REDUCED-ORDER MODEL FOR MICROSCALE ATMOSPHERIC DISPERSION COMBINING MULTI-FIDELITY LES AND RANS DATA

Bastien X Nony, Mélanie C. Rochoux, Thomas Jaravel, Didier Lucor

UNCECOMP 2023
5th ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Athens, Greece, 12—14 June 2023

REDUCED-ORDER MODEL FOR MICROSCALE ATMOSPHERIC DISPERSION COMBINING MULTI-FIDELITY LES AND RANS DATA

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

[1]CECI, Université de Toulouse, CNRS, CERFACS
42 Avenue Gaspard Coriolis, 31100 Toulouse, France
e-mail: bastien.nony,melanie.rochoux,thomas.jaravel@cerfacs.fr

[2] Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
91400 Orsay, France
e-mail: didier.lucor@lisn.upsaclay.fr

Keywords: Multi-fidelity, Co-kriging, Proper Orthogonal Decomposition, Convolutional Autoencoder, Atmospheric Boundary-Layer Flows, Computational Fluid Dynamics

Abstract. *Because of the complex turbulent flow dynamics caused by interactions between the atmospheric surface layer and the urban topography, forecasting microscale pollutant concentration fields is crucial for monitoring plume dispersion in urban regions. To capture these dynamics, large-eddy simulation (LES) is regarded as a high-fidelity numerical approach, however, it lacks real-time capabilities and remains costly because highly dimensional. Reynolds-averaged Navier-Stokes (RANS) approach is a cruder path to modelling these phenomena based on a statistical description of turbulent phenomena, leading to averaged transport equations. It is much less computationally expensive but also less accurate in representing interactions between the atmospheric boundary-layer flow and the buildings in urban areas. Designing an efficient reduced-order model (ROM) with a level of precision similar to the LES approach and accounting for atmospheric and point-source emission uncertainties is of primary importance. In this paper, we propose a novel multi-fidelity ROM, which combines two levels of data fidelity - LES and RANS - and we evaluate it in the context of atmospheric flow dispersion. RANS data are obtained by injecting detailed flow information from LES into a lower-fidelity tracer transport equation, in the RANS formalism, using physics-constrained machine learning. There are two steps in our approach: i) dimension reduction is performed using a convolutional autoencoder trained using pre-training, and ii) mapping the uncertain parameters to the autoencoder latent space is obtained using co-kriging and an autoregressive model. We consider a case of a simplified two-dimensional flow configuration around a surface-mounted obstacle, where the main physical quantity of interest is the time-average tracer concentration field and its variability with respect to uncertain atmospheric forcing and emission source location. We show that the multi-fidelity approach achieves increased performance compared to the LES and RANS single-fidelity approaches at equivalent computational budget. This is a promising approach to designing an efficient ROM applicable to realistic field-scale atmospheric dispersion cases.*

1 INTRODUCTION

Accidental short-term pollutant emissions occurring in urban areas or industrial sites can significantly degrade air quality and pose a risk to public health (*e.g.* 2011 Fukushima power plant explosion –Tsuruta et al., 2014; 2019/2020 Australian bushfires – Graham et al., 2021). The ability to accurately predict pollutant dispersion in urban areas is essential for making informed decisions in emergency situations (Da Silva et al., 2021; Mendil et al., 2022). Unfortunately, urban areas, due to highly specific topographies and complex interactions with meteorology, present unique challenges when attempting to predict the range of possible scenarios for near-source air pollutant dispersion (Philips et al., 2013; Dauxois et al., 2021).

Computational fluid dynamics has emerged as a powerful tool to accurately model the complex microscale atmospheric dispersion processes in urban environments (Tominaga and Stathopoulos, 2013; Dauxois et al., 2021), using both Reynolds-averaged Navier-Stokes (RANS) approaches (Milliez and Carissimo, 2007; García-Sanchez et al., 2017) and large-eddy simulations (LES) (Philips et al., 2013; Vervecken et al., 2015; García-Sanchez et al., 2018; Grylls et al., 2019). LES is of high interest due to its greater accuracy to represent flow/obstacle interactions and its ability to provide not only mean statistics but also higher-order statistics. However, this comes with a high computational cost, making it challenging to use in uncertainty quantification applications to account for uncertainties associated with large-scale atmospheric conditions (Sousa and Gorlé, 2019) and the lack of information about the emission source in an accidental context (Mons et al., 2017). To address these uncertainties, a multi-query ensemble must be built but requires overwhelming computational resources.

In this context, reduced-order models (ROM) aim to accurately capture the behaviour of high-dimensional systems while reducing computational costs. Recent studies have explored the use of machine-learning algorithms for atmospheric dispersion and natural convection emulation (Margheri and Sagaut, 2016; García-Sanchez et al., 2017; Lamberti and Gorlé, 2021; Lucor et al., 2022), which is a particularly challenging problem for ROMs due to the advective nature of the system and the source location uncertainty. In this context, we have designed and thoroughly evaluated in a previous study, a data-driven ROM combining Proper Orthogonal Decomposition (POD) with Gaussian Process Regression (GPR) to forecast tracer concentration field first- and second-order statistics for a simplified turbulent atmospheric boundary-layer flow over a surface-mounted obstacle (Nony et al., 2023). Bayesian optimisation aided by POD optimised the GPR hyperparameters through the hierarchy of POD modes. Our analysis showed that the POD-GPR model was able to capture the large range of spatial scales seen in the POD modes, even with a limited training budget of approximately a hundred LES snapshots, while maintaining acceptable accuracy. However, when further reducing the LES training budget, the emulated tracer concentration fields could lose their physical consistency, in particular in regions where advection processes are dominant, for instance near the emission sources upstream of the obstacle (Nony, 2023). Based on this finding, the present study aims at improving the physics consistency and accuracy of POD-GPR capabilities under a small training LES data budget. To this end, more advanced deep-learning approaches based on convolutional autoencoder are utilised to overcome POD data compression limitations, and all of the LES statistics are exploited to explore the potential of multi-fidelity ROM combining high-fidelity, expensive LES solutions with lower-fidelity, less expensive RANS solutions.

The outline of this study is as follows. Section 2 presents the selected test case and describes the numerical configuration to generate the LES database. Section 3 introduces the ROM approaches using LES training data only. First, we recall the POD-GPR standard approach

(Sect. 3.1) and improve it by substituting POD with a convolutional autoencoder (Sect. 3.2). Such an approach still demonstrates a lack of physics consistency when trained on a sparse LES database. In response, Sect. 3.3 introduces an alternative approach, which aims to combine emulation of relevant flow quantities from LES that are injected in a lower fidelity Reynolds-average representation of tracer dispersion. Similar ideas of exploiting the rich LES information with a simplified transport equation (TE) for tracer dispersion are found in the literature: for instance, Du et al. (2020) proposed a simplified TE-ROM model based on flow statistics precomputed from LES data to give access to fast tracer dispersion inference, while preserving the rich information from the LES dataset. Finally, Sect. 4 explores a multi-fidelity ROM approach mixing sparse training 50 LES and 450 TE-ROM solutions using convolutional autoencoders alongside co-kriging and autoregressive processes to reduce the computational cost, while maintaining an acceptable level of accuracy.

2 DATABASE OF PARAMETERISED LARGE-EDDY SIMULATIONS AND PERFORMANCE EVALUATION CRITERIA

In this section, we present the two-dimensional test case for which we have generated a LES parameterised simulation database. This database corresponds to an ensemble of LES snapshots with multiple parameter entries, from which we aim at learning a ROM to accurately predict the dispersion quantities of interest. In this perspective, we also present the criteria to evaluate the quality of ROM predictions using the LES simulations as references.

2.1 Description of the test case

We simulate pollutant dispersion induced by the interactions between a fully-developed neutral turbulent boundary-layer flow and an isolated square-shaped obstacle (Fig. 1) in a two-dimensional domain, as in Nony et al. (2023). The pollutant is a passive tracer released at a constant rate from a point-source emission (the emission source axial position is x_{src} and height is z_{src}).

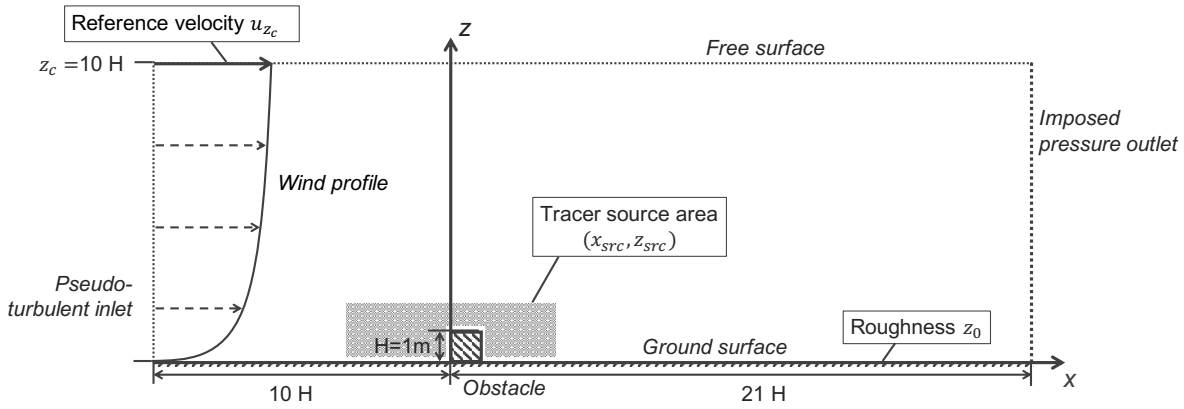

Figure 1: Schematic of an atmospheric turbulent boundary-layer flow around a surface-mounted square obstacle. Text boxes point to the uncertain parameters, *i.e.* pollutant source location, ground roughness and incoming flow velocity.

The turbulent boundary-layer flow is obtained by adding synthetically-generated wind fluctu-

ations to a time-averaged vertical wind profile corresponding to neutral atmospheric conditions:

$$\begin{cases} \overline{u}_{inlet}(z) = \dfrac{u_\tau}{\kappa} \log\left(1 + \dfrac{z}{z_0}\right), \\ \overline{u}_{inlet}(z = z_c) = u_{z_c}, \end{cases} \tag{1}$$

where $u_\tau\,(\mathrm{m\,s^{-1}})$ is the friction velocity, $\kappa = 0.41$ is the dimensionless von Kármán constant, $u_{z_c}\,(\mathrm{m\,s^{-1}})$ is the velocity magnitude at reference height ($z = z_c = 10$ m) and $z_0\,(\mathrm{m})$ is the aerodynamic roughness length. Assuming equilibrium with the inlet profile, the ground boundary condition ($z = 0$ m) is defined using a rough law-of-the-wall treatment based on the same aerodynamic roughness length z_0.

2.2 Choice of parametric uncertainty

As in Nony et al. (2023), we consider four uncertain parameters gathered into the vector $\boldsymbol{\mu} = (u_{z_c}, z_0, x_{src}, z_{src})^T \in \mathbb{R}^4$, which acts as input to the ROMs (Sects. 3–4). Based on typical ranges of parameters found in atmospheric dispersion studies, we consider the following log-uniform distribution for z_0 and uniform distribution for u_{z_c}:

$$\log(z_0) \sim \mathcal{U}\left(\log(10^{-3}), \log(10^{-1})\right)\ \mathrm{m}, \qquad u_{z_c} \sim \mathcal{U}([3, 9])\ \mathrm{m\,s^{-1}}. \tag{2}$$

The emission source can be located upstream, above or downstream of the obstacle. Its location uncertainty is characterised by the following uniform distributions from which the obstacle area is removed:

$$x_{src} \sim \mathcal{U}([-3.5, 3.5])\ \mathrm{m}, \quad z_{src} \sim \mathcal{U}([0.2, 2.0])\ \mathrm{m}. \tag{3}$$

2.3 Ensemble of large-eddy simulations snapshots

In this work, we rely on the AVBP[1] code developed by CERFACS (Gicquel et al., 2011) to generate the LES parameterised simulations (flow and tracer concentration field statistics) denoted by $\mathbf{K}_{\text{les}} = \{K_1, \ldots, K_{N_h}\}^T \in \mathbb{R}^{N_h}$, where N_h represents the number of grid elements in the domain of interest ($N_h \approx 240,000$ to properly resolve the flow/obstacle interactions).

The AVBP code solves the filtered compressible Navier-Stokes equations for flow dynamics and additional advection-diffusion equations for passive scalar dispersion on unstructured mesh. Subgrid-scale turbulence is represented using the standard Smagorinsky model (Smagorinsky, 1963). Each solution snapshot is obtained by averaging the LES instantaneous fields over a time window corresponding to 40 vortex shedding periods in the wake of the obstacle to guarantee statistical convergence for any inflow boundary condition (Nony, 2023).

A large database of 600 LES-based snapshots (corresponding to a total cost of 0.6 million CPU hours) is built, in order to train and test the ROM approaches. Each snapshot corresponds to a different realisation of the input parameter vector $\boldsymbol{\mu}$ following Halton's low-discrepancy sequence. For each snapshot, different quantities of interest are stored: the time-averaged tracer concentration field but also airflow field statistics such as the time-averaged airflow field and the turbulent kinetic energy. The analysis of this LES database shows that varying the tracer source location leads to very different regions of tracer accumulation, implying that there is a strong nonlinearity of the concentration field response to changes in the emission source location and to a lesser extent to the other parametric changes, which is a challenge for ROM (Nony et al., 2023).

[1] AVBP documentation, see http://www.cerfacs.fr/avbp7x/

2.4 Reduced-order model performance evaluation

Dataset splitting The full LES database made of N snapshots is split into two subsets following Halton's sequence ordering: *i)* a training dataset (80%, *i.e.* $N_{\text{train}} = 450$) to calibrate the ROM parameters; and *ii)* a test dataset to evaluate the ROM capacity to predict LES quantities of interest for new scenarios of input parameters (20%, *i.e.* $N_{\text{test}} = 150$). We analyse in this study how the ROM performance is modified when reducing the size of the training dataset to only 50 LES snapshots ($N_{\text{train}} = 50$) to move closer to a feasible budget for a multi-query LES framework on a realistic application. In this restricted framework, the test dataset remains the same to avoid introducing bias.

Performance evaluation To quantify the ROM performance in the physical space, we evaluate the Q^2 criterion on each feature. The reconstruction/prediction error is weighted by the variance over the LES samples for each grid element i:

$$Q_i^2 = 1 - \frac{\|\mathbf{K}_{\text{les},i} - \mathbf{K}_{\text{rb},i}\|_2^2}{\|\mathbf{K}_{\text{les},i} - \hat{\mathbb{E}}[\mathbf{K}_{\text{les},i}]\|_2^2}\,,\ \forall i = 1,\ \ldots\ , N_h, \tag{4}$$

where $\mathbf{K}_{\text{rb},i}$ corresponds to the ROM prediction at the ith grid element, whose objective is to best represent the LES counterpart $\mathbf{K}_{\text{les},i}$. The Q_i^2 criterion is estimated over the training dataset for verification and over the test dataset for evaluating the ROM prediction capacity. To help with the analysis, we also derive a global score from the variance-weighted local Q^2 criterion as

$$Q^2 = \sum_{i=1}^{N_h} \omega_i\, Q_i^2, \quad \omega_i = \frac{\hat{\mathbb{V}}(\mathbf{K}_{\text{les},i})}{\sum_{j=1}^{N_h} \hat{\mathbb{V}}(\mathbf{K}_{\text{les},j})}, \tag{5}$$

where $\hat{\mathbb{V}}(\mathbf{K}_{\text{les},i})$ corresponds to the variance unbiased estimation over the snapshots. This weighted average matches the usual explained variance criterion of the POD (Nony, 2023).

3 SINGLE-FIDELITY REDUCED-ORDER MODELLING APPROACHES

The very high computational cost associated with LES predictions motivates the offline construction of accurate ROM, which can produce in the online phase new ensemble predictions of the field statistics (*i.e.* for unexplored values of the uncertain parameters) at a very low computational cost. However, this is a challenging task since the LES quantities of interest are of very large dimension N_h, the number of available LES snapshots is limited ($N \ll N_h$), and the LES response to the input parameters $\boldsymbol{\mu}$ may be subject to strong nonlinearity due to the complexity of flow topology and tracer concentration patterns induced by the obstacle.

In this section, we introduce two ROM approaches for learning the parametric model response. We first recall a purely data-driven ROM approach directly learning the mapping between the uncertain parameters and the time-averaged tracer concentration fields based on LES data as in Nony et al. (2023), and we propose an improvement in the form of a convolutional neural network. We then propose and investigate an alternative physics-constrained ROM approach, which injects emulated flow quantities learned from LES data into a lower-fidelity Reynolds-averaged tracer transport equation; this equation corresponds to RANS formalism and predicts ensemble-averaged tracer concentration fields.

3.1 Data-driven POD-GPR reduced-order modelling approach based on LES data

3.1.1 Methodology

Due to the high dimension of the time-averaged tracer concentration fields, the data-driven ROM is built in a two-step approach following the choices made in Nony et al. (2023). The first step consists in projecting the LES fields onto a reduced-basis space of dimension L (with $L < N \ll N_h$) using a standard proper orthogonal decomposition (POD) approach (Sirovich, 1987; Berkooz et al., 1993). Within this POD framework, the approximated LES fields $\mathbf{K}_{\mathrm{rb}}$ are represented as linear combinations of the POD modes $\{\boldsymbol{\psi}_l\}_{l=1,\cdots,L}$:

$$\mathbf{K}_{\mathrm{rb}} = \sum_{l=1}^{L} k_l(\boldsymbol{\mu})\, \boldsymbol{\psi}_l(x,z)\,, \tag{6}$$

where $k_l(\boldsymbol{\mu}) \in \mathbb{R}$ is the l-th reduced coefficient to emulate with respect to the uncertain parameters $\boldsymbol{\mu}$ and to whiten (Kessy et al., 2018).

Based on the comparative study in Nony (2023) showing the better performance of Gaussian processes compared to other metamodeling techniques such as polynomial chaos and decision trees, emulation of each reduced coefficient k_l is carried out using a Gaussian process regression (GPR) model (Rasmussen and Williams, 2005). We consider L independent and noisy GPR models based on anisotropic Matérn covariance kernels (here characterised by the smoothness hyperparameter $\nu = 5/2$); anisotropy means that we consider a different length-scale for each input parameter using the automatic relevance determination (ARD) formulation. If we note $(\mathcal{U},\, \mathcal{K}_l) \,=\, \{(\boldsymbol{\mu}^{(n)}, k_l^{(n)}),\ n \,=\, 1,\ldots,N\}$ the training dataset dedicated to the l-th reduced coefficient and $(\mathcal{U}^*,\, \mathcal{K}_l^*)$ the test dataset counterpart, the GPR inference formula for the test reduced coefficients results from the posterior distribution

$$\mathcal{K}_l^* \;|\; \mathcal{U}, \mathcal{K}_l, \mathcal{U}^* \sim \mathcal{N}\left(m_l^*, \mathrm{cov}(\mathcal{K}_l^*)\right), \tag{7}$$

with

$$\begin{cases} m_l^* & = r_l(\mathcal{U}^*,\mathcal{U})\ [r_l(\mathcal{U},\mathcal{U}) + s_l^2\, I]^{-1}\ \mathcal{K}_l \\ \mathrm{cov}(\mathcal{K}_l^*) & = r_l(\mathcal{U}^*,\mathcal{U}^*) - r_l(\mathcal{U}^*,\mathcal{U})\ [r_l(\mathcal{U},\mathcal{U}) + s_l^2\, I]^{-1}\ r_l(\mathcal{U},\mathcal{U}^*), \end{cases} \tag{8}$$

where s_l^2 is the noise variance and r_l is the Matérn kernel. The hyperparameters of each GPR model may be optimised separately to best fit the characteristic length-scale of each POD mode Nony et al. (2023); this is done using a ten-restart gradient descent on the marginal log-likelihood (Rasmussen and Williams, 2005).

In this work, POD is implemented using the randomised truncated singular value decomposition (Halko et al., 2011) from the scikit-learn library (Pedregosa et al., 2011). GPR is also implemented using scikit-learn. It was shown in Nony et al. (2023) that a large number of reduced basis modes are required to accurately represent the LES ensemble variance at all points of the domain of interest (in particular, in the area upstream of the obstacle where the emission sources is sometimes located – Fig. 1).

3.1.2 POD-GPR prediction results

To quantify the performance of the POD-GPR approach, we employ the Q^2 ensemble performance metrics to accurately assess the ROM performance (Sect. 2.4). For all ROM implementations, Q^2 results can be found in Table 1.

To further validate the POD-GPR ROM, we present in Fig. 2 the emulation results for a test snapshot with nominal uncertain parameter values (this nominal snapshot has been selected to represent the ensemble mean of uncertainty distributions and to capture the various physical phenomena occurring near the obstacle such as recirculation areas, turbulent diffusion-dominated regions, and advection-dominated regions). Figure 2a shows the LES reference field; Fig. 2b shows the emulated field obtained with the POD-GPR ROM trained on the full dataset of 450 LES. This example demonstrates the capability of the ROM to accurately reproduce the main tracer concentration structures, particularly in regions where turbulent diffusion effects are dominant (for instance, in the wake of the obstacle far away from the emission source). These regions carry the majority of the ensemble variance and are easily predicted, thus leading to a high Q^2 global score of 96.7% on the test ensemble. Still, this global score masks some prediction errors that occur locally: *i)* in the vicinity of the emission source, where the estimated tracer concentration is substantially underestimated; *ii)* in the recirculation area upstream of the obstacle, where the tracer concentration levels are overestimated; and *iii)* in tracer-free regions upstream of the obstacle due to the noisy high-order POD modes (Nony et al., 2023).

We now examine how the POD-GPR prediction changes when the ROM is trained on the limited dataset of 50 LES. The associated global Q^2-score of 83.1% may first appear satisfactory but when examining the nominal snapshot prediction in Fig. 2c, it becomes clear that the ROM fails to accurately emulate the tracer concentration magnitude near the emission source, with noisy spurious structures appearing in tracer-free regions upstream of the emission source (white lines). Furthermore, the tracer concentration is significantly underestimated at the emission source and upstream of the obstacle, thus indicating a low quality of the ROM prediction. This result illustrates that physical consistency is lost upstream of the obstacle when reducing the training dataset. This non-physical behaviour is worsened by further reducing the training dataset (Nony, 2023) and is mainly due to POD, which is unsuitable for strongly nonlinear patterns that require a very large number of POD modes in the reduced basis ($L = 100$ for the full training dataset) while being incompatible with sparse training data. To mitigate physically-inconsistent emulations, we propose to improve the ROM structure by replacing POD with a deep learning convolutional autoencoder (AE).

3.2 Improving the data-driven reduced-order modelling approach using deep learning

In this section, we explore the ability of convolutional autoencoders (Murata et al., 2020; Fukami et al., 2020) to enhance the data-driven ROM performance by improving average concentration field compression, while meeting the constraint of reduced training dataset for LES.

3.2.1 Convolutional autoencoder implementation

Convolutional autoencoders cannot be directly applied to unstructured mesh since their convolutional kernels rely upon a Cartesian grid. For this reason, before proceeding to the learning stage, we linearly interpolate the LES field data onto a uniform Cartesian grid, whose spatial discretisation is consistent with that of the unstructured mesh, resulting in a total of 25,960 grid points and corresponding to tensors of dimension $\mathbb{R}^{295\times88}$.

Figure 3 shows the convolutional autoencoder architecture we use in this work and that is inspired by the work by Murata et al. (2020). This architecture includes convolutional layers at the periphery and dense layers in the core layers close to the latent vector. Without lack of generality, the size of the latent space is empirically set to $L = 10$.

The convolutional autoencoder was implemented using the Keras/Tensorflow libraries. Train-

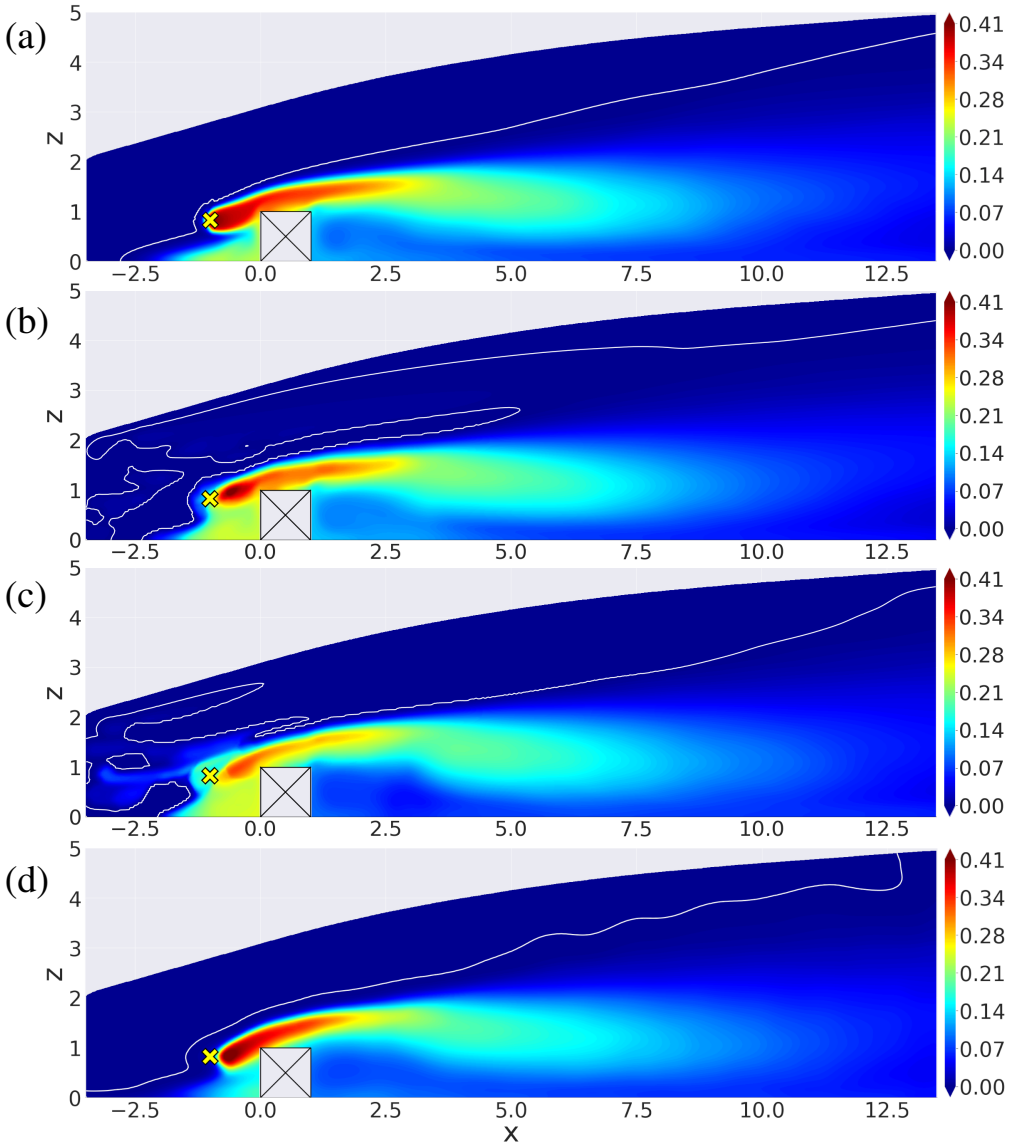

Figure 2: Normalised average tracer concentration field obtained for the nominal snapshot corresponding to $\boldsymbol{\mu} = (5.8\ \mathrm{m\,s^{-1}}, 2.8 \times 10^{-2}\ \mathrm{m}, -1.0\ \mathrm{m}, 0.8\ \mathrm{m})$. (a) LES reference. (b) POD-GPR ROM prediction when considering the full training dataset (450 LES snapshots) to compare to (c) POD-GPR ROM prediction obtained for the limited training dataset (50 LES snapshots). (d) AE-GPR ROM prediction also obtained for the limited training dataset (50 LES snapshots). White lines correspond to the 5×10^{-4}-contour line to highlight the presence of low-magnitude noisy structures.

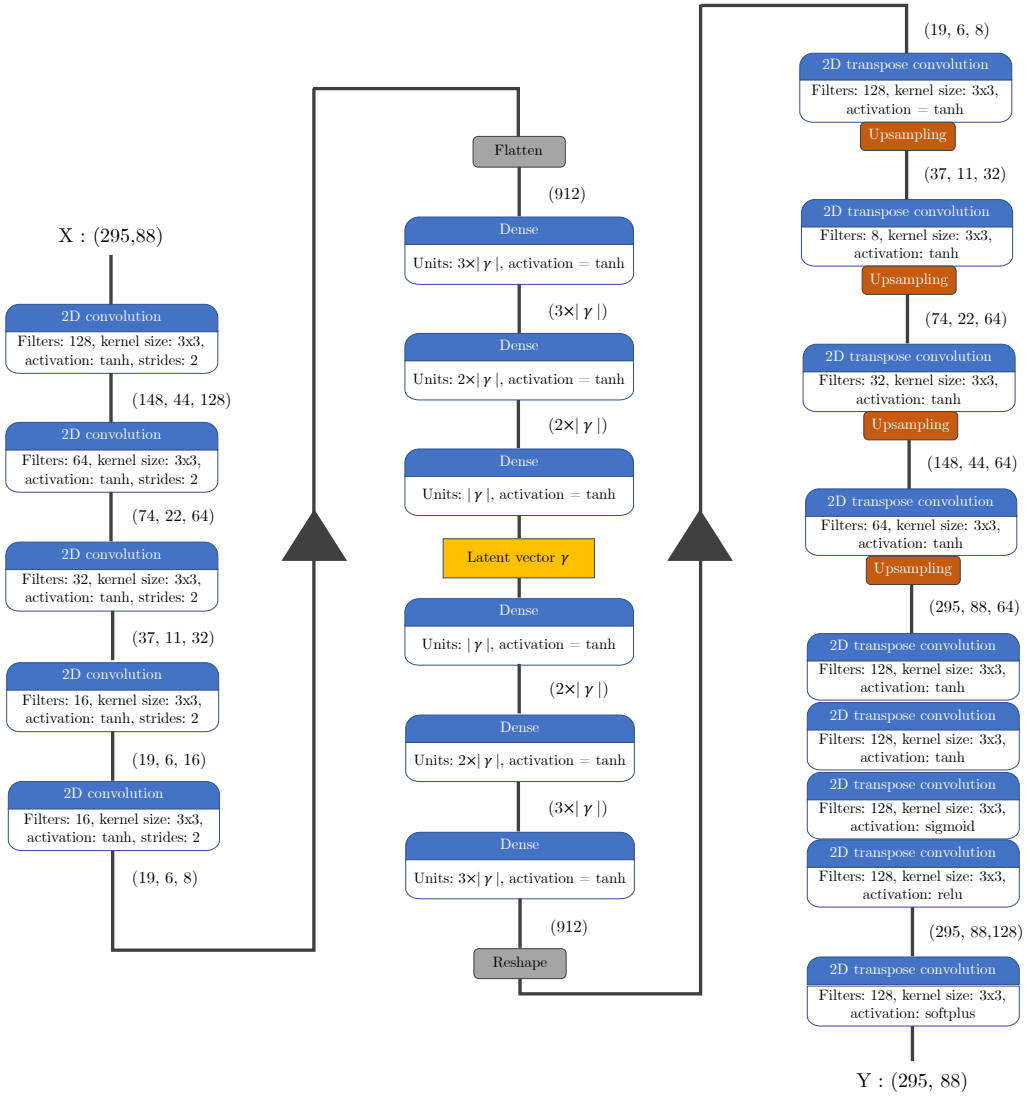

Figure 3: Convolutional autoencoder architecture consisting of i) convolutional layers (first and last columns) for feature extraction and dimension reduction, and ii) a dense multilayer perceptron (middle column) for high-level feature abstraction. γ corresponds to the latent vector of dimension L. Each rounded box describes a layer operation and the corresponding input/output tensor dimensions are specified underneath in brackets.

ing was performed by minimising the mean-square error via the ADAM descent scheme, with an initial learning rate of 10^{-3}, which was subsequently manually decayed to 10^{-4} and 10^{-6}. The convolutional autoencoder is combined with a GPR model for each latent variable, forming the AE-GPR ROM.

3.2.2 AE-GPR prediction results

Figure 2d shows the emulated nominal snapshot obtained with the AE-GPR ROM when considering sparse training data (*i.e.* using only 50 LES snapshots for the ROM training). This new ROM version improves the prediction results obtained with the POD-GPR approach at equivalent computational budget (Fig. 2c). The plume shape better matches that of the reference LES solution (Fig. 2a), both downstream of the obstacle and upstream near the emission source. Furthermore, the low-concentration contour line is found to be more faithful to the LES solution.

Still, it should be noted that near the emission source and near the upstream face of the obstacle, the tracer concentration peak is slightly shifted downstream from the actual source location, which is in violation of basic consideration of the physics of this problem.

Globally, the AE-GPR ROM achieves a global Q^2 score (84.8%) that is slightly higher than the POD-GPR ROM (83.1% – Table 1). The increase in performance may seem modest in view of the prediction improvements in the tracer concentration patterns (Fig. 2d), but these improvements are localised and do not represent a large fraction of the total variance. This motivates the implementation of scores that characterise position and shape errors, which is beyond the scope of the present work. It also points to the importance of obtaining physically sound predictions, in particular in regions with strong flow gradients, recirculations, etc.
In the following, we focus on designing a ROM approach that better satisfies physical constraints and gives access to more plausible tracer concentration predictions.

3.3 Physics-constrained reduced-order model based on LES-informed Reynolds-averaged tracer transport equations

In this section, we introduce an innovative hybrid approach referred to as RANS-TE, combining a ROM approach with physical equations. It consists of two steps: *i)* emulating relevant *airflow* statistics (*e.g.* mean flow field, turbulent kinetic energy) based on a POD-GPR ROM trained on LES data (as in Sect. 3.1), and *ii)* integrating these emulated statistics into a simplified scalar transport model via a Reynolds-average transport equation for the tracer, whose numerical resolution is much cheaper than simulating a full LES. In addition to saving computational time, this hybrid approach allows the flow dynamics to be decoupled from the tracer transport, thus separating the atmospheric parametric uncertainties from the source location parametric uncertainties. As tracer uncertainties are no longer sampled in the LES training database, the data-driven ROM approach can solely focus on atmospheric uncertainties (*i.e.* the reference velocity magnitude u_{z_c} and the aerodynamic roughness length z_0) in order to map the relevant airflow statistics. These emulated airflow statistics are then injected into the scalar transport equations to predict the average tracer concentration fields for different scenarios of source location.

3.3.1 LES-informed Reynolds-averaged tracer transport formulation

The key idea of the proposed RANS-TE hybrid approach is to solve the tracer transport equations by feeding some relevant information from the data-driven ROM trained onto the LES database for various flow scenarios. It is worth noting that supplementing DNS- or LES-based data into lower-order equations such as RANS is a growing strategy found in the literature (Steiner et al., 2022; Amarloo et al., 2022). Additionally, inverse modelling can be used to infer corrective fields to improve RANS closures accuracy (Parish and Duraisamy, 2016).

In the present work, the main flow statistics (velocity components, kinetic energy, and turbulent dissipation), denoted with a * superscript, are estimated from 50 LES snapshots and predicted using a POD-GPR framework based on $L = 10$ POD modes and an anisotropic Matérn kernel for GPR as detailed in Sect. 3.1. We note that much fewer POD modes are required here compared to the data-driven ROM approach presented in Sect. 3.1 as the learning only deals with atmospheric uncertainties and does not need to directly handle source uncertainties. For the tracer concentration, a more conventional Reynolds-averaged transport equation is solved

as

$$\overline{u_j}^{\star} \frac{\partial \overline{\mathbf{K}}}{\partial x_j} = -\frac{\partial}{\partial x_j} \left(\overline{\mathbf{K}' u_j'} \right) . \tag{9}$$

In this equation, the carrier velocity field $\overline{u_j}^{\star}$ is extracted from LES (as denoted by the * superscript), instead of being obtained via the resolution of the full RANS equations. Unclosed terms related to second-order velocity-tracer cross-correlations (turbulent diffusion term $\overline{\mathbf{K}' u_j'}$) remain to be modelled. A standard closure based on the Boussinesq assumption involving the mean tracer gradient is adopted here:

$$\overline{\mathbf{K}' u_j'} = -\frac{\nu_T^{\mathrm{RANS}}}{\mathrm{Sc}_T^{\mathrm{RANS}}} \left(\frac{\partial \overline{\mathbf{K}}}{\partial x_j} \right) , \tag{10}$$

where ν_T^{RANS} is the turbulent eddy-viscosity, and $\mathrm{Sc}_T^{\mathrm{RANS}}$ is the turbulent Schmidt number. As proposed by Yoshizawa et al. (2012), ν_T^{RANS} is solved through an additional transport equation for the turbulent eddy-viscosity that takes the following form:

$$\overline{u_j}^{\star} \frac{\partial \nu_T^{\mathrm{RANS}}}{\partial x_j} = C_{\mu P} \, \boldsymbol{k}_{\mathrm{tke}}^{\star} - C_{\mu \epsilon} \, \frac{1}{\tau_T^{\star}} \, \nu_t^{\mathrm{RANS}} + \nabla \cdot \left(\frac{\nu_T^{\mathrm{RANS}}}{\sigma_\nu} \nabla \nu_T \right) , \tag{11}$$

where $C_{\mu P}$, $C_{\mu \epsilon}$, σ_ν are modelling constants associated with turbulent production, dissipation, and diffusion; $\overline{u_j}^{\star}$ is the time-averaged velocity jth component; and $\boldsymbol{k}_{\mathrm{tke}}^{\star}$ is the turbulent kinetic energy. τ_T^* is a relevant turbulence time scale Yoshizawa et al. (2012) extracted from LES based on the ratio of turbulent kinetic energy $\boldsymbol{k}_{tke}^*$ to turbulent dissipation $\boldsymbol{\epsilon}^*$. More details about this RANS modelling and implementation approach can be found in Nony (2023).

3.3.2 RANS-TE prediction results

In this section, the RANS-TE ROM is implemented, for which airflow closure terms are approximated from a limited training dataset of 50 LES snapshots sampling the atmospheric uncertainties and the averaged concentration fields are obtained through the numerical resolution of the scalar transport equations for each source location sample. To evaluate the RANS-TE prediction performance, we rely on the quantitative Q^2 metrics and the nominal snapshot example as before.

Figure 4 shows the tracer concentration prediction obtained with the RANS-TE approach (Fig. 4a), and compares it to the reference LES solution (Fig. 4b). There is a good agreement between the two solutions upstream of the obstacle, although the RANS-TE model slightly overestimates tracer concentration. The shape of the wake is also accurately represented by the ROM. However, downstream of the obstacle, the ROM prediction diverges significantly from the LES solution, especially in the recirculation area near the obstacle. This induces a lower global Q^2 score for the RANS-TE approach (69.6%) than for the POD-GPR ROM approach (83.1% – Table 1).

The loss of performance in the downstream recirculation area is mostly due to the tracer turbulent flux closure, which is not intended to represent accurately turbulence transport for a two-dimensional setup, but it is not due to the airflow ROM lack of accuracy (Nony, 2023). Indeed, the emulation of the closure fields using POD-GPR reduces the Q^2 performance by only 2%.

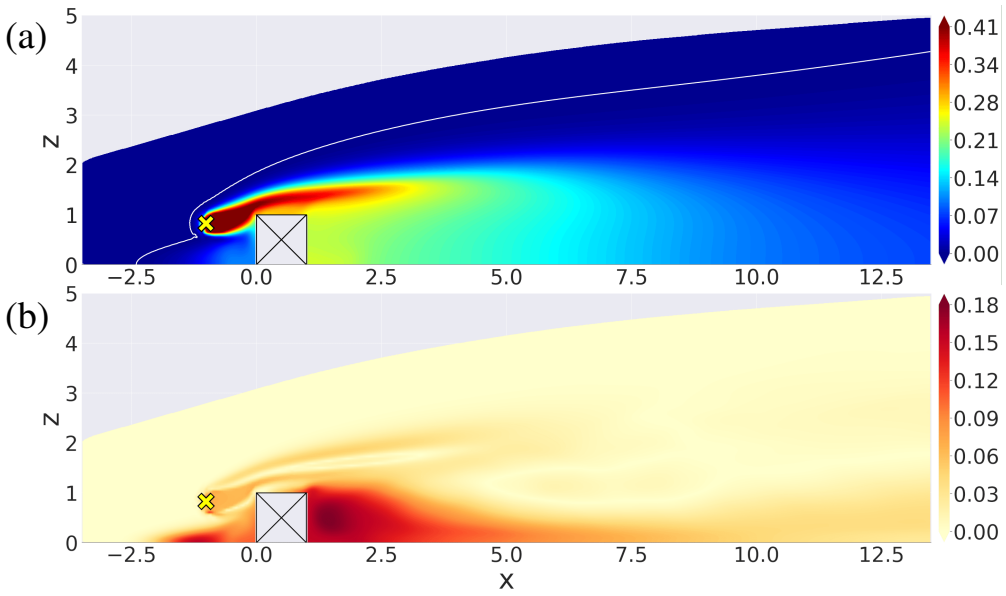

Figure 4: Nominal snapshot normalised average tracer concentration field. (a) RANS-TE ROM prediction. White lines indicate the presence of low-magnitude noisy structures (5×10^{-4}-contour lines). (b) Prediction absolute error measuring the discrepancies between the RANS-TE prediction and the LES reference solution.

3.4 Discussion about single-fidelity approaches

So far we have designed single-fidelity ROM approaches using LES data, cf. Table 1 for quantitative results. We first demonstrated the capability of a standard POD-GPR approach to accurately predict tracer concentration intensity and wake structure in regions dominated by turbulent transport. However, this approach had difficulty in capturing tracer concentration magnitude and patterns in advection-dominated regions upstream of the obstacle, where the concentration plumes are narrow, as nonlinearities are poorly handled by POD. The AE-GPR approach (based on a convolutional autoencoder for dimension reduction instead of POD) achieved better performance under small training dataset constraints. In a second step, to both reduce the training computational cost and better handle the tracer emission source uncertainties, the hybrid RANS-TE approach combining a data-driven ROM approach for airflow statistics and RANS transport equations was proposed. Although Q^2 performances were lower than for the AE-GPR ROM approach, the RANS-TE hybrid approach demonstrated its ability to accurately predict tracer concentration patterns in advection-dominated regions upstream of the obstacle. The complementarity of the RANS-TE and AE-GPR approaches in turbulent diffusion- and advection-dominated regions motivates the exploration of multi-fidelity ROM approaches, which could benefit from the advantages of each single-fidelity approach to reduce ROM prediction errors.

4 MULTI-FIDELITY REDUCED-ORDER MODELLING APPROACH

This section examines the potential for a more robust and efficient data-driven ROM by combining LES (Sect. 2.3) and RANS with transport equation (RANS-TE) modelling (Sect. 3.3). This type of approach mixing solutions of different fidelity levels is known as multi-fidelity (Goodfellow et al., 2016). The primary motivation is to harness the cheap computational cost of the low-fidelity RANS-TE model in order to more accurately predict average tracer concentration fields. In particular, multi-fidelity takes advantage of generating a larger training database at a reduced computational cost (compared to single-fidelity approaches presented in

Table 1: Single- and multi-fidelity ROM Q^2 global scores (in %) obtained by comparison to the LES test database. (size and nature of training simulations database are mentioned in parentheses).

ROM architectures	Training data	$Q^2(\%)$
POD-GPR	450 LES	96.7
POD-GPR	50 LES	83.1
AE-GPR	100 LES	94.0
AE-GPR	50 LES	84.8
RANS-TE	50 LES	69.6
MF-ROM	50 LES + 450 RANS-TE	92.6

Sect. 3). The RANS-TE model is here considered as the "low-fidelity" solver with faster evaluation, while the "high-fidelity" LES solver is very accurate but computationally expensive. The proposed multi-fidelity ROM is referred to as MF-ROM in the following.

4.1 Methodology

The present framework builds its multi-fidelity approach on a mixed training database made of 50 LES simulations along with 450 RANS-TE solutions. These RANS-TE solutions may be viewed as somewhat biased data as discussed in Sect. 3.3.2 as they themselves rely on LES data to be generated in the first place. In practice, the same 50 LES snapshots are also used to emulate the airflow field statistics for the RANS-TE approach (Sects. 3.3-3.3.2).

We focus on several extensions to multi-level response ROMs, which feature a similar non-intrusive structure to the POD-GPR approach. Multi-fidelity is introduced to reduce the dimensionality of the high-dimensional mean tracer concentration fields $\mathbf{K}_{\mathrm{les}}(\boldsymbol{\mu})$ and $\mathbf{K}_{\mathrm{te}}(\boldsymbol{\mu})$ obtained from LES and RANS-TE, respectively. Additionally, a multi-fidelity regression model is used to map the uncertainty parameters $\boldsymbol{\mu}$ onto the resulting compressed coefficients $\boldsymbol{k}_{\mathrm{les}}(\boldsymbol{\mu})$ and $\boldsymbol{k}_{\mathrm{te}}(\boldsymbol{\mu})$ in the latent space. Statistical approaches such as POD and GPR are initially tailored for single-level data. As an alternative, more sophisticated GPR procedures, such as those combining co-kriging and autoregressive models (Le Gratiet, 2013; Brevault et al., 2020), as well as convolutional autoencoders (AEs), which employ pre-training techniques (Goodfellow et al., 2016), may be utilised for this multi-fidelity purpose.

4.1.1 Pre-training for convolutional autoencoder

We implement the convolutional AE introduced in Sect. 3.2 using a pre-training transfer learning procedure (Goodfellow et al., 2016). The transfer learning here consists of taking a model that has already been trained for a given task and completing the training on new data to solve a different, but similar task. In our case, this procedure involves splitting the network training process into two steps. Firstly, the network weights are trained on the large dataset of 450 low-fidelity RANS-TE solutions $\mathbf{K}_{\mathrm{te}}$ until loss function convergence is reached. The best weights obtained in this step are then saved and used to initialise the second step of the training process, which is conducted based on the high-fidelity solutions until loss function convergence is reached again. In the end, the final network weights are the ones that best approximate the original 50 high-dimensional $\mathbf{K}_{\mathrm{les}}$ statistical fields. ADAM gradient descent is employed in both steps with an initial learning rate set to 10^{-3}, and manually stepwise-decreased to a final value of 10^{-6}. The resulting decoder is then used in the final multi-fidelity ROM (MF-ROM)

provided some latent space variables.

4.1.2 Multi-fidelity Gaussian processes using co-kriging and autoregressive models

The GPR framework is enhanced to model the mapping from the uncertain input parameters to the latent variables, *i.e.* $\boldsymbol{\mu} \mapsto \boldsymbol{k}_{\mathrm{MF-ROM}} \approx \boldsymbol{k}_{\mathrm{les}}$. Here, co-kriging with autoregressive models (Kennedy and O'Hagan, 2000; Le Gratiet, 2013) is used for multi-fidelity by training $L = 10$ independent GPR models on a multi-level formulation. Co-kriging combines data from different levels of fidelity using a weighted combination of the lower-fidelity data and a non-linear mapping of the higher-fidelity data. Autoregressive models use the lower-fidelity data to learn correlations between input variables, and model the nonlinear mapping of the higher-fidelity data.

In the first step, a standard GPR model (as in Sect. 3.1) is trained to emulate the low-fidelity latent variables $\boldsymbol{k}_{l,\mathrm{te}}(\boldsymbol{\mu})$. Then, using co-kriging, assuming the joint process $(\boldsymbol{k}_{l,\mathrm{les}}(\boldsymbol{\mu}), \boldsymbol{k}_{l,\mathrm{te}}(\boldsymbol{\mu}))$ is Gaussian given some parameters, and the Markov property, the high-fidelity latent variables $\boldsymbol{k}_{l,\mathrm{les}}(\boldsymbol{\mu})$ are derived using autoregression from the low-fidelity latent variables $\boldsymbol{k}_{l,\mathrm{te}}(\boldsymbol{\mu})$ (Le Gratiet, 2013):

$$\boldsymbol{k}_{l,\mathrm{les}}(\boldsymbol{\mu}) \approx \boldsymbol{k}_{\mathrm{l,MF-ROM}}(\boldsymbol{\mu}) = \rho_l(\boldsymbol{\mu})\, \boldsymbol{k}_{l,\mathrm{te}}(\boldsymbol{\mu}) + \delta_l(\boldsymbol{\mu}), \quad \text{for } l = 1, \ldots, L, \tag{12}$$

where $\delta_l(\boldsymbol{\mu})$ and $\boldsymbol{k}_{l,\mathrm{te}}(\boldsymbol{\mu})$ are assumed to be independent standard Gaussian processes with Matérn 3/2 kernels. The hyperparameters for $\delta_l(\boldsymbol{\mu})$ and $\boldsymbol{k}_{l,\mathrm{te}}(\boldsymbol{\mu})$ are denoted by $\boldsymbol{\theta}_{l,\delta}$ and $\boldsymbol{\theta}_{l,\mathrm{te}}$, respectively. The scaling factor $\rho_l(\boldsymbol{\mu})$ may be assumed to be constant, and is calibrated during the training process.

From such modelling assumptions, a Bayesian prediction model for the high-fidelity latent variables $\boldsymbol{k}_{l,\mathrm{MF-ROM}}(\boldsymbol{\mu}^*)$ can be derived from the posterior Gaussian distribution of $\boldsymbol{k}_{\mathrm{MF-ROM}}(\boldsymbol{\mu})$ given the dataset collection of RANS-TE and LES solutions $\mathcal{D}$, the scale factor ρ, and the hyperparameters $\boldsymbol{\theta}_{l,\mathrm{te}}$ and $\boldsymbol{\theta}_{l,\delta}$:

$$\boldsymbol{k}_{l,\mathrm{MF-ROM}}(\boldsymbol{\mu}^*) \mid \mathcal{D}, \rho_l, \boldsymbol{\theta}_{l,\mathrm{te}}, \boldsymbol{\theta}_{l,\delta} \;\sim\; \mathcal{N}(m_l(\boldsymbol{\mu}^*), r_l(\boldsymbol{\mu}, \boldsymbol{\mu}^*)), \tag{13}$$

with m_l and r_l the mean and covariance function of the Gaussian process. A closed form for the expression of $m_l(\boldsymbol{\mu}^*)$ is described in the work of Le Gratiet (2013), allowing for a fast numerical implementation of the emulation strategy.

Co-kriging and autoregressive models can then be used to model the mapping between uncertain parameters $\boldsymbol{\mu}$ and the latent mode representation of the time-averaged tracer response in a multi-fidelity context. This can be performed by optimising the Gaussian process hyperparameters using the Maximum Likelihood Likelihood (MLL) procedure, with multiple restarts (10 restarts) for improved accuracy.

4.2 Multi-fidelity results

To assess the performance of the multi-fidelity ROM, we use a Q^2 score evaluation on the test dataset and analyse the emulated nominal snapshot as in previous ROM prediction evaluations.

Figure 5 shows the prediction results for the nominal snapshot. This illustrates the superior performance of the multi-fidelity approach in comparison to the single-fidelity procedures trained under the sparse data constraint (Figs. 2cd). Namely, the emulated plume shape is in better agreement with the LES solution (Fig. 2a) and does not replicate the discrepancies of the RANS-TE, POD-GPR, and AE-GPR solutions. Downstream, the MF-ROM does not replicate

the strong errors obtained from the RANS-TE procedure. The upstream region is also better reconstructed than the AE-GPR solution. In particular, the concentration peak is accurately located at the emission source position, which was not the case for the single-fidelity AE-GPR approach (for which the peak concentration was slightly shifted downstream – (Fig. 2d). Furthermore, fewer artificial noise structures were generated in the upstream area for the multi-fidelity ROM prediction compared to the POD-GPR solution (Fig. 2c). Still, the presence of a noisy low-concentration $K = 10^{-4}$ concentration isoline in Fig. 5 implies that the convolutional autoencoder has not yet achieved convergence.

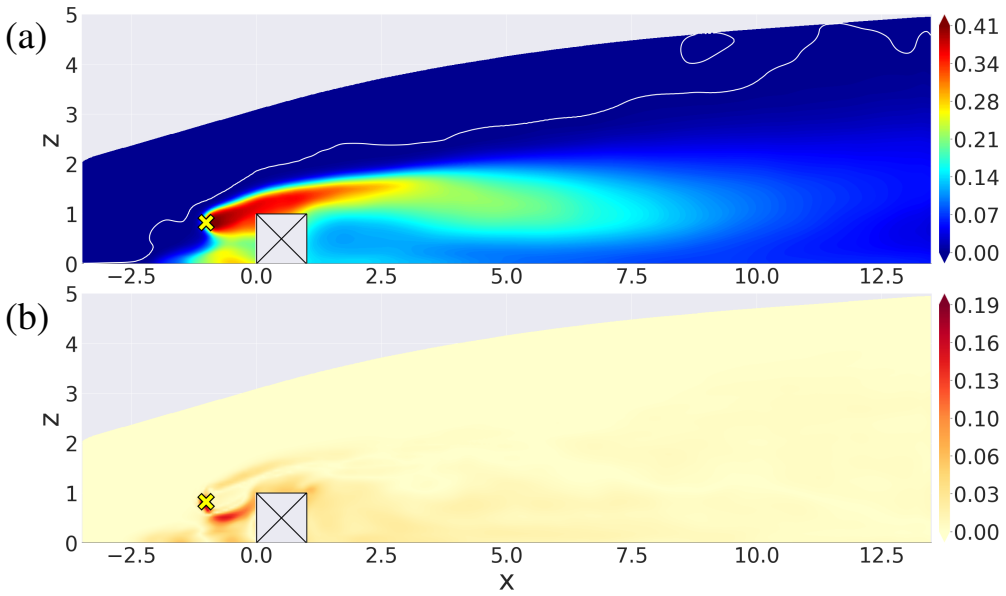

Figure 5: Nominal snapshot normalised average tracer concentration field. (a) MF-ROM prediction. The white lines indicate the presence of low-magnitude noisy structures, with the 5×10^{-4}-contour line indicating the magnitude of these structures. (b) Prediction absolute error relative to the reference LES solution.

The MF-ROM model achieves a global score $Q^2_{\text{global}} = 92.6\%$, which is a substantial improvement over the results obtained with single-fidelity ROM approaches based on 50 training LES snapshots (Table 1). This suggests that the MF-ROM approach benefits from the complementary information coming from both LES and RANS-TE snapshots, thereby increasing the accuracy of the convolutional AE and the parametric mapping of the latent space variables. Interestingly, the performance of the MF-ROM approach is even relatively close to that of the single-fidelity AE-GPR approach trained on 100 LES snapshots (Table 1). However, the latter is almost twice as expensive as the MF-ROM in terms of training cost, which makes the multi-fidelity approach promising for field-scale pollutant dispersion applications based on LES data.

5 CONCLUSION

In this study, we proposed a series of improvements to the finely tuned POD-GPR ROM approach introduced in Nony et al. (2023), using cutting-edge data-driven techniques based on LES data, for a simplified problem of pollutant dispersion in a turbulent boundary-layer flow. We observed that POD was not optimal to map the highly nonlinear relationship between the uncertain emission source location and the tracer concentration field values, leading to poor compression and a high required number of POD modes in the reduced basis. Convolutional autoencoders allow to significantly reduce the latent space dimension (by a factor of about ten) while improving ROM prediction accuracy. Still, ROM predictions can suffer in this case

from a lack of physical consistency for sparse LES training data. To overcome this limitation, which is an issue in field-scale pollutant dispersion applications, we have introduced a hybrid TE-ROM approach combining machine learning and a transport equation in order to constrain the emulated tracer concentration fields. Although this hybrid approach improves the physical consistency of the emulated fields in regions dominated by advection close to the emission source, the emulated fields are degraded in regions dominated by turbulent diffusion processes in the downstream recirculation area due to the limitations of the TE closure model formulation.

Interestingly, we have demonstrated that this novel TE-ROM approach still presents an interesting opportunity for multi-fidelity modelling. Using an augmented dataset of only 50 LES and 450 TE-ROM solutions, we have designed a multi-fidelity ROM combining a convolutional autoencoder with transfer learning and multi-fidelity GPR models. This novel TE-ROM achieves the best results of all tested ROM approaches at equivalent computational budget.

In future work, we plan to optimise the ratio between high-fidelity and low-fidelity solutions by leveraging modern techniques, taking advantage of active learning in order to maximise performance with minimal computational cost to move towards realistic field-scale pollutant dispersion applications.

References

A. Amarloo, P. Forooghi, and M. Abkar. Frozen propagation of reynolds force vector from high-fidelity data into reynolds-averaged simulations of secondary flows. *Physics of Fluids*, 34(11):115102, 2022.

G. Berkooz, P. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annu. Rev. Fluid Mech.*, 25(1):539–575, 1993. doi: 10.1146/annurev.fl.25.010193.002543.

L. Brevault, M. Balesdent, and A. Hebbal. Overview of Gaussian process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems. *Aerospace Science and Technology*, 107:106339, 2020.

F. T. Da Silva, N. C. Reis Jr, J. M. Santos, E. V. Goulart, and C. E. de Alvarez. The impact of urban block typology on pollutant dispersion. *J. Wind Eng. Ind. Aerodyn.*, 210:104524, 2021. doi: 10.1016/j.jweia.2021.104524.

T. Dauxois, T. Peacock, P. Bauer, C. P. Caulfield, C. Cenedese, C. Gorlé, G. Haller, G. N. Ivey, P. F. Linden, E. Meiburg, N. Pinardi, N. M. Vriend, and A. W. Woods. Confronting grand challenges in environmental fluid mechanics. *Phys. Rev. Fluids*, 6:020501, Feb 2021. doi: 10.1103/PhysRevFluids.6.020501. URL `https://link.aps.org/doi/10.1103/PhysRevFluids.6.020501`.

Y. Du, B. Blocken, and S. Pirker. A novel approach to simulate pollutant dispersion in the built environment: Transport-based recurrence CFD. *Build. Environ.*, 170:106604, 2020. doi: https://doi.org/10.1016/j.buildenv.2019.106604.

K. Fukami, T. Nakamura, and K. Fukagata. Convolutional neural network based hierarchical autoencoder for nonlinear mode decomposition of fluid field data. *Phys. Fluids*, 32(9):095110, 2020. doi: https://doi.org/10.1063/5.0020721.

C. García-Sanchez, G. V. Tendeloo, and C. Gorlé. Quantifying inflow uncertainties in rans simulations of urban pollutant dispersion. *Atmospheric Environment*, 161:263–273, 2017. doi: 10.1016/j.atmosenv.2017.04.019.

C. García-Sanchez, J. van Beeck, and C. Gorlé. Predictive large eddy simulations for urban flows: Challenges and opportunities. *Build. Environ.*, 139:146–156, 2018. ISSN 0360-1323. doi: j.buildenv.2018.05.007.

L. Gicquel, N. Gourdain, J.-F. Boussuge, H. Deniau, G. Staffelbach, P. Wolf, and T. Poinsot. High performance parallel computing of flows in complex geometries. *Comptes Rendus Mécanique*, 339(2-3):104–124, 2011.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT press, 2016.

A. M. Graham, K. J. Pringle, R. J. Pope, S. R. Arnold, L. A. Conibear, H. Burns, R. Rigby, N. Borchers-Arriagada, E. W. Butt, L. Kiely, C. Reddington, D. V. Spracklen, M. T. Woodhouse, C. Knote, and J. B. McQuaid. Impact of the 2019/2020 Australian megafires on air Quality and health. *GeoHealth*, 5(10):e2021GH000454, oct 2021. ISSN 2471-1403. doi: 10.1029/2021GH000454. URL `https://doi.org/10.1029/2021GH000454`.

T. Grylls, C. M. L. Cornec, P. Salizzoni, L. Soulhac, M. E. Stettler, and M. van Reeuwijk. Evaluation of an operational air quality model using large-eddy simulation. *Atmos. Environ.*, 3:100041, 2019. ISSN 2590–1621. doi: 10.1016/j.aeaoa.2019.100041.

N. Halko, P.-G. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

M. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

A. Kessy, A. Lewin, and K. Strimmer. Optimal whitening and decorrelation. *Am. Stat.*, 72(4): 309–314, 2018. doi: https://doi.org/10.1080/00031305.2016.1277159.

G. Lamberti and C. Gorlé. A multi-fidelity machine learning framework to predict wind loads on buildings. *J. Wind Eng. Ind. Aerodyn.*, 214:104647, 2021. doi: 10.1016/j.jweia.2021.104647.

L. Le Gratiet. *Multi-fidelity Gaussian process regression for computer experiments*. PhD thesis, Université Paris-Diderot-Paris VII, 2013.

D. Lucor, A. Agrawal, and A. Sergent. Simple computational strategies for more effective physics-informed neural networks modeling of turbulent natural convection. *J. Comput. Phys.*, 456:111022, 2022. URL `https://doi.org/10.1016/j.jcp.2022.111022`.

L. Margheri and P. Sagaut. A hybrid anchored-ANOVA–POD/Kriging method for uncertainty quantification in unsteady high-fidelity CFD simulations. *J. Comput. Phys.*, 324:137–173, 2016. doi: 10.1016/j.jcp.2016.07.036.

M. Mendil, S. Leirens, P. Armand, and C. Duchenne. Hazardous atmospheric dispersion in urban areas: A deep learning approach for emergency pollution forecast. *Environ. Model. Softw.*, 152:105387, 2022. doi: 10.1016/j.envsoft.2022.105387.

M. Milliez and B. Carissimo. Numerical simulations of pollutant dispersion in an idealized urban area, for different meteorological conditions. *Boundary-Layer Meteorol.*, 122(2):321–342, 2007. doi: 10.1007/s10546-006-9110-4.

V. Mons, L. Margheri, J.-C. Chassaing, and P. Sagaut. Data assimilation-based reconstruction of urban pollutant release characteristics. *J. Wind Eng. Ind. Aerodyn.*, 169:232–250, 2017. doi: 10.1016/j.jweia.2017.07.007.

T. Murata, K. Fukami, and K. Fukagata. Nonlinear mode decomposition with convolutional neural networks for fluid dynamics. *Journal of Fluid Mechanics*, 882, 2020.

B. Nony, M. Rochoux, T. Jaravel, and D. Lucor. Reduced-order modeling for parameterized large-eddy simulations of atmospheric pollutant dispersion. *Stochastic Environmental Research and Risk Assessment*, 2023. URL `https://doi.org/10.1007/s00477-023-02383-7`.

B. X. Nony. *Reduced-order models under uncertainties for microscale atmospheric pollutant dispersion in urban areas: exploring learning algorithms for high-fidelity model emulation*. Phd thesis, Université de Toulouse, France, 2023.

E. Parish and K. Duraisamy. A paradigm for data-driven predictive modeling using field inversion and machine learning. *Journal of Computational Physics*, 305:758–774, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

D. Philips, R. Rossi, and G. Iaccarino. Large-eddy simulation of passive scalar dispersion in an urban-like canopy. *J. Fluid Mech.*, 723:404–428, 2013. doi: 10.1017/jfm.2013.135.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN 9780262256834.

L. Sirovich. Turbulence and the dynamics of coherent structures. I - Coherent structures. II - Symmetries and transformations. III - Dynamics and scaling. *Q. Appl. Math.*, 45:561–571, Oct. 1987. doi: 10.1090/qam/910462.

J. Smagorinsky. General circulation experiments with the primitive equations: I. The basic experiment. *Mon. Weather Rev.*, 91(3):99–164, 1963. doi: 10.1175/1520-0493(1963)091⟨0099:GCEWTP⟩2.3.CO;2.

J. Sousa and C. Gorlé. Computational urban flow predictions with bayesian inference: Validation with field data. *Build. Environ.*, 154:13–22, 2019. doi: 10.1016/j.buildenv.2019.02.028.

J. Steiner, A. Viré, and R. P. Dwight. Classifying regions of high model error within a data-driven rans closure: Application to wind turbine wakes. *Flow, Turbulence and Combustion*, 109(3):545–570, 2022.

Y. Tominaga and T. Stathopoulos. CFD simulation of near-field pollutant dispersion in the urban environment: A review of current modeling techniques. *Atmos. Environ.*, 79:716–730, 2013. doi: 10.1016/j.atmosenv.2013.07.028.

H. Tsuruta, Y. Oura, M. Ebihara, T. Ohara, and T. Nakajima. First retrieval of hourly atmospheric radionuclides just after the Fukushima accident by analyzing filter-tapes of operational air pollution monitoring stations. *Sci. Rep.*, 4(1):6717, 2014. doi: 10.1038/srep06717. URL https://doi.org/10.1038/srep06717.

L. Vervecken, J. Camps, and J. Meyers. Dynamic dose assessment by Large Eddy Simulation of the near-range atmospheric dispersion. *J. Radiol. Prot.*, 35(1):165–178, jan 2015. doi: 10.1088/0952-4746/35/1/165.

A. Yoshizawa, H. Abe, Y. Matsuo, H. Fujiwara, and Y. Mizobuchi. A Reynolds-averaged turbulence modeling approach using three transport equations for the turbulent viscosity, kinetic energy, and dissipation rate. *Physics of Fluids*, 24(7):075109, 2012.