Reduced-order modeling for parameterized large-eddy simulations of atmospheric pollutant dispersion

B. X. Nony • M. C. Rochoux • T. Jaravel • D. Lucor

Accepted: 18 January 2023 / Published online: 15 February 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
Mapping near-field pollutant concentration is essential to track plume dispersion in urban areas. By solving most of the turbulence spectrum, large-eddy simulations (LES) can accurately represent concentration spatial variability. Synthesizing this large amount of information to improve the accuracy of lower-fidelity operational models is particularly appealing. Due to the computational cost of LES, it becomes a challenge in multi-query contexts, to quantify how tracer dispersion changes with atmospheric and source parametric variations. We propose a data-driven reduced-order model (ROM) combining proper orthogonal decomposition (POD) and probabilistic Gaussian process regression (GPR) to efficiently predict tracer concentration field first- and second-order statistics. The novelty of this work lies in the original approach used to efficiently tune the GPR based on the available reduced-basis information and properties. In practice, the GPR hyperparameters are optimized through the hierarchy of POD modes thanks to Bayesian optimization informed by POD. The robustness of the approach is also addressed under small data constraint. A detailed analysis of our model performance is carried out in the case of a turbulent atmospheric boundary-layer flow over a surface-mounted obstacle for which the convective character of the system together with the uncertainty of the pollutant source location make it very challenging for ROM. The model is able to capture the wide range of spatial scales across the POD modes. The ROM accuracy remains acceptable for a minimal budget of about one hundred LES snapshots, providing a first estimation moving towards more realistic atmospheric dispersion applications.

Keywords Air pollutant dispersion • Boundary-layer flow • Large-eddy simulation • Parametric uncertainty • Proper orthogonal decomposition • Gaussian process regression

1 Introduction
Accidental short-term pollutant emissions (e.g. 2011 Fukushima power plant explosion – Tsuruta et al. 2014; 2019/2020 Australian bushfires – Graham et al. 2021) can significantly degrade air quality and impact public health. In urban areas, peak pollutant concentrations are difficult to track due to the complex interactions between meteorology and urban topography (Philips et al. 2013). As a result, predicting the range of possible scenarios for near-source air pollutant dispersion in urban areas is critical for decision support in emergency situations (Da Silva et al. 2021; Mendil et al. 2022).

Computational fluid dynamics (CFD) is a complementary approach to field campaigns (e.g. Oklahoma City Joint Urban 2003 Experiment – Allwine and Flaherty 2006) to study microscale atmospheric dispersion processes in a
complex urban environment (Tominaga and Stathopoulos 2013; Dauxois et al. 2021). CFD has first been tackled through Reynolds-averaged Navier-Stokes (RANS) approaches (Milliez and Carissimo 2007; García-Sanchez et al. 2017), for which a time-averaged solution to the Navier-Stokes equations is sought, and turbulence is fully modeled via an approximate closure. By explicitly solving the large turbulent scales characterizing the flow, large-eddy simulation (LES) has emerged in recent years as an accurate reference solution to represent spatio-temporal variability of turbulent atmospheric flows in urban environment (Philips et al. 2013; Verwey et al. 2015a; García-Sanchez et al. 2018; Grylls et al. 2019). LES has a superior prediction capability compared to RANS, as it accurately captures highly unsteady and complex flow topologies typically found in the wake of buildings in urban canopies (Tominaga and Stathopoulos 2010). It also provides higher-order statistics such as concentration fluctuation and turbulent scalar flux fields that are not available with RANS. Yet, LES demands a much larger computational cost, which is a bottleneck for applications demanding uncertainty quantification.

Many sources of uncertainty in microscale CFD simulations come from large-scale atmospheric conditions that are difficult to represent due to their intrinsic variability (Sousa and Gorlé 2019), but also from limited knowledge of emission source properties (Mons et al. 2017). To deal with these uncertainties and build a multi-query uncertainty quantification framework suitable for CFD, one promising approach consists in substituting the parametric CFD model with a reduced-order model. The reduced-order model can be seen as a statistical model that has learned from a reduced CFD database the relationship between uncertain input parameters and the flow/tracer quantities of interest and that can then estimate the CFD model response over a wide range of parametric variation (Margheri and Sagaut 2016; García-Sanchez et al. 2017; Lamberti and Gorlé 2021). García-Sanchez et al. (2017), for example, used ensemble RANS simulations combined with polynomial chaos expansion to investigate how atmospheric uncertainties (inlet wind magnitude/direction and ground roughness) propagate on plume dispersion in Oklahoma City.

Key issues when building a CFD reduced-order model relate to i) the high dimension of the CFD outputs, and ii) the possible nonlinear input-output relationship that is difficult to learn from a limited training database. This is especially true in a LES setting that requires more computational resources than RANS, implying that the training stage can only rely on a limited LES database. To make the reduced-order model approach tractable in a LES framework, we adopt a two-step reduced-order model approach. The first step is to reduce the LES output dimension, i.e. by projecting the quantities of interest on a reduced basis. For this step, proper orthogonal decomposition (POD), also known as principal component analysis, has emerged as the benchmark approach in the literature due to its non-intrusive formulation, low computational cost, and hierarchical structure of the information that gives access to data interpretation (My-Ha et al. 2007; Marrel et al. 2015; van den Bos and Sanderse 2017). The second step consists of regressing or interpolating (i.e. metamodelling) the POD reduced coefficients with respect to the uncertain input parameters. The resulting reduced-order model in the form POD-regression can then be used to predict the quantities of interest for new input parameter entries. Since a reduced basis generally accounts for non-affine input/output relationships, basic interpolation techniques for the POD reduced-coefficient mapping may fail unless a sufficient number of LES samples is used during the training stage. In the literature, a wide variety of methods have been applied for predictive regression learning on numerical data such as polynomial chaos expansion (García-Sanchez et al. 2017; El Garroussi et al. 2022), decision trees (Xiang et al. 2021), Gaussian process regression (GPR) models (Marrel et al. 2015; Margheri and Sagaut 2016; Guo and Hesthaven 2018; Xiao et al. 2019; Ortali et al. 2022), and neural networks (Hesthaven and Ubbiali 2018; Swischuk et al. 2019; Ma et al. 2021; Lucor et al. 2022).

In the present work, the proposed reduced-order modeling framework is built from POD and GPR since both appear as standard choices to ensure result accuracy and interpretability. Typical practices found in the literature to explain the POD-GPR model behavior usually consist of POD mode visualization, POD explained variance, or POD-regression predictive performance quantification. However, the inner interactions between POD and GPR are rarely studied, while they may be of importance when the quantity of interest (such as tracer concentration) varies over many orders of magnitude, making the POD reduced coefficients very different from each other. In this study, we explore how GPR adapts to the POD reduced-coefficient regression, i.e. how GPR hyperparameters take into account the specific structures generated by POD. The novelty of the proposed method lies therefore in the design and evaluation of a Bayesian optimization procedure of the GPR model based on the maximum a posteriori. POD is used to draw non-uniform a priori distributions for the GPR hyperparameters to improve optimization convergence and thus improve the reliability of the offline learning stage. For this purpose, GPR hyperparameters are specific to each POD mode and calibrated accordingly.

The resulting POD-GPR reduced-order model is used to emulate time-averaged tracer concentration field statistics obtained with LES for a two-dimensional boundary-layer flow interacting with a surface-mounted obstacle. The
uncertain parameters considered in this study include the inflow boundary condition and the emission source location. We consider this simplified case study compared to actual urban canopy to have the capacity to generate a coherent LES database while remaining near to validation experiments that have been well-documented through field trials and physical modeling (Peterka et al. 1985; Martinuzzi and Tropea 1993). Such a LES database is valuable for carrying out a thorough evaluation of the reduced-order model performance depending on the POD and GPR configuration, on the choice of the quantity of interest, and on the size of the training database. Our analysis mainly focuses on time-averaged LES tracer concentration statistics but an extension to velocity-concentration cross-quantity (i.e. turbulent scalar flux) is also performed to investigate the capacity of the reduced-order model to generalize to coupled LES statistics. In addition, to address the small data constraint associated with the emulation of LES data, the sensitivity of the POD-GPR model accuracy to the size of the training database is also studied. These questions are of primary importance for accidental dispersion applications as it is crucial to have confidence in the reduced-order model predictions and to benefit from high-fidelity information coming from LES. This simplified case study therefore aims at providing guidelines for implementing a reduced-order model strategy suitable for future field-scale atmospheric dispersion cases in urban areas.

The outline of this paper is as follows. Section 2 presents the selected test case and the numerical configuration to generate the LES database. In Sect. 3, we present our POD-GPR reduced-order modeling approach. Section 4 analyzes in detail the performance of our reduced-order model with a focus on POD truncation, prior information for GPR hyperparameter optimization, extension to turbulent scalar flux, and sensitivity to training data.

2 Large-eddy simulation of turbulent flow over a surface-mounted obstacle

This section presents the two-dimensional test case and the numerical configuration considered in this study, with a focus on the modeling of inflow boundary conditions and parametric uncertainties. We run an ensemble of simulations with multiple parameter entries to collect LES snapshots from which a reduced-order model can be learned. Despite the fact that the solver we use does not resolve all flow scales (specificity of LES solvers), we will refer to it as a full-order model.

2.1 Numerical solver

The LES database is obtained with the AVBP\(^1\) LES code developed by CERFACS (Schönfeld and Rudgyard 1999; Gicquel et al. 2011). AVBP solves the filtered compressible Navier-Stokes equations on unstructured mesh. Additional advection-diffusion equations are solved for passive scalar dispersion. It is widely used to predict non-reactive and reacting unsteady flows in simple or complex geometry and is applicable to pollutant formation and atmospheric dispersion (Paoli et al. 2020).

In this work, the numerical discretization is based on an explicit, centered scheme from the continuous Taylor-Galerkin family called TTG4A, which is third-order in space and fourth-order in time on unstructured grids (Colin and Rudgyard 2000). In the low-Mach context of atmospheric dispersion, an artificial compressibility approach, also known as pressure gradient scaling (Ramshaw et al. 1985), is used to artificially reduce the speed of sound and thereby increase the time-step resulting from the Courant-Friedrichs-Lewy (CFL) condition.

The contribution of subgrid turbulence to the resolved flow variables is modeled with the Smagorinsky closure (Smagorinsky 1963) which influence on the LES prediction might also be of interest from a more theoretical point of view (Lucor et al. 2007). The subgrid scalar flux is modeled with the Boussinesq hypothesis with a subgrid turbulent Schmidt value equal to 0.7.

2.2 Case description and boundary conditions

This selected test case corresponds to an isolated square-shaped obstacle that is mounted on a rough ground surface (Vinçont et al. 2000; Gamel 2015). As depicted in Fig. 1, the obstacle interacts with a fully-developed neutral turbulent boundary-layer flow (from the left boundary).

The height of the obstacle is \(H = 1\) m. The two-dimensional computational domain is 31-m long (\(x\)-axis, streamwise direction) by 10-m high (\(z\)-axis, vertical direction). The domain height is ten times the obstacle height following guidelines for urban flow simulations (Franke et al. 2011). It is discretized with a uniform triangular mesh comprising 240,000 elements with a characteristic size \(\Delta x = \Delta z = H/10\).

The left side of the domain (at \(x = -10\) m) represents a turbulent inlet boundary where unsteady wind conditions are injected to mimic atmospheric boundary-layer turbulence as detailed in Section 2.3. The right side (at \(x = 21\) m) and the upper side (at \(z = 10\) m) correspond to an outlet with imposed pressure to mimic the atmosphere. The ground (at \(z = 0\) m) is modeled as a rough surface.

\(^1\) AVBP documentation, see http://www.cerfacs.fr/avbp7x/.
with a law of the wall based on the roughness length $z_0$ [m]. The obstacle surfaces (the obstacle is centered in $(x, z) = [0.5, 0.5] \times [m \times m]$) are modeled with the standard law of the wall.

The passive gas tracer is released from a point source emission, which can be either located upstream, above, or downstream of the obstacle as seen from the tracer source area in Fig. 1. The release is constant in time and modeled by a Gaussian shape in space with a standard deviation of 0.1 m around the center of release $(x_{src}, z_{src})$.

### 2.3 Inflow boundary condition modeling

#### 2.3.1 Average boundary condition modeling

The inlet wind condition imposes an average vertical profile $u_{inlet}$, based on the Monin-Obukhov similarity theory in neutral conditions:

$$u_{inlet}(z) = \frac{u_s}{\kappa} \log \left( \frac{z}{z_0} \right),$$

$$u_{inlet}(z = z_e) = u_{z_e},$$

where $u_s$ [m s$^{-1}$] is the friction velocity, $\kappa = 0.41$ is the dimensionless von Kármán constant, and $z_0$ [m] is the aerodynamic roughness length. In this study, the velocity $u_{z_e}$ enforced at height $z_e = 10$ m and the characteristic surface roughness length $z_0$ are used as input parameters in order to mimic operational conditions, where velocity measurements are typically obtained at some arbitrary reference height (Sousa and Gorlé 2019). From $u_{z_e}$ and $z_0$, Eq. 1 can be inverted to obtain the corresponding friction velocity $u_t$:

$$u_t = \frac{\kappa}{\log \left( 1 + \frac{z_e}{z_0} \right)} u_{z_e}.$$  

Consistently with the inlet condition, the same surface roughness length $z_0$ is considered assuming a fully-developed inlet flow over rough terrain.

#### 2.3.2 Inlet wind fluctuations

In addition to the average inlet wind profile from the Monin-Obukhov similarity theory, wind fluctuations are superimposed on the average profile to obtain a turbulent inlet boundary condition that mimics boundary-layer turbulence. The synthetic fluctuations are generated with the Kraichnan method implemented in the AVBP solver (Daviller et al. 2019) and follow the Passot-Pouquet turbulence spectrum (Passot and Pouquet 1987). The target turbulent kinetic energy $K$ at the inlet is estimated as $K = \frac{u_t^2}{\sqrt{C_\mu}}$ with $C_\mu = 0.09$ (Richards and Hoxey 1993).

### 2.4 Flow and tracer dispersion main features

#### 2.4.1 Output variable normalization

To analyze the LES simulations, we normalize the output velocity and concentration variables $u$ and $K$ by a reference velocity $u_t^{(ref)} = 3.7 \times 10^{-1}$ m s$^{-1}$ and a reference length scale (the obstacle height $H = 1$ m) as follows:

$$\tilde{u}(x, z) = \frac{u(x, z)}{u_t^{(ref)}}, \quad \tilde{K}(x, z) = \left( \frac{u_t^{(ref)} H^2}{Q_s} \right) \times K(x, z),$$

where the normalized velocity and tracer concentration variables are denoted by $\tilde{u}$ and $\tilde{K}$, respectively, and $Q_s$ is the flow rate. The value of $u_t^{(ref)}$ represents the mean friction velocity for the flow conditions explored in the parametric study (see Sect. 2.5 for further details). In the following, we drop the tilde notation for the sake of simplification.

#### 2.4.2 Instantaneous flow and tracer dispersion features

LES is used here to simulate unsteady flow and tracer features, and to obtain reference statistics (temporal average) of the flow velocity and tracer concentration fields across the domain. Figure 2 shows several instantaneous normalized tracer concentration fields along with
instantaneous vertical profiles of streamwise wind velocities for a given snapshot of the LES dataset. This snapshot can be considered as the nominal solution since it corresponds to the average atmospheric conditions considered in this study. This means that the inlet wind profile for the nominal snapshot is close to the ensemble mean in the parametric study (Sect. 2.5).

In this nominal case, the emission source is located upstream and at the height of the square. The tracer concentration patterns illustrate the complexity of the dispersion process. Upstream of the obstacle, the perturbed logarithmic profile (at $x = -3.5 \, H$) highlight the flow’s turbulent nature. As a result of the flow fluctuations, the tracer is either trapped in the recirculation zone on the windward face of the obstacle or advected downstream of the obstacle. Downstream of the obstacle, the flow is driven by a combination of the quasi-periodic vortex shedding induced by the flow-obstacle interaction and the background turbulence propagating from the inlet. The velocity profiles at $x = 5 \, H$ and $x = 10 \, H$ indicate that a reverse flow occurs near the ground, associated with a second recirculation region, transporting the tracer back towards the obstacle.

These qualitative observations match the typical patterns observed for flow and tracer dispersion in the vicinity of a square obstacle. In spite of the two-dimensional representation of turbulence, the numerical results are qualitatively similar to several experiments with bidimensional square section obstacles (infinitely long in the spanwise direction) (Vinçont et al. 2000; Gamel 2015).

2.4.3 Flow and tracer dispersion field statistics

The time-averaged tracer concentration is the primary output of interest for a dispersion study. Nevertheless, instantaneous vortex interactions along the streamwise direction may affect differently the time-averaged solution depending on the scenario as in the case of a spatially-developing plane mixing layers subject to uncertainties in the inflow boundary conditions (Ko et al. 2008). Other flow statistics, such as mean airflow statistics or second-order statistics (kinetic energy, momentum, and scalar turbulent fluxes), are useful for plume dispersion physical understanding and modeling. These quantities can be extracted from LES data, as they offer a rich description of the turbulent flow. However, it implies defining a time-averaging window to collect the statistics.

**Time-averaging process**

To initialize the LES simulations, the initial flow field is specified in adequation with the inflow boundary condition (Sect. 2.3). A spin-up is then necessary to establish turbulence throughout the computational domain and overcome the initial time transient. Following the spin-up, the LES model should be run for a certain time period to obtain LES field statistics (this is the statistically-stationary phase). In this study, the time-averaging window is based on the characteristic time scale of the vortex shedding. It
corresponds to 36 vortex shedding periods (the LES simulation is run for a total duration corresponding to 40 vortex shedding periods, and the first four periods are excluded from the averaging process). The corresponding physical time to simulate depends on the choice of the inflow parameters. For the nominal snapshot, this process leads to 414 s of physical time.

**First-order statistics: the average tracer concentration field**

The main quantities of interest in this study are the time-averaged normalized tracer concentration fields obtained with LES. They correspond to spatially-varying quantities. Figure 3 shows examples of normalized tracer concentration and streamwise velocity average fields for three parametric scenarios.

The first scenario presented in Fig. 3a corresponds to the nominal snapshot. The features already observed in the instantaneous fields in Fig. 2 are present in the average tracer concentration field. Close to the windward wall of the obstacle close to the ground, there is a first tracer accumulation area, which extends to axial locations upstream of the emission source due to the reverse flow velocity in this region. Above the obstacle top wall, the tracer is deflected and convected downstream, following the mean flow deviation induced by the obstacle. In the wake of the obstacle, the plume disperses due to the unsteady motion induced by vortex shedding.

The two other scenarios correspond to the case of an emission source placed downstream of the obstacle in a recirculation area where the tracer accumulates in large amounts (Fig. 3b) and to the case of an emission source located higher than the obstacle where there is no significant influence of the obstacle on the dispersion (Fig. 3c). This highlights that varying the tracer source location leads to different regions of tracer accumulation, implying that there is a strong nonlinearity of the concentration field response to changes in the emission source location, which is a challenge for the reduced-order model.

**Second-order statistics: the turbulent scalar flux field**

LES also provides access to higher-order quantities that are of interest for modeling tracer concentration statistics. For instance, the Reynolds averaging of the momentum,
and scalar transport equations yields second-order terms associated with correlations between flow features. For the scalar transport equation, the second-order term is the turbulent scalar. Figure 4 shows the resolved vertical turbulent scalar flux defined and normalized as \((\overline{wK} - \pi K)(\overline{\nabla'})\), where \(w\) is the vertical component of the velocity and \(\cdot\) represents the Reynolds time-averaging operator, for the nominal snapshot. The turbulent scalar transport features a horizontally elongated pattern and sharp sign changes along critical lines related to flow topology (shear layer, recirculation regions), highlighting the potential difficulty of the reduced-order model to emulate such quantities. An extension of the reduced-order modeling approach to the turbulent scalar flux is presented in this study to evaluate its capacity to handle both first-order and coupled statistics.

2.5 Uncertainty modeling

2.5.1 Sources of uncertainty

The objective of this study is to provide a methodology to explore how input parameter uncertainties propagate on the simulated tracer concentration statistics in the context of boundary-layer flow LES. These parameter uncertainties relate to the specification of the inflow boundary conditions (Sect. 2.3) and of the tracer emission source. We consider here four uncertain parameters: i) the reference velocity magnitude \(u_{cr}\) at domain’s height \(z_e = 10H\); ii) the aerodynamic roughness length \(z_0\); iii) the emission source position \(x_{src}\); and iv) the emission source height \(z_{src}\).

2.5.2 Parametric uncertainty

Distribution over the inlet wind conditions

The uncertain roughness length \(z_0\) and the reference velocity magnitude \(u_{cr}\) impact the average inlet wind profile, implying that the inlet wind \(u(x_{inlet}, z)\) becomes uncertain.

The range for the roughness length is set based on small obstacles between fallow ground and low mature agricultural crops (Wieringa 1992), leading to \(z_0 \in [10^{-3}, 10^{-1}]\) m. Its distribution is log-uniform with probability density function:

\[
\log(z_0) \sim U(\log(10^{-3}), \log(10^{-1}))
\]

corresponding to a mean value \(\mathbb{E}[z_0] \approx 0.021\) m and a standard deviation \(\sigma(z_0) \approx 0.025\) m (the coefficient of variation, i.e. the ratio of the mean value to the standard deviation, is 1.16). The distribution on \(z_0\) is defined as log-uniform so that the marginal distribution \(u(x_{inlet}, z)|u_{cr}\) is close to uniform. This wide range of variation for the roughness length means that a few snapshots in the LES database are near the validity limit of the LES approach (i.e. for large values of the roughness length). Still, this wide of variation was intended to allow for the reduced-order model evaluation over a wide range of the LES approach operability, and due to the log-uniform distribution the number of snapshots involved remains very small.

Streamwise velocity magnitude at domain’s height \(u_{cr}\) is supposed to follow a uniform distribution so that the marginal distribution \(u(x_{inlet}, z)|z_0\) is also uniform:

\[
u_{cr} \sim U([3,9]) \text{ m s}^{-1},
\]

corresponding to a mean value \(\mathbb{E}[u_{cr}] = 6.0\) m s\(^{-1}\) and a standard deviation \(\sigma(u_{cr}) \approx 1.7\) m s\(^{-1}\) (the corresponding coefficient of variation is 0.29). The range of variation for \(u_{cr}\) is chosen in agreement with urban air dispersion studies in the literature (García-Sánchez et al. 2014).

We define the reference velocity \(u_{ref}^{(\text{ref})}\) as the mean value of \(u_{cr}\) (Eq. 2) over the ensemble obtained by perturbing both the roughness length \(z_0\) and the streamwise velocity magnitude \(u_{cr}\) (using Monte Carlo random sampling and following the statistical distributions in Eqs. 4–5). This leads to \(u_{ref}^{(\text{ref})} = \mathbb{E}[u_{cr}] \approx 3.7 \times 10^{-1}\) m s\(^{-1}\), which is used to normalize the quantities of interest (Eq. 3).

The probability distribution on the average inlet wind profile \(\overline{u}(x_{inlet}, z)\) depends on the marginal distributions.

![Fig. 4 Vertical turbulent scalar flux obtained with LES for the nominal snapshot. Dashed lines correspond to the contour line along which the average normalized tracer concentration is equal to one-tenth of the maximum concentration to highlight the area of high concentration values. The yellow cross corresponds to the emission source location.](image-url)
chosen on the two uncertain parameters \(z_0\) and \(u_z\). Figure 5 shows how this distribution varies with respect to the vertical axis \(z\). At the domain’s height \(z = 10\) m, it follows a uniform distribution corresponding to the distribution on \(u_z\). But this is no longer the case when getting closer to the ground surface (i.e. when \(z < 10\) m), the probability distribution support decreases as \(z\) decreases. Several average inlet wind profiles based on varying entries for \(u_z\) and \(z_0\) are also plotted in Fig. 5 to illustrate the variety of profiles in the LES ensemble.

**Distribution over the tracer emission source characteristics**

Tracer uncertainty relates to the horizontal and vertical coordinates of the source location \((x_{src}, z_{src})\), also referred to as position and height, respectively. In this study, tracer emission can occur upstream, downstream, or above the obstacle. The range of variation is chosen so as to cover a broad panel of existing experimental studies (Mavroidis et al. 2007; Gamel 2015; Du et al. 2020). The marginal distributions associated with the source position and height are set uniform as:

\[
x_{src} \sim \mathcal{U}([-3.5, 3.5]) \text{ m, } z_{src} \sim \mathcal{U}([0.2, 2.0]) \text{ m.}
\]  

The range of variation for the source height is slightly away from the ground surface to avoid boundary issues. The area \([-0.2, 1.2] \times [0.2, 1.2] \text{ m}^2\) is removed from the range of variation to avoid truncation of the spatially Gaussian source term. In practice, uniform sampling of the source position and height is first carried out considering the whole area \([-3.5, 3.5] \times [0.2, 2.0] \text{ m}^2\). Then, samples located inside the obstacle area are removed from the ensemble used in this study.

**2.6 Full-order model database**

**2.6.1 Dataset acquisition**

In this study, we consider different sources of uncertainty in the LES full-order model: \(i)\) uncertainties associated with the large-scale atmospheric flow conditions \(\mu_{atm}\) (affecting the inflow and surface boundary conditions), and \(ii)\) uncertainties with the tracer emission source characteristics \(\mu_{tr}\). These uncertainties are described by uncertain scalar parameters that are considered as inputs to the LES problem (Sect. 2.5), and that form the input vector parameter:

\[
\mu = (\mu_{atm}, \mu_{tr})^T = (u_z, z_0, x_{src}, z_{src})^T \in \mathbb{R}^4.
\]  

The input parameters impact the simulated flow response and drive the quantities of interest \(K_{les} = \{K_1, \ldots, K_N\}^T \in \mathbb{R}^N\), namely the average tracer concentration or the turbulent scalar flux predicted at the \(N_h\) grid elements of the discretized computational domain (the nominal snapshot example is given in Figs. 3a and 4, respectively). Note that we restrict the area of interest on the \(x\)-axis to the interval \([-3.5, 13.5] \text{ m}\), containing 99.9% of the overall ensemble variance (this area still encloses the space of uncertainty on the source horizontal position \(x_{src}\) and height \(z_{src}\)). Moreover, all LES were conducted on the same numerical setting (same grid, convection scheme, LES model, etc.). Since the LES setup is robust, the non-
intrusive procedure is well decoupled from the LES model to only handle the parametric variability.

The LES full-order model provides an accurate prediction of the quantities of interest but at a significant computational cost, which prevents real-time outputs and requires extensive computational resources. These issues motivate the use of a reduced-order model, which can result in significant speedups to predict the quantities of interest for any new value of the input vector \( \mu \). Therefore, the objective of the reduced-order model is to approximate the LES model response on the space of the input parameters \( \mu \). This is carried out in the statistical learning framework introduced by Guo and Hesthaven (2018), where the learning stage (offline phase) corresponds to the construction (training) and evaluation (validation) of the reduced-order model. Later, multi-query evaluations of the reduced-order model (inline phase) allow predictions of additional scenarios and are thereby independent of the full-order model dimension. We focus next on the learning stage.

### 3.1 Principle of a reduced-basis approach

The purpose of reduced-basis approaches is to provide statistically efficient approximate solutions of the full-order model, decreasing the computational burden while minimizing the loss of accuracy. In this work, we aim at learning the mapping between the space of the uncertain input parameters \( \mu \in \mathbb{R}^d \) and the LES model response \( \mathbf{K}_{\text{les}} \in \mathbb{R}^{N_h} \). The reduced-order model builds an efficient representation of the LES high-fidelity dataset, i.e. a reduced-basis space of representation described by reduced-basis functions (or modes) \( \mathcal{V}_{\text{rb}} = \text{Span}(\{\psi_l\}_{l=1}^L) \subset \mathbb{R}^{N_h} \). The space \( \mathcal{V}_{\text{rb}} \) is assumed to be of low dimension compared to the number of grid elements (i.e. \( L \ll N_h \)). Reduced-basis solutions, denoted by \( \mathbf{K}_{\text{rb}} \), consist of the projected \( \mathbf{K}_{\text{les}} \) fields on \( \mathcal{V}_{\text{rb}} \). If the projection stands as a linear operator, \( \mathbf{K}_{\text{rb}} \) can be expressed as:

\[
\mathbf{G}_{\text{rb}} : \mathcal{P} \longrightarrow \mathcal{V}_{\text{rb}}
\]

\[
\mu \mapsto \mathbf{K}_{\text{rb}} = \sum_{l=1}^{L} k_l(\mu) \psi_l(x, z),
\]

where \( \{\psi_l\}_{l=1}^L \) correspond to the modes, and \( k_l(\mu) = \langle \mathbf{K}_{\text{les}}, \psi_l(x, z) \rangle \in \mathbb{R} \) is the \( l \)th reduced coefficient (or component). This approach provides a procedure to split the parametric dependency in \( \mu \) from the spatial dimension \((x, z)\) since local dissimilarity is now carried by the modes.

In this work, the modes are obtained from POD. This reduced-basis approach transforms LES data, i.e. projects the LES snapshots in a reduced space that is spanned by a set of parameter-independent functions \( \{\psi_l\}_{l=1}^L \). They thereby return discrete reduced-basis coefficients from the original output fields (further details on POD are given in Sect. 3.2).

Once the reduced basis is identified, regression meta-models can be trained to map the uncertain parameters \( \mu \) onto the reduced coefficients \( \{k_l\}_{l=1}^L \). Various methodologies have been explored in the literature, ranging from intrusive procedures such as POD-Galerkin (e.g.Afkham and Hesthaven 2017) to non-intrusive least-square learning such as polynomial chaos, neural networks, decision trees (e.g.Hesthaven and Ubbiali 2018; Mouradi et al. 2021;
Xiang et al. 2021), and Bayesian algorithms such as GPR (e.g. Guo and Hesthaven 2018). In this work, the regression component is performed using GPR (further details on GPR are given in Sect. 3.3). A preliminary comparative study (not presented here) indicated a clear performance gain with GPR compared to polynomial chaos and decision trees.

The resulting POD-GPR reduced-order model (Eq. 8) can then be used inline to estimate new quantities of interest $K_\text{rb}$ at new parameter values $\mu^*$ (i.e. at parameter values that are not included in the LES training database). The issues in building the reduced-order model for LES fields addressed in this work are three-fold: i) the quantities of interest simulated using LES are of very large dimension ($N_h$ is on the order of $10^5$); ii) the number of LES snapshots is limited due to the computational cost of a single LES (i.e. $N \ll N_h$); and iii) the mapping between the quantities of interest $K$ and the input parameters $\mu$ may be subject to nonlinearity. Additional difficulties associated with LES such as flow unsteadiness or the possibly large dimension of the input uncertainty space are beyond the scope of this work.

3.2 Proper orthogonal decomposition

3.2.1 Optimal reduced basis

In practice, POD (Sirovich 1987; Berkooz et al. 1993) computes an estimate of the optimal solution $V_{rb}^{n_{les}}$ from a collection of LES snapshots gathered in the snapshot matrix:

$$S = \left[ K_{les}^{(1)} | \ldots | K_{les}^{(N_h)} \right] \in \mathbb{R}^{N_h \times N}. \quad (9)$$

where $K_{les}^{(n)} \in \mathbb{R}^{N_h}$ corresponds to the $n_{th}$ LES snapshots in the database. POD seeks the optimal reduced basis of rank $L$ that stands as the optimal orthogonal projection manifold with respect to the Frobenius norm:

$$V_{rb}^{n_{les}} = \arg\min_{\mathbb{R}^{N_h}} \| S - S_{rb} \|_F, \quad (10)$$

$$S_{rb} = p(S; V_{rb}), \quad \text{rank}(V_{rb}) = L$$

with $p(\cdot; V_{rb})$ the projection from $\mathbb{R}^{N_h}$ to $V_{rb}$.

The idea behind POD is to find an orthonormal basis maximizing the variance of the projected field ensemble. Therefore, POD is usually implemented on the centered snapshot matrix. We note $T(\cdot)$ the linear operator applying an affine transformation (centering and normalization) to the snapshot data:

$$T : \mathbb{R}^{N_h} \longrightarrow \mathbb{R}^{N_h}$$

$$K_{les} \longmapsto \frac{1}{\sqrt{N-1}} (K_{les} - \overline{E}(K_{les})), \quad (11)$$

where

$$\overline{E}(K_{les}) = \left[ \overline{E}(K_{les,1}), \ldots, \overline{E}(K_{les,N_h}) \right] \in \mathbb{R}^{N_h}$$

is the empirical mean of the quantity of interest computed over the $N$ snapshots for each grid element, i.e. for the $i_{th}$ element:

$$\overline{E}(K_{les,i}) = \frac{1}{N} \sum_{n=1}^{N} K_{les}^{(n)}(i). \quad (12)$$

In the following, $T(S)$ refers to the matrix of transformed snapshots $\left( T(K_{les}^{(n)}) \right)_n$.

3.2.2 Dimension reduction

POD may now solve the diagonalization of the covariance matrix:

$$\text{Cov}(K_{les}, K_{les}) = T(S) T(S)^T = \Psi \Sigma^2 \Psi^T,$$

with

$$\Psi = [\psi_1 | \ldots | \psi_{N_h}] \in \mathbb{R}^{N_h \times N_h} \text{ an orthonormal matrix}$$

$$\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{N_h}) \in \mathbb{R}^{N_h \times N_h}, \quad \sigma_1 \geq \cdots \geq \sigma_{N_h} > 0. \quad (13)$$

Orthonormal vectors of $\Psi$ (the reduced-basis modes) carry some fraction of the ensemble variance quantified by the related eigenvalues in $\Sigma$ denoted by $\{\sigma_i\}_{i=1,\ldots,N_h}$ and satisfying $T(S) T(S)^T \psi_i = \sigma_i \psi_i$.

Since the number of mesh elements $N_h$ is very large, it becomes advantageous to keep the $L$ first modes (among the $N_h$ modes) that preserve the maximum variance of the original ensemble. The resulting truncated matrices are denoted by $\Psi = [\psi_1 | \ldots | \psi_{L}] \in \mathbb{R}^{N_h \times L}$ and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{L}) \in \mathbb{R}^{L \times L}$. The number of modes $L$ to retain to obtain an accurate reduced-order model is problem-dependent, and this is an open question for microscale pollutant dispersion. In this study, we therefore investigate the number of modes $L$ necessary to achieve a good level of accuracy in the different areas of interest (peak tracer concentration near the emission sources, tracer accumulation upstream and downstream of the obstacle, tracer dispersion downstream of the obstacle).

3.2.3 Interpretation

POD acts as a change of basis and the meaning of each mode can be interpreted by a correlation analysis between
the POD reduced coefficient \(k_l\) and the spatial concentration field \(\mathbf{K}_{\text{les}}\). In practical terms, the correlation field between the \(l\)th random variable \(k_l(\bm{\mu})\) and the random field \(\mathbf{K}_{\text{les}}(\bm{\mu})\) (varying between -1 and 1 by definition) may be stated as both terms depend on uncertain inputs \(\bm{\mu}\). For each mode \(l\), the correlation field can be expressed as:

\[
\text{Corr}(\mathbf{K}_{\text{les}}, k_l) = \left[ \frac{(N-1)\sigma_l}{\sqrt{\text{Var}(\mathbf{K}_{\text{les}})}} \right]_i,
\]

where the indices \(i\) correspond to a given element of the field \(\text{Corr}(\mathbf{K}_{\text{les}}, k_l) \in \mathbb{R}^N\), where \(\psi_{li}\) corresponds to the \(i\)th element of \(\Psi_l \in \mathbb{R}^N\), the \(l\)th function in \(\bar{\Psi}\), and where \(\bar{\Psi}(\mathbf{K}_{\text{les}})\) represents the variance unbiased estimation over the snapshots for the \(l\)th grid element. This variance is estimated as:

\[
\bar{\Psi}(\mathbf{K}_{\text{les}}) = \frac{1}{N-1} \sum_{n=1}^{N} (K_{\text{les},i}^{(n)} - \bar{\mathbf{E}}(K_{\text{les}}))^2.
\]

The linearity in the operator \(T\) makes it simple to express the inverse reconstruction operator from which the original quantity of interest field may be recovered from the compressed information carried by the POD reduced coefficients:

\[
\mathbf{K}_{\text{les}} \mapsto k = \Sigma^{-1/2} \bar{\Psi}^T T(\mathbf{K}_{\text{les}}),
\]

where \(\Sigma^{-1/2} = \text{diag}(1/\sqrt{\sigma_1}, \ldots, 1/\sqrt{\sigma_N}) \in \mathbb{R}^{L \times L}\) and \(\bar{\Psi} \in \mathbb{R}^{N \times L}\) are the matrices restricted to the \(L\) first modes (Sect. 3.2.1), and \(k = [k_1, \ldots, k_L] \in \mathbb{R}^L\) is the vector of POD reduced coefficients (Eq. 8).

The transformation in Eq. 16 may now be applied to the original dataset (expressed in the grid space) to create a new dataset of POD reduced coefficients \(k\) for varying input parameters \(\bm{\mu}\). The compressed information is suitable for modeling (through a metamodel) the relationship between the average tracer concentration field and the uncertain input parameters. This is detailed in the next section.

### 3.3 Gaussian process regression

In this work, rather than training a cumbersome metamodeling procedure involving the high-dimensional LES output fields, we train metamodels to predict the reduced coefficients \(k = [k_1, \ldots, k_L]\) from the input parameters \(\bm{\mu}\) through GPR (Rasmussen and Williams 2005).

#### 3.3.1 Metamodel formulation

**Learning task**

POD decomposes ensemble variance into hierarchical information carried by the \(L\) reduced coefficients. The first modes carry the large energetic structures contained in the data, whereas the higher modes focus more and more on local effects. Moreover, the parametric variability induces strong changes in the modal decomposition distribution. It is for instance obvious that a change in the geometric position of the tracer emission is going to drastically affect the amplitude and/or the signs of the POD coefficients. As a result, the associated response surfaces \(k = k(\bm{\mu})\) tend to reflect the characteristic scales of the tracer concentration patterns.

We aim to construct GPR models that take into consideration this prior knowledge (e.g. typical length-scales) in order to be efficient at accurately and robustly describing the POD reduced coefficients. Therefore, one key aspect of the metamodeling stage deals with the choice of the kernel (Sect. 3.3.2) and the hyperparameter optimization (Sect. 3.3.3) to adapt the Gaussian processes to the variety of length-scales across the POD modes.

Because the POD reduced coefficients are decorrelated (Eq. 18), we design \(L\) independent GPR models, i.e. we learn the relation between each reduced coefficient \(k_l\) and the input parameters \(\bm{\mu}\) for \(l\) varying from 1 to \(L\) (Fig. 6).

**Noisy training data**

In this work despite the fact that we consider the LES quite accurate, we wish to keep the opportunity to treat the
data as noisy. For instance, we may consider the LES data to be somewhat noisy due to time averaging. To account for this noise, the non-interpolating GPR model for the $l$th reduced coefficient $k_l$ is written as:

$$k_l = f_l(\mu) + \epsilon_l,$$

where $\epsilon_l \sim \mathcal{N}(0, s_l^2)$ is an additive noise term with variance $s_l^2$.

### Reduced-coefficient prediction

The GPR framework assumes that the mapping $f_l$ (Eq. 19) is a Gaussian stochastic process such that:

$$f_l(\mu) \sim \mathcal{GP}(m_l(\mu), r_l(\mu, \mu')) \ \forall (\mu, \mu') \in \mathcal{P} \times \mathcal{P},$$

with $m_l(\mu) = \mathbb{E}[f_l(\mu)]$ the mean function of the Gaussian process and

$$r_l(\mu, \mu') = \mathbb{E}[(f_l(\mu) - m_l(\mu))(f_l(\mu') - m_l(\mu'))]$$

its associated covariance function. GPR starts with a prior distribution over the mean and covariance function. In this work, the prior mean is assumed to be 0 because of POD data transformation (Sect. 3.2), and the prior covariance is specified by choosing a kernel (Sect. 3.3.2). The learning stage consists in updating the mean and covariance by integrating the information from the reduced-coefficient dataset (Sect. 3.2.4).

We note $(\mathcal{U}, K_i) = \{(\mu^{(n)}, k^{(n)}_i), \ n = 1, \ldots, N\}$ the data-set dedicated to the $l$th reduced coefficient. We use the raw samples of the Halton’s sequence over the interval $[0, 1]^d$ as the input data to GPR in order to ensure that the samples are adequately distributed over the uncertain space. This is motivated by the fact that the raw Halton’s sequence has suitable statistical properties (same parameter range $[0, 1]$, approximately identical distribution), while the input parameters are of a very different nature. In particular, the surface roughness $z_0$ is logarithmically distributed and orders of magnitude smaller in comparison to the other input parameters.

From now on, we make the distinction between the training sample and the test sample: the corresponding $N_{\text{train}}$ and $N_{\text{test}}$-collection matrices are denoted by $(\mathcal{U}, K_i)$ and $(\mathcal{U}^*, K_i^*)$. It should be noted that the test samples are unknown during the GPR training stage (they are used as an independent dataset from the training sample for validation purposes). The goal is to predict the output values $K_i^*$ from the predictive Gaussian distribution (Eq. 22) evaluated at $\mathcal{U}^*$. Using these notations, the joint distribution between the training dataset and some new test evaluations is expressed with respect to the kernel as:

$$\left[ \begin{array}{c} K_i \\ K_i^* \end{array} \right] \sim \mathcal{N}\left(0, \left[ \begin{array}{cc} r_l(\mathcal{U}, \mathcal{U}) + s_l^2 I & r_l(\mathcal{U}, \mathcal{U}^*) \\ r_l(\mathcal{U}^*, \mathcal{U}) & r_l(\mathcal{U}^*, \mathcal{U}^*) + s_l^2 I \end{array} \right] \right),$$

(21)

where $s_l$ is the noise variance (Eq. 19), and $I$ stands for the identity matrix. This equation is standard for the GPR framework (Rasmussen and Williams 2005), and is valid under the assumptions of noisy observations and gaussianity of Gaussian process pointwise values. Following again Rasmussen and Williams (2005), the predictive distribution from which values of $K_i^*$ can be sampled is computed as:

$$K_i^* | \mathcal{U}, K_i, \mathcal{U}^* \sim \mathcal{N}(m_i^*, \text{cov}(K_i^*)),$$

(22)

where

$$\left\{ \begin{array}{l} m_i^* = r_l(\mathcal{U}^*, \mathcal{U}) [r_l(\mathcal{U}, \mathcal{U}) + s_l^2 I]^{-1} K_i \\ \text{cov}(K_i^*) = r_l(\mathcal{U}^*, \mathcal{U}^*) + s_l^2 I - r_l(\mathcal{U}^*, \mathcal{U}) [r_l(\mathcal{U}, \mathcal{U}) + s_l^2 I]^{-1} r_l(\mathcal{U}, \mathcal{U}^*) \end{array} \right.$$

(23)

All terms in Eq. 23 are known. The covariance formulation depends on prior variance $r_l(\mathcal{U}^*, \mathcal{U}^*)$ over the test dataset refined by information from the training dataset. In this formulation, the matrix $[r_l(\mathcal{U}, \mathcal{U}) + s_l^2 I]$ is inverted using a computationally-efficient Cholesky decomposition (Rasmussen and Williams 2005). Therefore, the posterior distribution for the $l$th reduced coefficient can be directly estimated using Eq. 22. The GPR estimator is set as the mean posterior, which is a linear combination of kernel
distances computed between the test point and all the training data.

### 3.3.2 Anisotropic kernel

The kernel is at the core of GPR as it entails specific assumptions on data covariance in the input space $\mathcal{P}$. We consider here the radial basis function (RBF), also known as the squared-exponential kernel, as it is a common choice in machine learning framework applied to numerical solutions (e.g. Guo and Hesthaven 2018; Ortali et al. 2022).

**Kernel definition**

The RBF kernel leads the underlying process to be stationary since it is expressed as a function of the $\ell_2$-norm distance $d = \| \mu - \mu' \|_2$ for $\mu, \mu' \in \mathcal{P}$:

$$r_{\text{RBF}}(d) = q_1 \exp \left(-\frac{d^2}{2\lambda_i^2}\right),$$

where $q_1$ is the signal variance parameter, and $\lambda_i > 0$ is the length-scale (or stability hyperparameter).

**Length-scale hyperparameters**

The length-scale parameter $\lambda_i$ represents the level of variability in the reduced coefficients as a function of distance in the input space $\mathcal{P}$. In this work, we consider different length-scales for each input parameter since the input parameters are of different nature (source location versus atmospheric inflow conditions). Anisotropy may be embedded using a distinct correlation length-scale per dimension:

$$d(\mu^{(m)}, \mu^{(n)}) = \sqrt{(\mu^{(m)} - \mu^{(n)})^T A_i (\mu^{(m)} - \mu^{(n)})},$$

where $\mu^{(m)}$ and $\mu^{(n)}$ are realizations of the input vector $\mu$, and where $A_i \in \mathbb{R}^{d \times d}$ corresponds to the length-scale matrix. In this work, this matrix will be of the form $A_i = \text{diag}(1/\lambda_{x_{\text{src}}}, 1/\lambda_{z_{\text{src}}}, 1/\lambda_{x_{\text{h}}}, 1/\lambda_{z_{\text{h}}})$, which assumes independent length-scales and discriminates Gaussian process instability according to each input dimension. This form of the Gaussian process length-scale matrix is referred to as automatic relevance determination (ARD) in the literature.

### 3.3.3 Hyperparameter optimization

Hyperparameter settings have a substantial impact on the GPR model prediction performance. An optimization process is usually used to determine an optimal value for the hyperparameters rather than simply specifying them. In this study, we optimize the hyperparameters for each of the $L$ GPR models to adapt to the characteristic length-scale of each POD mode. For this purpose, we compare different optimization approaches in terms of accuracy and efficiency.

For a noisy GPR framework, the set of hyperparameters $\theta_i$ includes the correlation length-scales, the noise variance, and the Gaussian process variance, meaning that

$$\theta_i = \{d_i, q_1, \lambda_{x_{\text{src}}}, \lambda_{z_{\text{src}}}, \lambda_{x_{\text{h}}}, \lambda_{z_{\text{h}}}, \lambda_{x_{\text{h}}}, \lambda_{z_{\text{h}}}\} \in \mathbb{R}^{10}.$$  

Bayesian maximization is used to determine the optimal set of the hyperparameters maximizing their posterior:

$$\theta_{i, \text{opt}} = \arg \max_{\theta_i} \log p(\theta_i | \mathcal{U}, \mathcal{K}_i) \Rightarrow \theta_{i, \text{opt}} = \arg \max_{\theta_i} \log p(\mathcal{K}_i | \mathcal{U}, \theta_i) + \log p(\theta_i).$$

The first term called the marginal log-likelihood is expressed as:

$$\log p(\mathcal{K}_i | \mathcal{U}, \theta_i) = -\frac{1}{2} \log \left[ \frac{1}{\lambda_i} \left( r_i(\mathcal{U}, \mathcal{U}) + s_i^2 I \right) \right]^{-1} \mathcal{K}_i$$

$$-\frac{1}{2} \log | r_i(\mathcal{U}, \mathcal{U}) + s_i^2 I | = 2 \pi,$$

observing $p(\mathcal{K}_i | \mathcal{U}, \theta_i)$ is Gaussian with Gaussian observation noise (Eq. 21). The second term in Eq. 26 involves the prior distribution over the hyperparameters $p(\theta_i)$. In the literature, gradient descent is widely used to find the local optimum of Eq. 26 (Rasmussen and Williams 2005).

**Maximum log-likelihood estimation**

Without making any further assumption about the noise or the length-scales, one may proceed with a naive optimization of the log-likelihood, assuming uniform prior distributions over the hyperparameters (i.e. the term $p(\theta_i)$ in Eq. 26). Unfortunately, gradient descent algorithms perform poorly in this case due to multiple local optima. To overcome this issue, a standard way is to perform multiple gradient descent iterations starting from different hyperparameter initial conditions. The final solution is then chosen as the one achieving the highest maximal log-likelihood (MLL) score. One drawback is that these multiple gradient descent iterations increase the computational cost of the GPR approach (recall that in this work we have $L$ GPR models to train and therefore we have to repeat this optimization procedure $L$ times).

**Maximum a posteriori estimation**

In this work, we propose an alternative approach to the standard MLL approach to avoid multiple gradient descent iterations. The key idea is to better pose the optimization problem by providing non-uniform prior distributions (the term $p(\theta_i)$ in Eq. 26) and thereby have a starting point for the hyperparameters that is not far from the sought solution. This approach is expected to converge to a physically-
consistent local minimum in a single iteration. We refer to this approach as the maximum a posteriori (MAP) estimation. The novelty here is that the prior information is directly inferred from the POD coefficients (Sect. 4.2). The resulting MAP approach will be compared to the standard N-restart MLL approach in the following.

3.4 Reduced-order model performance evaluation

3.4.1 Training, calibration, and test

In this study, LES data (made of 750 snapshots – Sect. 2.6.2) are split into three subsets following Halton’s sequence ordering: i) a training dataset to learn the POD modes and the POD reduced coefficients using GPR (63%, i.e. \(N_{\text{train}} = 472\)); ii) a calibration dataset to estimate GPR hyperparameter prior distributions (7%, i.e. \(N_{\text{calib}} = 53\)); and iii) a test dataset to evaluate the capacity of the POD-GPR model to predict LES quantities of interest for new samples of the uncertain input parameters (30%, i.e. \(N_{\text{test}} = 225\)). Having a calibration dataset that is independent of the proper training dataset is essential to provide an unbiased estimation of the Gaussian process prior distributions.

3.4.2 Performance metrics

**Individual Gaussian process regression model**

In this work, we quantify the individual GPR model performance using a \(Q^2\) (explained variance) criterion for each POD reduced coefficient. With this criterion, the \(l\)th GPR prediction error is weighted by the variance over the \(m_l\) of the \(l\)th POD mode:

\[
Q^2_l = 1 - \frac{\|k_l - m^*_l\|^2}{\|k_l - \mathbb{E}[k_l]\|^2}, \quad \forall l = 1, \ldots, L,
\]

where \(m^*_l\) is the \(l\)th GPR mean prediction (Eq. 23). In this study, \(Q^2_l\) is estimated over the test dataset. Since we use whitening, reduced coefficients have zero-mean and unit-variance, meaning that this per-mode \(Q^2\) metric directly reflects the mean squared error (MSE) on the predicted reduced coefficients (i.e. \(Q^2 \approx 1 - \text{MSE}\)).

**Reduced-order model**

To quantify the performance of the reduced-order model in the physical space, we also evaluate the \(Q^2\) criterion on each grid point of the domain of interest. For the \(i\)th grid element, the reconstruction/prediction error is weighted by the variance over the LES samples as follows:

\[
Q^2_i = 1 - \frac{\|K_{\text{les},i} - K_{\text{rb},i}\|^2}{\|K_{\text{les},i} - \mathbb{E}[K_{\text{les},i}]\|^2}, \quad \forall i = 1, \ldots, N_i.
\]

In this study, the \(Q^2_i\) criterion is estimated over the LES training dataset for verification as well as over the LES test dataset for evaluating the reduced-order model prediction capacity. To help with the analysis, we also derive a global score from the variance weighted local \(Q^2\) criterion as:

\[
Q^2_{\text{global}} = \sum_{l=1}^{N_l} a_l Q^2_l, \quad a_l = \frac{\tilde{V}(K_{\text{les},l})}{\sum_{j=1}^{N_l} \tilde{V}(K_{\text{les},j})},
\]

where \(\tilde{V}(K_{\text{les},l})\) corresponds to the variance unbiased estimation over the snapshots defined in Eq. 15. This weighted average matches the usual explained variance criterion of the POD, in agreement with the problem of \(\ell_2\)-matrix norm maximization:

\[
Q^2_{\text{global}} = \sum_{l=1}^{L} \sigma_l / \sum_{l=1}^{N_l} \sigma_l.
\]

3.4.3 Learning procedure

The steps for building our POD-GPR reduced-order model can be summarized as follows:

**Training stage (learning)**

1. Extract the POD modes (Eq. 13) and truncate the reduced basis to the first \(L\) modes (Eq. 31) from the training data: \(\{\psi_i\}_{l=1,...,L}\).
2. Calibrate the GPR hyperparameter prior distribution from POD information: \(\{\theta_l\}_{l=1,...,L}\).
3. Estimate the GPR hyperparameter posterior distribution using either MLL or MAP (Eq. 26): \(\{\theta_{l,\text{opt}}\}_{l=1,...,L}\).

**Prediction stage (validation)**

1. Compute the POD reduced coefficients \(\{k_l(\mu^*)\}_{l=1,...,L}\) for the test sample of input parameters \(\mu^*\) (Eqs. 22–23).
2. Perform inverse POD to recover the predicted field of interest \(K_{\text{rb}}^*\) from the POD coefficients for the test sample (Eq. 17).
3. Compare the emulated fields with the reference LES test samples \(K_{\text{les}}^*\) (Eqs. 29–30).

In this work, POD is implemented using the randomized truncated singular value decomposition (Halko et al. 2011) from the scikit-learn library (Pedregosa et al. 2011). As for GPR, the MLL approach is implemented using scikit-learn, while the MAP approach is implemented using GPyTorch (Gardner et al. 2018).
4 Results

In the following, we analyze POD modes and quality of representation (Sect. 4.1) and give details about how to inform GPR models using POD (Sect. 4.2). In Section 4.3, we analyze the POD-GPR model accuracy on average tracer concentration fields using the full dataset. We provide an analysis of the behavior of the GPR models through hyperparameter optimums displayed in the hierarchical order of the POD modes. We then present an extension of the reduced-order model to the turbulent scalar flux (Sect. 4.4). We end this section with a numerical analysis of the reduced-order model robustness with a smaller training dataset (Sect. 4.5).

4.1 Analysis of proper orthogonal decomposition

This section investigates the quality of representation of the ensemble variance as well as the mode spatial structures for time-averaged quantities of interest. We explain the choices that lead to the selection of the number of POD modes $L$. The number of POD modes directly determines the number of GPRs to perform (Eq. 8).

Cumulative explained variance

Several criteria are available in the literature for selecting the number of modes, the vast majority relying on the explained variance metrics evaluated on training data (e.g. the cumulative explained variance, the Kaiser and elbow rules – Jolliffe 2002). Figure 7 shows the cumulative explained variance of the full training dataset (Sect. 3.4.1) decomposed on the POD reduced basis as a function of the $L$ POD modes. The first mode alone contributes to about 50% of the explained variance, the first fifteen modes explain more than 95%. Alternatively, the Kaiser rule applied to 70% of the mean eigenvalue suggests keeping only twenty-five modes; this corresponds to 97.3% of the total ensemble variance. The elbow rule estimates the truncation threshold from the sign change in the eigenvalue second-order derivative: only the first five modes shall be kept following this rule, which corresponds to 85.6% of the total ensemble variance. These different criteria lead to very different truncation choices. A finer analysis of the POD modes is necessary to determine an appropriate truncation level $L$.

Mode interpretation

We now analyze the variance structures carried by the POD modes from the correlation fields (Eq. 14) plotted in the physical space. Figure 8 presents such fields for five out of the first hundred POD modes, which carry patterns that are representative of the whole set of POD modes. We use the three snapshots presented in Fig. 3 to provide insights into the correlation patterns displayed in Fig. 8.

The first correlation fields have spatially widespread structures with horizontally elongated shapes, looking like streaks aligned with the streamwise direction. For the first POD coefficient (Fig. 8a), the correlation field consists of two opposite horizontal layers. High tracer concentrations in the lower part of the domain are unlikely to occur along with high concentrations in altitude – and vice versa – because of the flow horizontal structure. This behavior is found both in the nominal and high-altitude emission source cases (Fig. 3ac). In the nominal case, high tracer concentration values remain localized near the ground. Differently, when the emission source is positioned higher, the plume remains far from the ground and from the obstacle. Indeed, the plume vertical dispersion is primarily controlled by the presence of the obstacle and the vortex shedding resulting from the flow interaction with the obstacle developing in the downstream recirculation region.

The second POD coefficient (Fig. 8b) is linked to the interactions between plume and recirculation areas, depending if the tracer source is upstream or downstream of the obstacle. When the source is located upstream and sufficiently close to the ground, the obstacle constrains the plume dispersion in an accumulation area close to its left boundary. Similarly, when the emission source is located downstream close enough to the obstacle (see example of Fig. 3b), the tracer is trapped in a second recirculation area where large tracer concentration values can be obtained.

The third coefficient (Fig. 8c) describes the area where the plume stops interacting with recirculation areas. This occurs when the emission source is located sufficiently far from the obstacle and from the ground. In that case, the
plume is transported above the obstacle and disperses vertically further downstream because of vortex shedding.

From the tenth POD mode (Fig. 8d), we can see narrower correlation structures that are associated with the near-source advection-dominated dispersion of the tracer by the average flow. The different streaks are directly related to the different source locations of the LES database since upstream emission source locations induce very refined, horizontally-elongated plume wakes. The correlation field of the fiftieth mode (Fig. 8e) highlights that the very high POD coefficients focus on the remaining ensemble variance heterogeneity. This coefficient features very localized bubble structures, which match tracer concentration peaks due to emission source locations present in the POD training database. The small structure values also raise the question of the potential noise induced in the POD modes and coefficients due to the small training data.

This analysis shows that the interpretation of the POD coefficients is strongly related to local plume structures, and thus to the source location.
**Field prediction example**

Figure 9 illustrates through the example of the nominal snapshot (Fig. 3a), the capacity of reconstructing the average tracer concentration fields from POD for different truncation levels $L$. Only accounting for the very first modes as suggested by the elbow rule ($L = 5$) provides a good representation of highly dispersed tracer areas downstream and of the tracer accumulation in the recirculation areas near the obstacle (Fig. 9d). However, it does not handle well sharp patterns resulting from tracer advection-dominated dispersion near the emission source. Moreover, the peak emission of the source is underestimated and not well located. The Kaiser rule (Fig. 9c) partially reconstructs the wake structures by including more modes ($L = 25$) but cannot correctly recover the plume structure close to the emission source. When including up to $L = 100$ modes in the reduced basis (Fig. 9b), improvements in the field reconstruction mainly relate to the intensity and location of the source peak emission as well as to the localized structures around the emission source.

To conclude this section on POD, high-order modes ($l \geq 50$ for instance) contain rich information on the near-source physics embedded in the LES snapshots. This large number of modes relates to the specificity of our problem, as changing the source location induces very sharp and localized plume structures near the emission source in the ensemble. For this reason, different snapshots might be very poorly correlated around the obstacle. It is essential to keep a large number of POD modes to well represent local spatial concentration structures. We therefore choose at this stage, to keep $L = 100$ POD modes in the reduced basis that represent 99.6% of the explained variance (Fig. 7). Using this reduced basis, we investigate how to efficiently
tune a GPR to map POD coefficients of verage tracer concentration fields from any set of uncertain parameters.

4.2 Gaussian process hyperparameter prior distribution

GPR requires optimizing hyperparameters \( \theta_l = \{ \lambda_1^l, \lambda_2^l, \lambda_3^l, \ldots \} \) for the \( l \)-th reduced-order model, which is carried out by gradient descent on the marginal log-likelihood (MLL) or on the posterior distribution (MAP). In the case of MAP, prior distributions are no longer assumed to be uniform (Sect. 3.3.3). This section presents how to calibrate Gamma and Gaussian distributions for the hyperparameters from POD modes and reduced-coefficient features that are required as input to the MAP optimization procedure.

Noise prior

POD often assumes that the variability of the low-order reduced coefficients is related to systematic behavior among the LES dataset, whereas the variability carried by the high-order reduced coefficients measures noise (Jolliffe 2002). This is equivalent to treating the reduced coefficients on the first modes as unbiased data, which is a very restrictive assumption (Sect. 3.3). In this study, GPR accounts for the noise on the \( l \)-th reduced coefficient through the hyperparameter \( s_l^2 \) (Eq. 19). We assume that a prior estimate of \( s_l^2 \) may be obtained from the noise introduced by the time-averaging process performed on the LES (Sect. 2.6.2). This may not be the only source of noise introduced during the LES database generation, but this provides a lower bound noise estimation that is useful to demonstrate the added value of our methodology.

The main parameter involved in the time-averaging process is the length of the time-averaging window used to acquire LES statistics. To evaluate the noise for the \( l \)-th reduced-order model, we compare the reduced coefficient obtained by averaging over only the first (respectively the last) 50% of the full simulation time window \( k_{l,0\rightarrow50\%} \) (respectively \( k_{l,50\rightarrow100\%} \)),

\[
s_l^2 \approx \frac{1}{2N} \sum_{n=1}^{N} \left( k_{l,50\rightarrow100\%}(\mu^{(n)}) - k_{l,0\rightarrow50\%}(\mu^{(n)}) \right)^2. \tag{32}
\]

Figure 10 shows the \( s_l^2 \)-estimates obtained on the calibration samples (thin line). The estimated noise is small compared to the variability of the reduced coefficients (normalized to unity) but varies over several orders of magnitude. For this reason, since \( s_l^2 \) is positive, we set a Gamma prior distribution for \( s_l^2 \), whose mode matches the regressed estimate for each of the \( L \) reduced-order models (thick line in Fig. 10).

In addition, we assume POD has extracted most of the noise from the data so that noise magnitude tends to be small compared to the total variability of the reduced coefficients. Since reduced coefficients are normalized to unit-variance, we assume that \( 0 \leq s_l^2 \leq 1 \). The assumption \( s_l^2 < 1 \) strongly depends on the quality of the training data. The Gamma distribution only applies a soft positivity constraint on \( s_l^2 \)-distribution; Bayesian estimation may eventually lead to \( s_l^2 > 1 \) if there are not enough training data.

Finally, Gamma distribution parameters are inferred from the mode estimate (thick line in Fig. 10) and the Gamma mean (set arbitrarily to 0.5, the mid-value of the assumed variation interval for \( s_l^2 \), although such choice may be worth further investigation). Note that for the hyperparameter optimization step, the starting point of the gradient descent for \( s_l^2 \) is taken as the Gamma distribution mode.

Mean and scaling priors

Now we establish prior information on the mean and variance of the Gaussian processes (Eq. 20). We assume a constant mean and that the variance is decomposed into a systematic component and a noise component. From the expression of the noisy GPR models, we obtain for each POD reduced coefficient \( l \):

\[
\mathbb{E}[k_l] = m_l, \quad \text{Var}(k_l) = \varphi_l + s_l^2, \tag{33}
\]

where the signal variance hyperparameter \( \varphi_l \) (Eq. 24) characterizes systematic variability, while \( s_l^2 \) characterizes noise variability.

Figure 11 shows the variance statistics of the reduced coefficients on the different LES datasets (Sect. 3.4.1). The variance evaluated on the training data is equal to one (this is expected due to whitening). This is no longer the case.

![Fig. 10 Estimation of the noise hyperparameter \( s_l^2 \) for each reduced-order model \( l \). The thin line corresponds to the noise estimation \( s_l^2 \) from the POD modes (Eq. 32). The thick line corresponds to the average trend found by regression and is defined by the following equation \( s_l^2 = 2.16 \times 10^{-4} \beta^{0.93} \).](image)
Correlation length-scale prior

We are now interested in the prior distributions of the RBF kernel correlation length-scales \( \{ \lambda_{l,xsrc}, \lambda_{l,zsrc}, \lambda_{l,uzc}, \lambda_{l,z0} \} \) (Sect. 3.3.2).

The analysis of the POD correlation fields (Fig. 8) reveals an increase in small-scale spatial heterogeneity for increasingly higher-order modes driven by the emission source location. Recall that the first correlation fields include horizontally elongated shapes looking like streaks aligned with the streamwise direction, and that when the POD mode index increases, the number of alternated streaks increases, and these streaks get narrower along the vertical direction. This implies that the correlation length-scales \( \lambda_{l,xsrc} \) and \( \lambda_{l,zsrc} \) associated with \( x_{src} \) and \( z_{src} \) are strongly influenced by these patterns. In particular, the typical correlation length along the source height \( \lambda_{l,zsrc} \) should be related to the number of streaks, and should therefore decrease as we consider higher POD modes. We therefore model a decrease in the length-scales \( \lambda_{l,xsrc} \) and \( \lambda_{l,zsrc} \) when moving to higher-order modes. Because the correlation length-scales are positive, we adopt a prior Gamma distribution for \( \lambda_{l,xsrc} \) and \( \lambda_{l,zsrc} \). The mode of each Gamma distribution is determined by the simple decreasing rule: \( 1/l \) for \( l = 1, \ldots, 100 \). The variance of the Gamma distribution is set to 1, which corresponds to the interval length associated with the normalized input parameters \([0, 1]^4\), Sect. 2.6.2).

Since there is no analogous interpretation of decreasing length-scales for \( z_{0} \) and \( u_{z} \), we retain prior Gamma distributions with constant mode and variance equal to 1 for \( \lambda_{l,u_{z}} \) and \( \lambda_{l,z0} \). Note that during the hyperparameter optimization step, the starting point of the gradient descent for \( \{ \lambda_{l,xsrc}, \lambda_{l,zsrc}, \lambda_{l,uzc}, \lambda_{l,z0} \} \) is set to the Gamma distribution mode.

To conclude this section, it is possible to define prior distribution for the GPR hyperparameters based on POD information. This information will then be used as a starting point for the gradient descent in the MAP optimization step.

4.3 Gaussian process regression performance

In this section, we compare the POD-GPR reduced-order model solutions obtained from MLL and MAP hyperparameter optimization procedures to show the added value of MAP. We also quantify the noise of the coefficients on each POD mode and carry out a performance analysis using the \( Q^2 \) metrics.

Comparison of optimized solutions

Figure 12 presents the modes of the hyperparameter prior probability distribution function (dashed lines) as well as the optimized hyperparameter solutions obtained with MAP (dotted lines) to compare with MLL (solid lines). Recall that MAP takes as input the GPR hyperparameter prior information (Sect. 4.2).

Figure 12abcd shows that MAP and MLL procedures converge to similar hyperparameter solutions. Estimated correlation length-scales are close to each other. Still, MAP estimates are systematically underpredicted compared to MLL solutions on the first fifty modes.

Noise estimates in Fig. 12ef are consistent for almost all POD modes. Only the estimates for the variance
hyperparameter $q_l$ differ between MLL and MAP procedures. The MLL procedure is not constrained by the prior distribution, and the $q_l$-estimates are free to converge far from the starting point value $q_l = 1$. Concerning MAP optimization, the prior $q_l$-distribution incorporates a soft constraint that causes $q_l$ to diverge only if the benefit on the posterior maximization is significant. Since the MAP procedure converges to values close to the prior distribution mode, this suggests that the scaling value has a negligible impact on the posterior maximization.

In the end, the difference between the MLL and MAP results mainly concerns $q_l$ on the first fifty POD modes and consequent length-scales. Nevertheless, since noise is negligible on the first modes (because $q_l \gg s_l^2$), the associated Gaussian process mean predictions are almost equal (Eq. 22). Hence, MLL and MAP procedures converge towards similar optima for the GPR hyperparameters, meaning that they correspond to equivalent predictive models but with an economy concerning numerical costs for MAP compared to MLL: the coarse prior distributions are sufficient to ensure MAP convergence in a single gradient descent when fifteen iterations are required for MLL (this factor of fifteen is all the more important as this optimization procedure is repeated for each of the $L = 100$ GPR models).

**Gaussian process noise analysis**

Figure 12f shows that the prior $s_l^2$-estimation is close to the optimal values obtained from both MLL and MAP procedures. This suggests that the temporal convergence error is the primary source of noise in the data. It also shows that there is a gradual increase of noise with increasing mode index. On the first POD modes, the noise is small compared to the signal variance. This is consistent with the fact that the first POD modes are almost noise free since they carry large variance structures. But the noise increases continuously on high-order modes such that it can be substantial compared to the signal variance (the maximum noise estimate is close to 0.2 for very high modes, which is of the same order of magnitude as the signal variance that is around 1). This suggests that there are not two distinct behaviors among the modes, namely those that carry systematic information on the one hand and those that carry noise on the other hand. The noise magnitude carried by each mode tends to grow continuously when sorted hierarchically and explains why there is not an obvious choice for POD truncation.
Correlation length-scale analysis

In terms of correlation length-scales, we observe in Fig. 12abcd that the correlation length-scales on inflow parameters $\lambda_{x_{\text{src}}}$ and $\lambda_{z_{\text{src}}}$ are larger than the correlation length-scales on position parameters $\lambda_{x_{\text{src}}}$ and $\lambda_{z_{\text{src}}}$. Furthermore, the gap increases when moving to high-order modes. Indeed, $\lambda_{x_{\text{src}}}$ and $\lambda_{z_{\text{src}}}$ decrease in a stable manner over the first hundred POD modes as in the prior solutions. This suggests that the variation of the reduced coefficients is mostly due to variations in $x_{\text{src}}$ and $z_{\text{src}}$ and that the higher-order reduced coefficients become more sensitive to small variations in source position and height while being relatively insensitive to $u_{z_c}$ and $z_0$. This explains why the optimization process for $\lambda_{x_{\text{src}}}$ and $\lambda_{z_{\text{src}}}$ becomes unstable for high-order POD modes (Fig. 12ab). This is also consistent with the small structures observed in the high-order modes and associated with tracer concentration wakes (Sect. 4.1).

Gaussian process regression accuracy

We now evaluate the accuracy of the GPR models over the test dataset for validation purposes. Figure 13 shows the $Q^2$-scores obtained for each POD mode for the MLL and MAP procedures. For the very first modes, the $Q^2$ score is close to one, which means that GPR models perform almost perfectly on large variance structures with limited overfitting. However, they cannot maintain this level of performance when moving to high-order modes that feature more complex and localized structures. The decrease in performance appears linear, and the higher noisier modes are more difficult to predict (with a $Q^2$ coefficient evolving between 50% and 60% from the eightieth mode onwards).

Results show that the adaptive procedure for GPR remains successful in predicting fine structures of the LES ensemble variance. Optimizing the hyperparameters’ mode per mode greatly improves the GPR accuracy compared to simply imposing a prior trend on hyperparameters. This gain in $Q^2$ is more pronounced when moving to high-order modes. Results also confirm that the reduced-order models obtained from MLL and MAP procedures are equivalent as anticipated from the equivalent correlation length-scales in Fig. 12. Globally, the $Q^2$-score of the reduced-order model evaluated on the test dataset is equal to 97.7%. This model slightly overfits data since the $Q^2$-score is equal to 99.3% on the training dataset (recall that the reduced-order model cannot achieve a global $Q^2$-score beyond 99.6%, which is the upper limit imposed by the POD reconstruction error for $L = 100$—Fig. 7). Note that the $\ell^2$-hyperparameter slightly reduces accuracy on the training dataset but this is not of the same magnitude as the decrease in performance observed on the test dataset. This hyperparameter is therefore not responsible for the observed regression error.

Field prediction examples

Figures 14, 15 and 16 show the average normalized tracer concentration fields predicted by the reduced-order model (resulting from MAP optimization) for three snapshots of the LES test dataset (Fig. 3).

For the nominal snapshot (Fig. 14), the largest prediction errors are made (i) close to the emission source with largely underestimated tracer concentration (about 50% of the LES reference concentration), and (ii) in the accumulation area upstream of the obstacle with overestimated coarser-structured tracer concentration levels. Further away from the source, the tracer concentration in the wake of the obstacle is well predicted. Upstream of the obstacle, we can distinguish slight noise in no-tracer areas due to high-order POD modes that carry small noisy structures (Fig. 8).

We also present the prediction result for a case where the emission source is far from the obstacle and the ground (Fig. 15). This case emphasizes what has already been observed in the nominal case. The tracer concentration at the actual emission source and along the wake is hard to predict. The prediction absolute error can reach up to 77% of the LES solution. The fine structures of the plume are hard to predict since information is mostly carried by multiple high-order POD modes in a region where there are only a few samples of the emission source in the LES dataset (this case is located at the boundary of the tracer source area—Fig. 1).

To complement the analysis, we finally present the prediction result for a case where the emission source is located in the recirculation area downstream of the obstacle (Fig. 16). This case is much better predicted by the reduced-order model. The areas of high tracer concentrations form a wide-spread structure in areas carried by the first POD modes. Most of the information can be therefore conveniently recovered from the first reduced coefficients with accurate GPR models, including in the recirculation area downstream of the obstacle. The largest prediction error is on the order of 10% of the LES solution.
These three snapshots highlight the capacity of the reduced-order model to predict the main tracer concentration structures, in particular when the emission source is located in recirculation zones near the obstacle. In this situation, the tracer concentration patterns are smooth as the dispersion is dominated by turbulent diffusion, making the prediction task easier. It is now interesting to evaluate the capacity of the reduced-order model to handle more LES field statistics and in particular coupled LES statistics.

### 4.4 Application to turbulent scalar fluxes

To better exploit the explicit resolution of turbulence by the LES approach (compared to other approaches such as RANS for instance), it is possible to rely on a similar reduced-order modeling approach for higher-order statistical quantities. In the context of dispersion, the turbulent scalar flux is a key quantity that is fully modeled in low-fidelity dispersion models (Vervecken et al. 2015b) but can be explicitly simulated using LES (see nominal snapshot example in Fig. 4). In this section, the POD-GPR reduced-order model is therefore extended to the resolved vertical turbulent scalar flux using the full training dataset.

As for the average quantities of interest, the maximum performance of the POD-GPR model on the test dataset is obtained for $L = 100$ modes. It achieves a global $Q^2$-score equal to 93.9% (compared to 97.7% for the average tracer concentration). This lower $Q^2$ performance can be explained by the finer structures of the vertical turbulent scalar flux, which make the prediction process slightly more complex than for the average quantities of interest. This is not due to the hyperparameter trends (not shown),

---

**Fig. 14** Normalized average tracer concentration field obtained for the nominal snapshot with a reduced-order model prediction (white lines correspond to the $5 \times 10^{-4}$-contour line to highlight the presence of low-magnitude noisy structures). **b** Prediction absolute error calculated with respect to the LES snapshot (Fig. 3a).

**Fig. 15** Same caption as in Fig. 14 for a LES snapshot with a high emission source (Fig. 3c).
which are similar for the average tracer concentration and the vertical turbulent scalar flux. However, there is a wider spread of information on the high POD modes for the vertical turbulent scalar flux. Indeed, the POD variance is spread over a slightly larger number of POD modes: one hundred modes carry 99.3% of the ensemble variance for the vertical turbulent scalar flux (compared to 99.6% for the average tracer concentration); and the Kaiser rule (Sect. 4.1) suggests to keep around 32 modes instead of 25 modes for the average quantities of interest.

As an illustration, Fig. 17 shows the vertical turbulent scalar flux prediction for the nominal snapshot (Fig. 17a) and the associated absolute error map with respect to the LES solution (Fig. 17b). We can observe the sharper localized structure compared to the associated average concentration field (Fig. 14a), with upward (positive) turbulent scalar flux in a large part of the plume and downward (negative) flux in the wake of the obstacle. The worst areas of predictions are again located upstream of the obstacle close to the source. Still, the sign and magnitude of the turbulent flux is well predicted by the reduced-order model, even in localized areas of strong dispersion, such as the windward corner of the obstacle.

This highlights that our POD-GPR reduced-order modeling approach is versatile and has the capacity to emulate a variety of LES field statistics depending on the context of microscale air pollutant assessment: it can perform direct prediction of the tracer concentration, or it can be used to extract physical quantities related to the phenomena that are directly relevant for the plume spread modeling such as turbulent scalar fluxes.
All the results presented so far were obtained by considering a large number of modes in the reduced basis with a rich training dataset ($N_{\text{train}} = 472$). Such a large LES database is unaffordable in all realistic atmospheric dispersion cases involving larger computational domains and three-dimensional effects. We now evaluate how much the reduced-order model accuracy degrades when the size of the training dataset is reduced to have a better idea of the minimum required budget for building a multi-query uncertainty quantification framework.

### 4.5 Robustness to training dataset

We now reduce the training dataset to one hundred snapshots ($N_{\text{train}} = 100$) and fifty snapshots ($N_{\text{train}} = 50$) to move towards a more realistic multi-query LES framework. Note that 90% of these training snapshots are used to determine the POD reduced basis, with the remaining 10% used for calibration as before. This implies that the maximum number of modes in the reduced basis is directly equal to the POD training size (90 for $N_{\text{train}} = 100$; 45 for $N_{\text{train}} = 50$). The only difference is that here the whole dataset is used to optimize the GPR models since the dataset is of very limited size. Note also that the test dataset remains the same as before (as described in Sect. 3.4.1) to avoid introducing bias during the validation stage.

Figure 18 compares the evolution of the per-mode $Q^2$-score for both full and reduced training datasets. The reduced-order model accuracy significantly decreases for the reduced database. There is a faster linear decrease towards $Q^2 = 0$ than with the full training dataset (the threshold $Q^2 = 0$ is approximately reached for the fiftieth mode for $N_{\text{train}} = 100$ and the twenty-fifth mode for $N_{\text{train}} = 50$). This suggests that the number of POD modes to consider in the reduced-order models should be reduced due to the too-limited size of the training dataset.

Including higher-order modes in the reduced-order model can even degrade its global performance. Here, the optimal choice is to keep $L = 47$ modes for $N_{\text{train}} = 100$, and $L = 30$ modes for $N_{\text{train}} = 50$. The associated global $Q^2$-score is equal to 91.8% and 83.5%, respectively, and can be compared to 97.7% for the full training dataset (Sect. 4.3). These scores may seem satisfactory but when looking at the nominal snapshot prediction, the prediction errors observed before are amplified. Figure 19 presents the field prediction result for $N_{\text{train}} = 100$. The shape of the tracer concentration wake is retrieved but the reduced-order model has difficulty predicting the correct tracer concentration magnitude near the emission source: non-physical noisy structures appear in tracer-free regions upstream of the emission source location (white lines); the tracer concentration is underestimated at the emission source and upstream of the obstacle (the largest prediction error is more than 50% of the LES solution value in this region); and the high tracer concentrations correspond to a much thinner region than for the full training dataset (see Fig. 14).

This highlights that the analysis of the global $Q^2$-score can be misleading about the reduced-order model accuracy since the prediction quality is spatially heterogeneous and can fastly degrade upstream of the obstacle due to the large tracer concentration gradients near the emission source. Even if the prediction performance is reduced when the training dataset includes a hundred snapshots, the reduced-order model predictions remain physically consistent. However, physical consistency is lost upstream of the obstacle when further reducing the training dataset (not shown): spurious structures are strengthened, the low concentration magnitude contour lines are much coarser, and this can go as far as having no clear emission source in the emulated fields. Hence, based on this work, our recommendation is to use a minimum budget of at least a hundred snapshots to emulate parameterized LES of urban pollutant dispersion when the location of the emission source is part of the uncertain parameters.

### 5 Conclusion

The goal of this study was to – propose a purely data-driven, non-intrusive reduced-order model combining POD and GPR for predicting turbulent field statistics of interest associated with near-source tracer concentration dispersion in a multi-query uncertainty quantification context; and – to validate accuracy and interpretability of the results, thanks to the construction of a solid LES database. A two-dimensional case study corresponding to a turbulent atmospheric flow over a surface-mounted obstacle was considered to generate a LES ensemble based on perturbed inflow boundary conditions and emission source location,
and to provide guidelines for future field-scale dispersion applications.

The specific patterns observed in the POD modes and components were found to translate into the GPR hyper-parameter distributions. We demonstrated that POD information can be used to perform an efficient and robust Bayesian mode-per-mode optimization of GPR hyperparameters by designing non-uniform prior distributions. Such a procedure was found to achieve similar accuracy than the standard N-restart MLL maximization approach from a single gradient descent, substantially accelerating the offline training stage and proving the reliability of the reduced-order model solutions. These results offer an interesting preliminary avenue for improving the scalability of GPR for high-dimensional problems, whose training cost frequently justifies its exclusion from the list of potential learning candidates.

The POD-GPR emulation process was found to accurately reproduce the LES predictions while reducing the computational cost to query a snapshot for a new set of parameters by several orders of magnitude. Concentration levels in the recirculation areas were well predicted by the POD-GPR model but it was more challenging to properly characterize the fine plume structures in the near-source regions caused by the uncertainty in the emission source location. POD analysis revealed that these structures were carried out by high-order POD modes and that it was therefore necessary to consider a large number of modes (hundred modes for the full training database) to accurately emulate near-source plume structures. This explains why a large LES training dataset was required to achieve accurate emulation without artifacts.

By reducing the number of training snapshots, a loss of consistency with physics principles was observed. For instance, non-physical noisy structures could appear in tracer-free regions. Still, the POD-GPR model prediction performance remained acceptable when considering at least a hundred LES snapshots in the training database, providing a first budget estimate for field-scale applications. Moreover, the POD-GPR approach is also applicable to high-order statistics such as turbulent scalar fluxes that are relevant for the RANS formalism for instance. This illustrated that LES high-fidelity solutions could be used to improve the accuracy of lower-fidelity dispersion models, for instance by providing better turbulence closure terms, which could be valuable for practical dispersion applications (Grylls et al. 2019).

In terms of perspectives, deep-learning-inspired techniques such as autoencoders offer an intriguing line of study for overcoming POD limitations, improving compression performance in presence of substantial nonlinearity (Fukami et al. 2020; Fresca and Manzoni 2022), and meeting the constraint of reduced training database. Future work also involves moving beyond a purely data-driven approach not constrained by physics principles and introducing physical constraints in the learning process to reduce prediction artifacts in case of small training data. From an application viewpoint, it is of primary interest to extend the approach to more realistic atmospheric boundary-layer flows for different stability conditions to evaluate its potential for air pollutant dispersion applications, and to exploit the posterior Gaussian distribution predicted by GPR for uncertainty quantification in a risk assessment.
context. In the long term, a reduced-order model informed by LES could be valuable to produce ensemble forecasts and assess human exposure to toxic air pollutants in the event of an accident.

**Acknowledgements** The authors acknowledge Amandine Marrel (CEA) and Lionel Soulhac (LMFA/INSA Lyon) for fruitful discussions on dimension reduction/Gaussian process regression and on dispersion modeling, respectively. They also acknowledge Thibault Gioud (Cerfacs) for support on the LEMMINGS Python package for managing ensemble LES runs, and Isabelle D’Ast from the Computer Support Group at Cerfacs for support on GPyTorch and TensorFlow library installations.

**Author Contributions** All authors contributed to the study’s conception, design, and analysis. Material preparation and data collection were performed by BXN and TJ. The first draft of the manuscript was written by BXN and MCR, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** Funding and access to supercomputing resources for this work were provided by CERFACS.

**Declarations**

**Conflict of interest** All authors have no competing interests to declare that are relevant to the content of this article.

**References**

Afkham BM, Hesthavn JS (2017) Structure preserving model reduction of parametric hamiltonian systems. SIAM J Sci Comput 39(6):A2616–A2644

Allwine KJ, Flaherty JE (2006) Joint Urban 2003: study overview and instrument locations. Tech. rep., Pacific Northwest National Laboratory (PNNL), Richland, WA, USA, https://doi.org/10.2172/890732

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

van den Bos L, Sandezer B (2017) Uncertainty quantification for wind energy applications–literature review. Tech. Rep. SC-170, Centrum Wiskunde & Informatica, Amsterdam (CWI), The Netherlands

Colin O, Rudyard M (2000) Development of high-order Taylor-Galerkin schemes for LES. J Comput Phys 162(2):338–371. https://doi.org/10.1006/jcph.2000.6538

Da Silva FT, Reis NC Jr, Santos JM et al (2021) The impact of urban block typology on pollutant dispersion. J Wind Eng Ind Aerodyn 210(104):524. https://doi.org/10.1016/j.jweia.2021.104524

Dauxois T, Peacock T, Bauer P et al (2021) Confronting grand challenges in environmental fluid mechanics. Phys Rev Fluids 6(020):501. https://doi.org/10.1103/PhysRevFluids.6.020501

Daviller G, Oztarlik G, Poinot T (2019) A generalized non-reflecting inlet boundary condition for steady and forced compressible flows with injection of vortical and acoustic waves. Comput Fluids 190:503–513. https://doi.org/10.1016/j.compfluid.2019.06.027

Du Y, Blocken B, Pirker S (2020) A novel approach to simulate pollutant dispersion in the built environment: transport-based recurrence CFD. Build Environ 170(106):604. https://doi.org/10.1016/j.buildenv.2019.106004

El Garroussi S, Ricci S, De Lozzo M et al (2022) Tackling random fields nonlinearities with unsupervised clustering of polynomial chaos expansion in latent space: application to global sensitivity analysis of river flooding. Stoch Environ Res Risk Assess 36:693–718. https://doi.org/10.1007/s00477-021-02060-7

Franke J, Hellsten A, Schlunzen KH et al (2011) The COST 732 best practice guideline for CFD simulation of flows in the urban environment: a summary. Int J Environ Pollut 44(1–4):419–427

Fresca S, Manzoni A (2022) Pod-dl-rom: enhancing deep learning-based reduced order models for nonlinear parametrized pdes by proper orthogonal decomposition. Comput Methods Appl Mech Eng 388(114):181. https://doi.org/10.1016/j.cma.2021.114181

Fukami K, Nakamura T, Fukagata K (2020) Convolutional neural network based hierarchical autoencoder for nonlinear mode decomposition of fluid field data. Phys Fluids 32(9):095,110. https://doi.org/10.1063/5.0020721

Gamell H (2015) Caracterización experimentale de l’écoulement et de la dispersion autour d’un obstacle bidimensionnel. PhD thesis, Ecole centrale de Lyon, France

García-Sánchez C, Philips D, Gorlé C (2014) Quantifying inflow uncertainties for CFD simulations of the flow in downtown Oklahoma City. Build Environ 78:118–129. https://doi.org/10.1016/j.buildenv.2014.04.013

García-Sánchez C, Tendeloo GV, Gorlé C (2017) Quantifying inflow uncertainties in runs simulations of urban pollutant dispersion. Atmos Environ 161:263–273. https://doi.org/10.1016/j.atmosenv.2017.04.019

García-Sánchez C, van Beeck J, Gorlé C (2018) Predictive large eddy simulations for urban flows: challenges and opportunities. Build Environ 139:146–156

Gardner J, Pleiss G, Weinberger KQ, et al (2018) Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. Adv Neural Inf Process Syst 31

Gicquel L, Gourdain N, Boussuge JF et al (2017) Quantifying inflow conditions. Phys Fluids 20(7):077,102

Grylls T, Cornec CML, Salizzoni P et al (2019) Evaluation of an operational air quality model using large-eddy simulation. Atmos Environ 3(100):041. https://doi.org/10.1016/j.aeaoa.2019.100041

Guo M, Hesthavn JS (2018) Reduced order modeling for nonlinear structural analysis using Gaussian process regression. Comput Methods Appl Mech Eng 341:807–826. https://doi.org/10.1016/j.cma.2018.07.017

Halko N, Martinsson PG, Tropp JA (2011) Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev 53(2):217–288. https://doi.org/10.1137/090771806

Hesthavn JS, Ubbiali S (2018) Non-intrusive reduced order modeling of nonlinear problems using neural networks. J Comput Phys 363:55–78. https://doi.org/10.1016/j.jcp.2018.02.037

Jolliffe IT (2002) Principal component analysis. Springer series in statistics, Berlin, p 29

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

Ko J, Lucor D, Sagaut P (2008) Sensitivity of two-dimensional spatially developing mixing layers with respect to uncertain inflow conditions. Phys Fluids 20(7):077,102
