road traffic in Great Britain. Atmos Environ. 2012;54:107–16. https://doi.org/10.1016/j.atmosenv.2012.02.028.

34. Kurtenbach R, Kleffmann J, Niedojadlo A, Wiesen P. Primary NO2 emissions and their impact on air quality in traffic environments in Germany. Environ Sci Eur. 2012;24(1):21. https://doi.org/10.1186/2190-4715-24-21.

35. Gillies JA, Gertler AW, Sagebiel JC, Dippel WA. On-road particulate matter (PM2.5 and PM10) Emissions in the Sepulveda Tunnel, Los Angeles, California. Environ Sci Technol. 2001;35(6):1054–63. https://doi.org/10.1021/es991320p.

Hammond *et al. Adv. Model. and Simul. in Eng. Sci.*(2020)7:37

Page 22 of 22

36. Querol X, Alastuey A, Ruiz CR, Artiñano B, Hansson HC, Harrison RM, Buringh E, ten Brink HM, Lutz M, Bruckmann P, Straehl P, Schneider J. Speciation and origin of PM10 and PM2.5 in selected European cities. Atmos Environ. 2004;38(38):6547–55. https://doi.org/10.1016/j.atmosenv.2004.08.037.

37. Soulhac L, Salizzoni P, Mejean P, Didier D, Rios I. The model SIRANE for atmospheric urban pollutant dispersion. PART II, validation of the model on a real case study. Atmos Environ. 2011;49:320–37. https://doi.org/10.1016/j.atmosenv.2011.11.031.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.