20th International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
14-18 June 2021, Tartu, Estonia

COMPOUND PARAMETRIC METAMODELLING OF LARGE-EDDY SIMULATIONS FOR MICROSCALE ATMOSPHERIC DISPERSION

Bastien X. Nony[1], Mélanie C. Rochoux[1], Didier Lucor[2] and Thomas Jaravel[3]

[1] CECI, Université de Toulouse, CNRS, CERFACS, 42 Avenue Gaspard Coriolis, 31100 Toulouse, France
[2] LISN, CNRS, Université Paris-Saclay, Campus universitaire Bâtiment 508, Rue John von Neumann, 91405 Orsay, France
[3] CERFACS, 42 Avenue Gaspard Coriolis, 31100 Toulouse, France

Abstract: In pollutant dispersion problems, mapping concentrations in the first tens or hundreds of meters from the source still remains a modelling challenge. Large-eddy simulations (LES) are able to represent time and space variability of turbulent atmospheric flow, which is of prime importance to assess public short-term exposure. However, they remain far from real time and subject to uncertainties, in particular to parametric uncertainties associated with the large-scale atmospheric forcing and the emission source position. In this work, we show that an efficient and accurate metamodel of the tracer concentration information provided by LES and encapsulating their associated uncertainties can be built using appropriate statistical tools combining machine learning and principal component analysis. We present a proof-of-concept study based on a simplified but representative flow configuration (two-dimensional flow around a surface-mounted cube) using the AVBP LES solver and testing a variety of metamodels (linear regression, Gaussian processes, random forest, gradient boosting, etc.). Results reinforce the idea that for sufficiently statistically-converged quantities of interest and for a sufficiently large LES data set, a compound surrogate model can succeed in synthesizing information from the LES in the whole computational domain (with a Q^2 predictivity coefficient above 90 %). Downstream of the obstacle, the Q^2 coefficient of all metamodels reaches excellent results over 90%. Upstream, the tracer concentration is subject to strong discontinuities; combining metamodels allows to guarantee a good predictivity coefficient over 75%.

Key words: *Computational Fluid Dynamics, Large-Eddy Simulation, Atmospheric Dispersion, Microscale, Surrogate Models, Parametric Uncertainty*

INTRODUCTION

Industrial accidents often involve a pollutant plume dispersion potentially harmful to human health, economy and/or environment. A new generation of decision support tools at the interface between modeling and statistical learning could be designed to help monitoring an emergency by accurately simulating the plume dispersion and accounting for a range of possible scenarios at future time frames.

Near-field plume dispersion is controlled by complex turbulent flow dynamics enhanced by complex geometry, for example in urban areas where separation and recirculation zones are induced by the presence of buildings of different height and geometry. Computational Fluid Dynamics (CFD) approaches, based on Reynolds-averaged Navier Stokes (RANS) and Large-Eddy Simulation (LES), is now becoming popular to accurately represent these processes (Philips *et al.,* 2013). However, CFD approaches remain computationally intensive and therefore far from real time. Replacing the CFD model by a metamodel (*i.e.* a statistical model that has learnt the relationship between uncertain inputs and some quantities of interest over a CFD data set) can partly overcome this limitation. García-Sánchez *et al.* (2017) have demonstrated the capacity to train a metamodel based on polynomial chaos expansion, which mimics the RANS model response for a wide range of parametric variations on three uncertain parameters: the upstream roughness height, the wind direction and magnitude. Once the training step is done using a RANS data set, the resulting metamodel can be used to estimate at no further cost, the metamodel model response for a new set of parameters. This was illustrated for the "Joint Urban 2003 field experiment" in Oklahoma City, which is fully representative of urban geometry complexity.
The objective of this work is to provide a detailed comparison of metamodels to identify which are the most relevant for pollutant dispersion and in particular if polynomial multi-linear regression approaches as in

García-Sánchez *et al.* (2017) are sufficient to reproduce the CFD response to parametric variations. We consider here a two-dimensional canonical case (*i.e.* a two-dimensional flow around a surface-mounted cube) to compare a variety of metamodels, evaluate their performance and assess their robustness with respect to the training set size and data noise (associated with the time-average process). The number of uncertain parameters is kept low: three scalar parameters are considered: the inlet wind speed and the emission source position. The quantity of interest is the time-averaged tracer concentration affected by unsteady flow features, such as vortex shedding. The numerical data set is built using the AVBP LES solver (http://www.cerfacs.fr/avbp7x/). In the present study, the objective is to design a robust and accurate metamodel to represent the AVBP response to variations in the three uncertain parameters. This paper is structured as follows: *i)* the numerical case study and setup are introduced; *ii)* the metamodelling strategy is presented; and *iii)* the metamodel results are analysed.

NUMERICAL CASE STUDY AND SETUP

The test case is representative of a passive tracer dispersion in the lower part of the atmospheric boundary layer, which is modelled as an open two-dimensional computational domain in the reference frame (x, z) (Figure 1). A single obstacle with side length $H = 1$ m is considered. The domain consists of a rectangle $31H$ long (inlet flow direction x) by $10H$ high (vertical direction z). The domain height of $10H$ was chosen to avoid upper boundary effects.

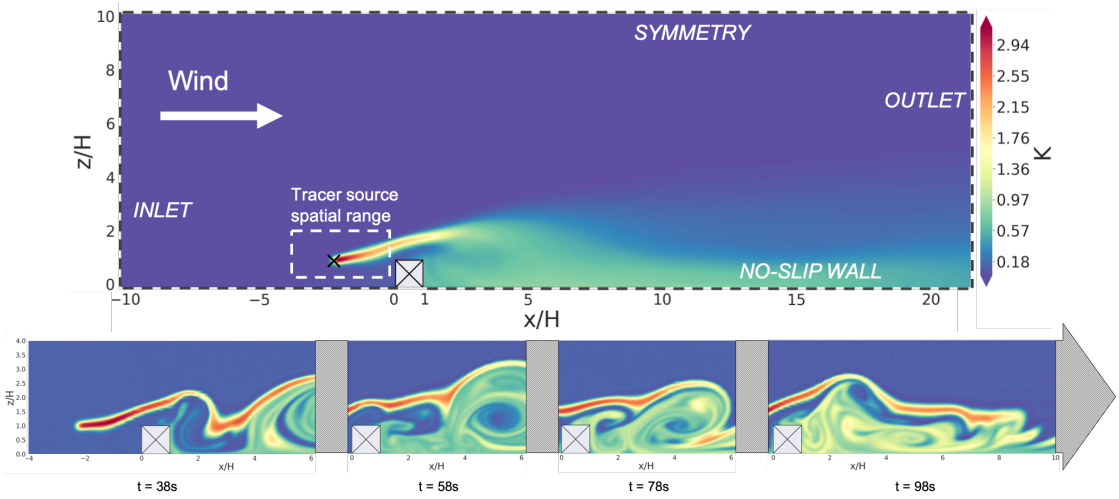

Figure 1. Scheme of the two-dimensional computational domain with boundary conditions. Example of a mean tracer concentration field is given in the background for a source centered at (-2.26, 0.98) and an inlet wind speed of 5.58 m s^{-1}. Associated instantaneous snapshots of the tracer concentration fields are also given.

A steady and uniform air flow is injected by the inlet boundary. The inlet wind speed is uncertain, with $U_{inlet} \in [1,10]$ m s^{-1}. The incoming flow is perturbed by the presence of the obstacle, inducing unsteady flow features, primarily vortex shedding. A tracer emission point-source (propylene here) is positioned upstream of the obstacle with a constant emission rate. The passive tracer is numerically injected in the form of a Gaussian function. The Gaussian source center is located upstream of the obstacle and is uncertain, with $x_{src} \in [-3.5, -0.2]H$ horizontally and $z_{src} \in [0.2, 2]H$ vertically.

AVBP is used to solve the compressible Navier-Stokes equations with an artificial compressibility approach and with a fine mesh grid made of 240,000 triangular elements. A uniform spatial resolution of $0.04H$ (4 cm) was adopted to solve the flow around the obstacle with 25 grid points in each direction. A third-order accurate advection of the Taylor-Galerkin family is used. NSCBC (Navier-Stokes Characteristic Boundary Conditions) are used to properly handle acoustics at the inlet and outlet. The upper boundary models an atmospheric free surface through a plane symmetry condition. The lower boundary (the ground surface and the obstacle) is modelled as an adiabatic wall with a zero velocity. A constant pressure condition is imposed at the outlet.

METAMODELLING STRATEGY

The objective in this work is to design a metamodel able to predict the mean (time-averaged) tracer concentration field (containing N_{nodes} mesh nodes) as a function of the three uncertain parameters:

$$M : [1,10]\times[-3.5,-0.2]\times[0.2,2] \rightarrow \mathbb{R}_{+}^{N_{nodes}} \quad (U_{inlet}, x_{src}, y_{src}) \rightarrow (K_i)_{i\in\{1,\dots,N_{nodes}\}} \tag{1}$$

M is referred to as the response surface where the quantities of interest are the scaled concentration coefficients at different mesh nodes obtained from the uncertain parameters. The scaled concentration is defined by $K = c/c_0$ with c the time-averaged concentration and c_0 the normalizing concentration:

$$c_0 = \frac{\rho_0 \times Q}{\bar{U} \times H^2} \begin{cases} \rho_0, \text{ air density} \\ Q, \text{ volume flow of tracer source} \\ \bar{U}, \text{ ensemble mean inlet wind velocity} \end{cases} \tag{2}$$

Only the nodes where the ensemble mean tracer concentration is above $K \geq 4.59 \times 10^{-2}$ (10^{-2} kg m^{-3}) inside the box $[-4, 10]H \times [0, 4]H$ are considered in the metamodeling procedure, resulting in $N_{nodes} \approx 16{,}000$. Outside this box, concentrations are very small, falling below the solver numerical accuracy, and are thus discarded for metamodelling.

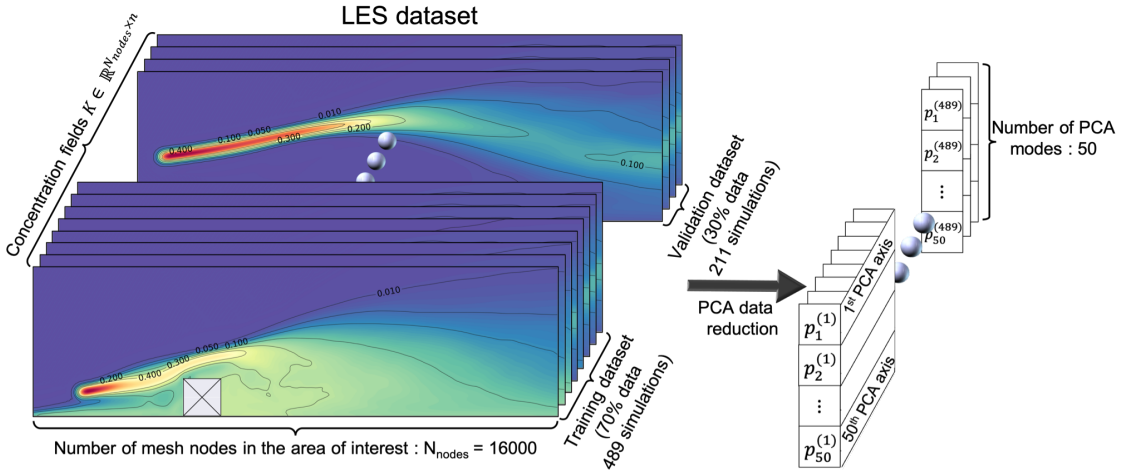

Figure 2. Scheme of the LES data set generation and PCA postprocessing before metamodeling.

To perform a detailed analysis of the metamodel performance, a very large ensemble of LES (n = 700) is generated. The response surface is sampled using a quasi Monte-Carlo approach, assuming uniform statistical distributions for the three parameters given the range of variations in Equation 1. A low-discrepancy Halton's sequence is adopted in this study to obtain a homogeneous sampling of the low-dimensional parameter space. AVBP is run for each sampled set of parameters. The ensemble is segmented into a training set (70% of the LES, *i.e.* 489) and a validation set (30% of the LES, *i.e.* 211). Each resulting mean tracer concentration field K is decomposed using Principal Component Analysis (PCA) to reduce the output dimension from $N_{nodes} \approx 16{,}000$ to 50 principal components (Figure 2). Tests have shown that this is enough to accurately represent the tracer concentration over the whole domain. To limit the computational cost related to the LES data set generation, the AVBP simulation time is adjusted according to the inlet flow velocity U_{inlet}. Higher inlet velocities imply faster asymptotic convergence of the mean tracer concentration. A LES lasts from 150 s to 1,600 s in physical time to reach approximately a fixed number of 30 vortex shedding based on an estimated Strouhal value of S_t = 0.02. Using this approach, running 100 LES requires about 40,000 CPU hours on CERFACS' supercomputer.

Different regression metamodels are trained to approximate the response surface in Equation 1 (via the SciKit Learn library) using the LES training data set: multi-linear regression with and without penalty (Ridge, LASSO/Least Absolute Shrinkage and Selection Operator, OMP/Orthogonal Matching Pursuit), decision trees (random forest and gradient boosting), Gaussian processes. All of them are put in competition to represent the relation between each PCA component and the three uncertain parameters. In the resulting

compound metamodel, we only keep a unique "champion" metamodel per PCA axis, i.e. the one that achieves the best performance according to the Q^2 predictivity coefficient (Figure 3):

$$Q^2 = 1 - \frac{||p_i(K) - p_i(\widehat{K})||^2}{Var(\, p_i(K)\,)} \tag{3}$$

with $p_i(K)$ and $p_i(\widehat{K})$ the projected components on the ith PCA axis of the AVBP prediction K and a given metamodel prediction $\widehat{K}$, respectively, over the validation data set. The Q^2 coefficient gives the percentage of the variance explained by the metamodel (a perfect metamodel features $Q^2 = 1$; in practice, a Q^2 coefficient above 90% is considered as excellent). In this work, the performance of the compound and standalone metamodels is quantified using the Q^2 metrics on tracer concentrations at the N_{nodes} nodes over the validation data set. It is therefore a spatial field. To help the analysis of the results, different Q^2 statistics are computed: the spatially-average value of the Q^2 coefficient over the computational domain (the global Q^2) but also over limited-area regions (upstream, around the obstacle and downstream).

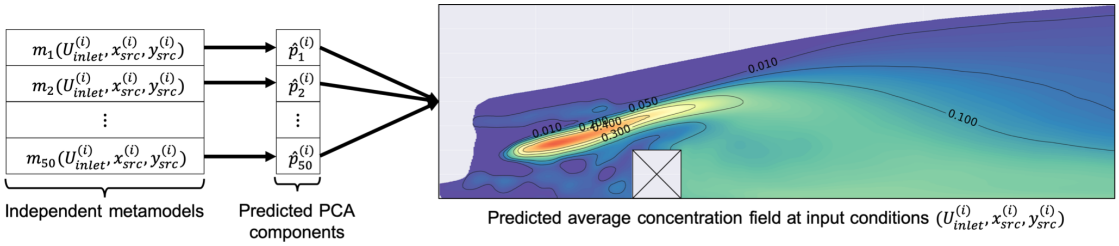

Figure 3. Scheme of the metamodeling procedure.

RESULTS

Figure 4 presents the spatial field of the Q^2 coefficient obtained for the compound metamodel. Table 1 presents the associated Q^2 statistics per limited-area domain. Results show a good performance of the compound metamodel in the whole computational domain, except in some upstream areas where the Q^2 is locally very low. This is consistent with a global Q^2 coefficient higher than 94%. The best Q^2 performance is obtained in the area downstream (the downstream Q^2 statistics is above 97%), while the Q^2 statistics decreases when moving upstream (the upstream Q^2 statistics remains above 80%). The dispersion process occurring downstream makes the mean tracer concentration field smoother for all sets of uncertain parameters. This is easier to capture by the tested metamodels than upstream of the obstacle, where the mean tracer concentration fields locally feature high values in distinct areas that depend on the tracer emission source location. Because only few realizations are considered in this area in the validation data set, the associated variance is low, prediction errors are penalizing, resulting in a poor Q^2 performance.

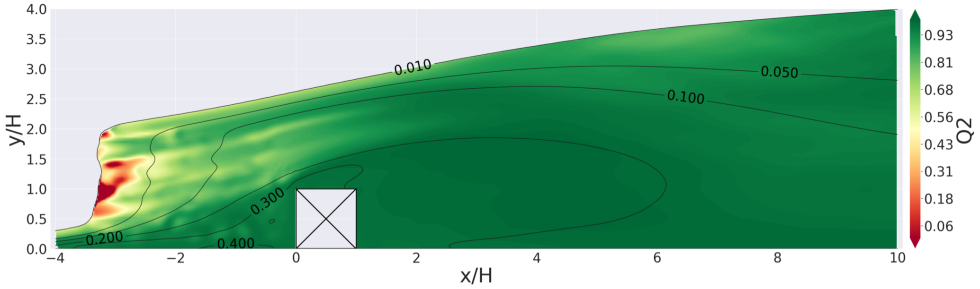

Figure 4. Spatial field of the Q^2 predictivity coefficient (Equation 3) using the compound metamodel.

Table 1 highlights the added value of the compound metamodel compared to standalone metamodels (when all PCA components are predicted using the same metamodel). The compound metamodel obtains the best performance for all Q^2 statistics. There is no much difference in the downstream area (where all metamodels reach a Q^2 statistics above 97%). However, having a compound metamodel limits the drop in performance in the upstream area: linear regression without penalty and Gaussian process regression feature an upstream Q^2 statistics equal or lower than 60%, while this statistics remains above 80% for the compound metamodel (gradient boosting follows closely at 78%). For the 90% quantile of concentration maximum residual errors, the compound metamodel value is $K = 3.94$, which is better than gradient boosting (4.49), linear (5.64) and Gaussian process regressors (6.60). Table 1 also shows that the performance of the compound metamodel

remains the best among all tested metamodels and relatively good when reducing the size of the training set (from 489 to 100 LES) or when introducing noise in the training set (by dividing by two the AVBP simulation time). There are now more differences between the performance of the compound metamodel and that of the gradient boosting metamodel, especially in the upstream area. The compound metamodel good behaviour is explained by its flexibility in adjusting to PCA components that are difficult to predict. The PCA components 1 to 4 are best predicted using Gaussian process regression, while the PCA components 5 to 11 are better predicted by gradient boosting. This implies that Gaussian processes better represent large-scale tracer concentration structures, while gradient boosting is more adapted to track sharp concentration gradients near the tracer emission source position. Figure 5 gives an example of the compound metamodel solution for a validation case. The predicted mean tracer concentration is compared to the AVBP solution. Downstream plume dispersion is well predicted. The upstream tracer concentration is slightly under-predicted in the wake of the emission source. There is also some noise due to accumulated prediction errors on the high-level PCA components upstream.

Table 1. Comparison of the metamodel Q^2 statistics for a validation data set made of 211 LES. In brackets are indicated the Q^2 statistics obtained when dividing by two the AVBP time window, in square brackets the statistics obtained when the training data set is reduced from 489 to 100 LES.

Metamodel/Q^2	**Global Q^2**	**Upstream Q^2** $x \in [-4, -0.5]H$	**Obstacle Q^2** $x \in [-0.5, 1.5]H$	**Downstream Q^2** $x \in [1.5, 10]H$
Compound	94% (93%) [85%]	80% (78%) [62%]	94% (93%) [80%]	97% (96%) [91%]
Linear regression (without penalty)	90% (89%) [66%]	60% (60%) [-43%]	83% (82%) [44%]	97% (96%) [93%]
Gradient boosting	93% (92%) [80%]	78% (78%) [52%]	92% (92%) [66%]	97% (95%) [88%]
Gaussian process regression	87% (86%) [2%]	41% (40%) [-320%]	81% (81%) [-37%]	98% (96%) [80%]

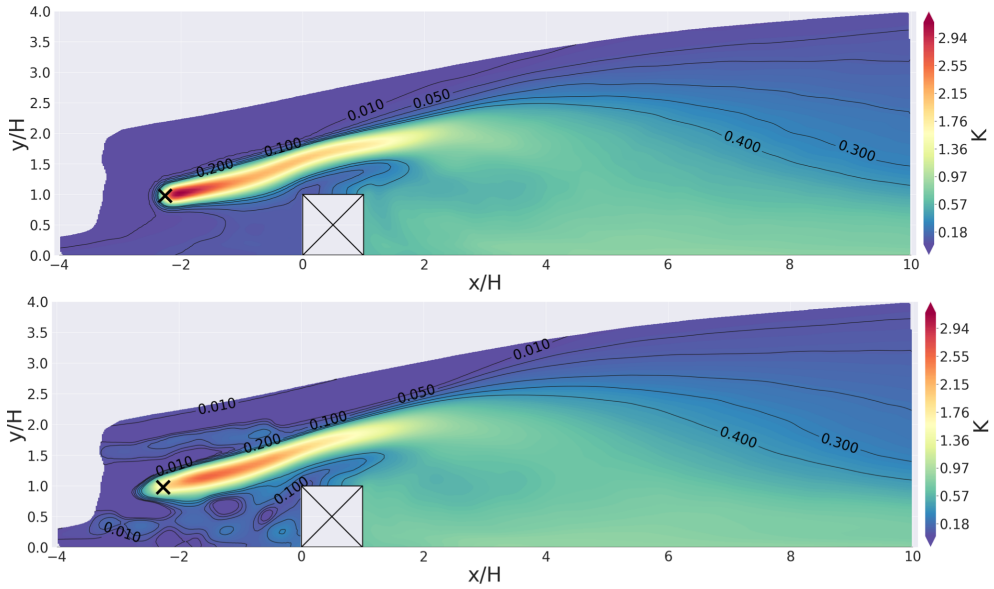

Figure 5. Mean tracer concentration field for U_{inlet} = 5.58, x_{src} = -2.26, y_{src} = 0.98 for (top panel) the AVBP solution and (bottom panel) the compound metamodel prediction.

CONCLUSION

The compound metamodel approach has shown its ability to reconstruct time-averaged information from LES on a simplified pollutant dispersion case. Mixing different metamodels (Gaussian process and gradient boosting) allows more PCA components to be considered upstream of the obstacle to improve the metamodel prediction and robustness to the training set size and noise. Future work includes adding a temporal dimension to fully take advantage of the LES, and applying it to more realistic cases.

REFERENCES

García-Sánchez, C., Tendeloo, G.V. and C. Gorlé, 2017: Quantifying inflow uncertainties in RANS simulations of urban pollutant dispersion. *Atmos. Environ.*, **161**, 263-273.

Philips, D.A., Rossi, R. and G. Iaccarino, 2013: Large-eddy simulation of passive scalar dispersion in an urban-like geometry. *J. Fluid Mech.*, **723**, 404-428.