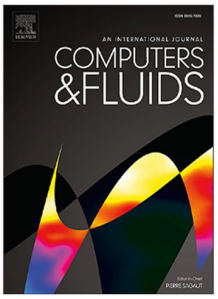


Augmented state estimation of urban settings using on-the-fly sequential Data Assimilation

L. Villanueva, M.M. Valero, A. Šarkić Glumac, M. Meldi

PII: S0045-7930(23)00343-2
DOI: https://doi.org/10.1016/j.compfluid.2023.106118
Reference: CAF 106118

To appear in: *Computers and Fluids*

Received date : 31 March 2023
Revised date : 26 September 2023
Accepted date : 8 November 2023

Please cite this article as: L. Villanueva, M.M. Valero, A.Š. Glumac et al., Augmented state estimation of urban settings using on-the-fly sequential Data Assimilation. *Computers and Fluids* (2023), doi: https://doi.org/10.1016/j.compfluid.2023.106118.

Augmented state estimation of urban settings using on-the-fly sequential Data Assimilation

L. Villanueva[a,*], M. M. Valero[b], A. Šarkić Glumac[c], M. Meldi[b]

[a] *Institut Pprime, CNRS - ISAE-ENSMA - Université de Poitiers, 11 Bd. Marie et Pierre Curie, Site du Futuroscope, TSA 41123, 86073 Poitiers Cedex 9, France*
[b] *Univ. Lille, CNRS, ONERA, Arts et Métiers ParisTech, Centrale Lille, UMR 9014- LMFL- Laboratoire de Mécanique des fluides de Lille - Kampé de Feriet, F-59000 Lille, France*
[c] *University of Luxembourg, Interdisciplinary Centre for Security, Reliability and Trust (SnT), 6 Avenue de la Fonte, Esch-sur-Alzette, 4364 Luxembourg*

Abstract

A data-driven investigation of the flow around a high-rise building is performed by combining heterogeneous experimental samples and numerical models based on the Reynolds-Averaged Navier–Stokes (RANS) equations. The experimental data, which include velocity and pressure measurements obtained by local and sparse sensors, replicate realistic conditions of future automated urban settings. The coupling between experiments and the numerical model is performed using techniques based on the Ensemble Kalman Filter (EnKF), including advanced manipulations such as localization and inflation. The augmented state estimation obtained via EnKF has also been employed to improve the predictive features of the RANS model via optimization of the free global model constants of two turbulence models used to close the equations, namely the $\mathcal{K}-\varepsilon$ and the $\mathcal{K}-\omega$ SST turbulence models. The optimized inferred values are far from the classical values prescribed as general recommendations and implemented in codes, but also different from other data-driven analyses reported in the literature. The results obtained with this new optimized parametric description show a global improvement for both the velocity and pressure fields. In addition, some topological improvements for the flow organization are observed downstream, far from the location of the sensors.

*Corresponding author
Email address: lucas.villanueva@ensma.fr (L. Villanueva)

Keywords: Urban settings, Data Assimilation, EnKF, CONES

1. Introduction

Among the open challenges in the field of fluid mechanics, the accurate prediction and control of the turbulent flows around bluff bodies is a timely subject in the current development of automated, data-informed urban areas. Bluff bodies are characterized by massive separation of the flow at high Reynolds numbers, which is responsible for the emergence of large, energetic wakes [1]. The global aerodynamic interactions are, in this case, characterized by complex concurring phenomena such as shear layers, flow separation, and reattachment and recirculation regions. Predicting such features is a complex task, owing to the extensive range of active dynamic scales that can be observed in fully developed turbulence. Computational resources required to completely represent turbulent flows via direct numerical simulation are prohibitive for Reynolds numbers observed in realistic applications dealing with urban settings. Reduced-order Computational Fluid Dynamics (CFD) such as models based on the Reynolds-Averaged Navier–Stokes (RANS) equations [2, 3] can provide a statistical description of complete urban areas with affordable resources, but the accuracy of such prediction is strongly affected by the features of the turbulence model needed to close such dynamic equations. The models, which are driven by a number of coefficients classically determined via empiric approaches, usually fail to represent interactions of different physical phenomena triggered by turbulence, such as the ones previously listed, which are observed in urban settings. Experimental approaches, which rely on measurement that can be obtained by various techniques, such as pressure sensors and hot wires, can provide a virtually exact characterization of the flow features in the form of pressure and velocity measurements. However, experimental data may be local in space and time, and a full-volume representation of flows is prohibitively expensive.

Studies in the last decades have tried to create a solid network between numerical simulation and experiments in order to exploit the intrinsic advantages of both methods. A rigorous mathematical background is provided by tools from Estimation Theory [4], which is a branch of statistics. Among the numerous methods available in the literature, Data Assimilation (DA) [5] is a vast family of tools tailored to combine experimental and numerical data to obtain a more accurate prediction of the flow. Sequential DA uses tools

from probability and statistics to target physical states with minimized uncertainty (state estimation), once different sources of information and their related level of confidence are provided. Among these methods, one can include the Kalman Filter (KF) [6] and its ensemble version, the Ensemble Kalman Filter (EnKF) [7, 8], which is arguably among the most powerful tools available in the literature for DA.

Tools from Estimation Theory can also exploit available information and state estimation to train or optimize an underlying model with the aim of better performing in operative conditions where reference data is unavailable. This task has been extensively performed in the literature to improve the predictive capabilities of reduced order models for CFD, particularly RANS [9]. Examples dealing with Uncertainty Propagation [10, 11], Data Assimilation [12, 13] and also Machine Learning [14, 15, 16, 17] are available in the literature. These works show that improved prediction accuracy using RANS modeling can be obtained. However, two main difficulties are observed for these applications. First, the optimization and training of the models usually demand an extensive spatially well-distributed set of data. Second, the models may underperform when used for extrapolation i.e. predict flow features that are different from those observed in the available data used for optimization/training. This last point is crucial for accurately predicting the statistical moments of turbulent flows, and it is an open challenge.

As previously mentioned, realistic configurations in urban settings exhibit non-linear interaction of several complex phenomena that cannot be isolated [18]. Among these, the well-known test case of the high-rise building [19] shows such features despite its simple geometry. The complex flow topology produced by the concurring effects of separation of boundary layers, instabilities triggered by shear effects, turbulent wakes, and recirculation regions is challenging to capture accurately even by advanced data-driven strategies. Among the work reported in the literature for this test case, Ben-Ali et al. [20] and Zhao et al. [21] propose an improvement for classical RANS models using experimental data for the data-driven procedures. This kind of application is still rare in the literature for fluid mechanics as most of the analyses in the literature rely on the usage of models of different resolution. Classical applications usually optimize reduced-order models such as RANS using high-fidelity data from direct numerical simulation. The main advantage of using numerical data relies on the complete control of boundary and initial conditions, which allows for excluding bias in fundamental parameters such as the mass flow rate. However, model optimization for realistic operating

conditions cannot bypass the usage of experimental data, and it is a key step to be unlocked towards fully automated digital twin applications [22]. Ben-Ali et al. [20] inferred the behavior of turbulence modeling in the open-source platform OpenFOAM using experimental pressure measurements on the surface of the building. They developed an adjoint method to perform such optimization. Their results showed that the accuracy of the RANS model could be improved. However, this improvement was associated with the development of advanced numerical techniques, including modification of the dynamic equations for the turbulence physical quantities (turbulent kinetic energy $\mathcal{K}$, and energy dissipation rate ε or specific energy dissipation rate ω). Optimizing global model constants, which are accessible to a standard user of the code, would not produce significant improvements in this case. On the other hand, Zhao et al. [21] used an EnKF code to infer the behavior of such model constants, always using the software OpenFOAM. In order to perform the optimization, they employed time-averaged velocity measurements available on 230 sensors sparsely distributed in the flow volume around the building. In this case, they obtained an efficient calibration of the model constants. However, such a global distribution of data is difficult to obtain in realistic conditions, where sensors are mostly clustered in local regions where measurements can be efficiently performed.

Another important aspect to consider is the influence of sensor placement on the performance of the data assimilation algorithm. In the context of variational data assimilation applied to the unsteady flows past a rotationally oscillating cylinder, Mons et al. [23] addressed this challenge known as sensor placement problem. They proposed a first-order adjoint-based procedure to maximize the sensitivity of the observations with respect to changes in initial and boundary conditions. A more comprehensive framework is presented in a separate work [24] where besides linear sensor placement approaches, second-order adjoint-based methodologies are explored. Very few velocity measurements at optimized locations were able to successfully reconstruct the field. Another advantage of this approach is its capability to estimate the minimum number of sensors required to attain a desired level of reconstruction accuracy. However, it is essential to note that the research relies on the laminar RANS equations and low Reynolds number, and the consideration of an actual turbulence model is identified as a prospective avenue for future investigation. Furthermore, the challenges related to resolving the sensor placement problem stem from limitations in the measurement process itself. For instance, obtaining dense measurements of the entire mean

flow can be problematic, and such comprehensive data may not always be available. In the case of hot-wire anemometry, due to interference of the equipment only a limited number of hot-wires can be used to perform measurements simultaneously. Additionally, certain sensor placement locations of significance may be excluded due to equipment limitations. For example, in the case of hot-wire sensors, high-turbulent and reverse flows within the domain can hinder accurate measurements [25] Similarly, placing pressure sensors near the edges of a model may also be restricted due to practical limitations. These considerations highlight the multifaceted nature of the sensor placement problem, where both algorithmic and measurement-related aspects must be carefully addressed.

In this article, RANS simulation for the flow around a high-rise building is augmented via integrating heterogeneous experimental data using tools based on the EnKF. The observation is provided in the form of time-averaged data from several pressure taps on the surface of the building and a limited amount of velocity measurements obtained via hot wires. One additional difficulty of this work is that data is obtained with different tools, and therefore the acquisition systems exhibit different features and challenges. This heterogeneous observation, which shares features with the data employed in the two studies previously discussed, gives the opportunity to perform a RANS model optimization using realistic data available in an urban setting. The DA augmentation is performed by optimizing several global free constants that determine the behavior of time-averaged closure. To this purpose, a dedicated C++ library is developed, which performs an on-the-fly coupling of the CFD runs with the data-driven algorithms, dramatically reducing the computational costs. It will be shown that approaches based on the EnKF, owing to the smoothing characteristics of the filter, are suitable for robust integration of experimental data within the reduced-order CFD formalism.

The article is organized as follows. In Sec. 2, the numerical strategies and the algorithms used in this work are presented. This includes a description of the numerical solver used as well as a presentation of the data-driven strategies, which are integrated into a specific C++ library. In Sec. 3, the setup of the DA analysis is presented. The different techniques are outlined and compared, selecting the best-performing technique. This section is also supported by the Appendix A. In Sec. 4, the results obtained are compared with data from a high-fidelity simulation and experiments. At last, in Sec. 5, the final remarks are drawn, and future perspectives are discussed.

2. Numerical Ingredients

2.1. Numerical code: OpenFOAM

Numerical simulations in this work are performed using a C++ open-source library known as *OpenFOAM* [26]. This library includes a number of solvers based on Finite Volumes (FV) discretization [27], as well as a number of utilities for preprocessing, postprocessing, and data manipulation. Owing to the free license and the very large number of modules available, allowing for extended multi-physics analyses, this code has been extensively used in the literature for research work in fluid mechanics [28, 29, 20].

For this work, the FV numerical discretization is performed for the RANS Navier-Stokes equations for stationary, incompressible flows and Newtonian fluids [1]:

$$\overline{u}_j \frac{\partial \overline{u}_i}{\partial x_j} = -\frac{\partial \overline{p}}{\partial x_i} + \frac{\partial \overline{\tau}_{ij}}{\partial x_j} - \frac{\partial \tau_{ij}^T}{\partial x_j} \qquad i = 1, 2, 3 \tag{1}$$

$$\nabla^2 \overline{p} = -\frac{\partial \overline{u}_j}{\partial x_i}\frac{\partial \overline{u}_j}{\partial x_i} - \frac{\partial}{\partial x_i}\left(\frac{\partial \tau_{ij}^T}{\partial x_j}\right) \tag{2}$$

where Eq. 1 is the momentum equation and Eq. 2 is the Poisson equation. The variables used are the velocity $\mathbf{u} = [u_1, u_2, u_3] = [u_x, u_y, u_z]$, the normalized pressure p, the viscous stress tensor τ_{ij} (which is modeled using the Newtonian fluid hypothesis) and the Reynolds stress tensor τ_{ij}^T. The overbar indicates the average operation performed to obtain Eqs. 1 - 2. The axes are oriented so that x is the streamwise direction, y is the spanwise direction and z is the vertical direction. Within the RANS framework, a turbulence closure must be used for τ_{ij}^T. The $\mathcal{K} - \varepsilon$ model [30, 2] uses the eddy viscosity hypothesis to create a link between τ_{ij}^T and the gradient of the averaged velocity $\overline{\mathbf{u}}$:

$$-\tau_{ij}^T = 2\nu_T \overline{S}_{ij} - \frac{2}{3}\mathcal{K}\delta_{ij} \tag{3}$$

where ν_T is the turbulent viscosity, $\mathcal{K}$ is the turbulent kinetic energy and $\overline{S}_{ij}$ is the mean strain rate:

$$\overline{S}_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right) \tag{4}$$

In the $\mathcal{K}-\varepsilon$ model, ν_T is expressed as an algebraic function of $\mathcal{K}$ and the energy dissipation rate ε:

$$\nu_T = C_\mu \frac{\mathcal{K}^2}{\varepsilon} \tag{5}$$

where C_μ is a model constant to be calibrated. To close the problem, two model equations for $\mathcal{K}$ and ε must be included:

$$\frac{\partial \mathcal{K}}{\partial t} + \overline{u}_j \frac{\partial \mathcal{K}}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_T}{\sigma_\mathcal{K}}\right)\frac{\partial \mathcal{K}}{\partial x_j}\right] + \mathcal{P} - \varepsilon \tag{6}$$

$$\frac{\partial \varepsilon}{\partial t} + \overline{u}_j \frac{\partial \varepsilon}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_T}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_j}\right] + C_{\varepsilon 1}\frac{\varepsilon}{\mathcal{K}}\mathcal{P} - C_{\varepsilon 2}\frac{\varepsilon^2}{\mathcal{K}} \tag{7}$$

where the production term $\mathcal{P} = \nu_T \overline{S}^2$, $\overline{S} = \sqrt{2\overline{S}_{ij}\overline{S}_{ij}}$. The model is complete once the five constants C_μ, $C_{\varepsilon 1}$, $C_{\varepsilon 2}$, $\sigma_\mathcal{K}$ and σ_ε are determined. Launder and Sharma [30] provided values that were calibrated via the analysis of academic test cases, such as the free decay of homogeneous isotropic turbulence or the turbulent plane channel.

A second popular turbulence model implemented in numerous CFD solvers is the $\mathcal{K}-\omega$ SST model [2]. Its formulation relies on blending functions which are governed by a number of parameters, including the distance to the nearest wall δ. The blending functions provide a hybrid modeling combining features of the two classical models $\mathcal{K}-\varepsilon$ and $\mathcal{K}-\omega$ [2]. More precisely, the $\mathcal{K}-\omega$ SST model behaves like the $\mathcal{K}-\omega$ model close to the surface of the immersed bodies, and it transitions to a $\mathcal{K}-\varepsilon$ behavior with increasing distance. OpenFOAM's formulation of this model is based on Ref. [31], where two model equations for $\mathcal{K}$ and ω must be included. A limiter for the production term in the turbulent kinetic energy equation is introduced to exclude very large turbulence levels in regions with high normal strain. Hence, the production term is $\widetilde{\mathcal{P}} = \min(\mathcal{P}, 10\,\beta^\star \mathcal{K}\omega)$. The turbulence dissipation rate $\varepsilon = \beta^\star \mathcal{K}\omega$ is defined with the restrictive parameter $\beta^\star$ for the same reason.

$$\frac{\partial \mathcal{K}}{\partial t} + \overline{u}_j \frac{\partial \mathcal{K}}{\partial x_j} = \frac{\partial}{\partial x_j}\left[(\nu + \sigma_{\mathcal{K}}\nu_T)\frac{\partial \mathcal{K}}{\partial x_j}\right] + \widetilde{\mathcal{P}} - \beta^\star \mathcal{K}\omega \tag{8}$$

$$\frac{\partial \omega}{\partial t} + \overline{u}_j \frac{\partial \omega}{\partial x_j} = \frac{\partial}{\partial x_j}\left[(\nu + \sigma_{\omega}\nu_T)\frac{\partial \omega}{\partial x_j}\right] + \alpha\overline{S}^2 - \beta\omega^2 + 2(1-F_1)\sigma_{\omega 2}\frac{1}{\omega}\frac{\partial \mathcal{K}}{\partial x_i}\frac{\partial \omega}{\partial x_i} \tag{9}$$

F_1 is a blending function characterized by:

$$F_1 = \tanh\left(\left\{\min\left[\max\left(\frac{\sqrt{\mathcal{K}}}{\beta^\star\omega\delta}, \frac{500\nu}{\delta^2\omega}\right), \frac{4\sigma_{\omega 2}\mathcal{K}}{CD_{\mathcal{K}\omega}\delta^2}\right]\right\}^4\right) \tag{10}$$

with $CD_{\mathcal{K}\omega} = \max\left(2\sigma_{\omega 2}\frac{1}{\omega}\frac{\partial \mathcal{K}}{\partial x_i}\frac{\partial \omega}{\partial x_i}, 10^{-10}\right)$. F_1 controls the blending between the $\mathcal{K}-\varepsilon$ and $\mathcal{K}-\omega$ models. More precisely, $F_1 \approx 0$ corresponds to a $\mathcal{K}-\varepsilon$ model far from the immersed body surface, and $F_1 \approx 1$ corresponds to the $\mathcal{K}-\omega$ model in the proximity of the body. To close the problem, the turbulent viscosity ν_T is defined as:

$$\nu_T = \frac{a_1\mathcal{K}}{\max(a_1\omega, \overline{S}F_2)} \tag{11}$$

where a_1 is a constant ($a_1 = 0.31$) and F_2 is another blending function:

$$F_2 = \tanh\left(\left[\max\left(\frac{2\sqrt{\mathcal{K}}}{\beta^\star\omega\delta}, \frac{500\nu}{\delta^2\omega}\right)\right]^2\right) \tag{12}$$

The nine global constants that determine the model are $\sigma_{\mathcal{K}1}$, $\sigma_{\mathcal{K}2}$, $\sigma_{\omega 1}$, $\sigma_{\omega 2}$, α_1, α_2, β_1, β_2 and $\beta^\star$. With the exception of the latter, one can see that every other coefficient is actually defined by two subscripts $1, 2$. The coefficients included in Eqs. 8 - 9 are determined via a linear interpolation between the two values provided, which is driven by the function F_1. For example, $\alpha = \alpha_1 F_1 + \alpha_2(1-F_1)$. From this last relation, one can deduce that the subscript 1 provides a value classical for the $\mathcal{K}-\omega$ model and, on the other hand, the subscript 2 gives back values for the $\mathcal{K}-\varepsilon$ model. Menter determined values for the nine parameters [32]. In practice, similarly to the

$\mathcal{K} - \varepsilon$ model, these coefficients are not constants but they are a function of the local dynamics of the flow and their interaction with global features (see discussion in Refs. [33, 9, 14]).

2.2. Wind tunnel experiments and the high-rise building model

The experiments were conducted in the atmospheric boundary-layer wind tunnel of the Ruhr-University Bochum, Germany. The wind tunnel has a cross-section of 1.6 × 1.8 m and a test section length of 9.4 m. The model has a square cross-section with edges $B = 0.133$ m, and the height of the building $H = 0.4$ m, representing a 120 m tall building in the full-scale. The building has a flat roof, and 0° wind direction is investigated so that the asymptotic velocity is aligned with the streamwise direction x. Fig. 1 a) shows the wooden building model mounted on a rotating table in the wind tunnel.

The mean incident wind profile, measured in the empty wind tunnel at the center of the turntable, matches that of a power law with the exponent of 0.2, as shown in Fig. 1 b). The mean velocity at the referenced model height is $u_{ref} = 16$ m/s, while the streamwise turbulence intensity is $I_u = 13\%$. This is representative of the terrain category II [34] simulating realistic conditions of the flow around isolated high-rise buildings, which can be used to approximate the flow pattern in urban areas with a dominant high-rise building surrounded by sparse low-rise buildings. Such an arrangement is common on the outskirts of large cities.

The wind tunnel measurements also included pressure measurements using 64 pressure taps on the roof and 26 taps on the facades, as well as velocity measurements at 28 locations above the roof, as shown in Fig. 2.

2.3. RANS simulation

The considered high-rise building case is a numerical representation of the wind tunnel tests. The dimensions of the computational domain are chosen by adopting the best practice guidelines given by Tominaga et al. [18]. The upstream domain length is $5H$. The resulting dimensions of the domain are length (x) × width (y) × height (z) $15.5H \times 4.5H \times 4H = 6.2$ m × 1.8 m × 1.6 m. For the z direction, the height has been chosen to match the height of the wind tunnel.

A structured grid is used near the high-rise building surfaces, as shown in Fig. 3. The distance from the center point of the wall adjacent cell to the building leads to an average $y^+ = 141$ and minimum $y^+ = 40$, which ensures

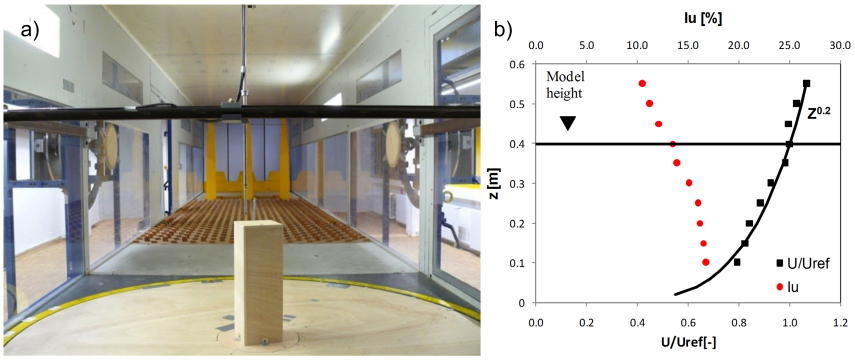


Figure 1: Wind tunnel test section used to produce experimental data. (a) Visualization of the high-rise building model and its installation in the wind tunnel and (b) shape of the average velocity and turbulence intensity profile in the streamwise direction x.

that the centers of the near-wall grid elements are located in the logarithmic layer. The total number of grid elements used to discretize the domain is equal to 513 266 cells.

A grid dependency study is performed comparing the numerical results against those obtained using a finer grid. This more refined grid is composed of 4.3×10^6 cells and characterized by a spatial resolution that is two times higher near the building model than the coarse case. The mean pressure predicted by the coarse and fine grid simulations is compared at the locations of the pressure taps in Fig. 2. The comparison showed that 86% of the points on the building surface have a relative difference below 10%.

Two turbulence models are chosen: $\mathcal{K} - \varepsilon$ and $\mathcal{K} - \omega$ SST. The inlet boundary conditions for the velocity field as well as for the statistical features of turbulence are set using these equations:

$$u_x(z) = u_{ref}\left(\frac{z}{z_{ref}}\right)^{0.2} \tag{13}$$

$$u_y(z) = u_z(z) = 0 \tag{14}$$

$$\mathcal{K}(z) = a(I_u(z)u(z))^2 \tag{15}$$

$$\varepsilon(z) = \frac{u_{ABL}^{*3}}{\kappa(z+z_0)} \tag{16}$$

$$\omega(z) = \frac{\epsilon(z)}{\beta'\mathcal{K}(z)} \tag{17}$$

For the $\mathcal{K}-\varepsilon$ model, the inlet boundary condition for the mean velocity $\overline{\mathbf{u}}$, turbulent kinetic energy $\mathcal{K}$ and turbulence dissipation rate ε are determined from the incident vertical wind tunnel profiles. The turbulent kinetic energy $\mathcal{K}$ is calculated using Eq. 15 from the measured $\overline{\mathbf{u}}$ and the measured streamwise turbulence intensity $I_u\left(I_u(z) = \frac{\sigma_u(z)}{\overline{\mathbf{u}}(z)}\right)$, where $\sigma_u(z)$ is standard deviation of streamwise velocity component. a is a parameter in the range between 0.5 and 1.5 [35, 36, 18], and in this study, $a = 1$ is chosen, as recommended by [18]. Eq. 16 and Eq. 17 are using the von Karman constant $\kappa = 0.42$ and $\beta' = 0.09$, respectively. The aerodynamic roughness length is defined as $z_0 = 0.002$ m, and the friction velocity $u_{ABL}^{*} = \kappa\frac{u_{ref}}{\ln\left(\frac{H+z_0}{z_0}\right)}$. Similarly, in the case of $\mathcal{K}-\omega$ SST turbulence model, the specific turbulent dissipation rate ω is calculated based on the turbulence dissipation rate from Eq. 16 and turbulence kinetic energy from Eq. 15. The inlet boundary conditions are assigned using the groovyBC library in OpenFOAM. The outlet is considered to be a pressure outlet with a constant relative pressure equal to zero and a zero-gradient boundary condition for the velocity.

The SIMPLE algorithm [27] was used for pressure-velocity coupling. Classical choices have been performed for the numerical schemes. First-order upwind schemes have been used for the convection terms, while second-order centered schemes have been used for viscous terms. Pressure interpolation from the cell center to the face center has been obtained via second-order linear schemes native to the OpenFOAM solver.

2.4. Observation: experimental data obtained in wind tunnels

The sampled data is heterogeneous as different sensors are used to capture features of the velocity field on the roof and pressure measurements on the surface of the building. The velocities above the roof are mainly measured at two different heights ($z/H = 1.1, 1.15$) at locations marked with black circles in Fig. 2 d). In addition, above two locations at the roof, including the center of the roof, Fig. 2 d), nine heights are considered with the spacing of $(z - H)/H = 0.025$. The measurements are performed using a hot-wire anemometer, which consists of two cross wires allowing for measurement of both stream-wise and vertical velocity components. All velocity data are sampled with the frequency of 2000 Hz.

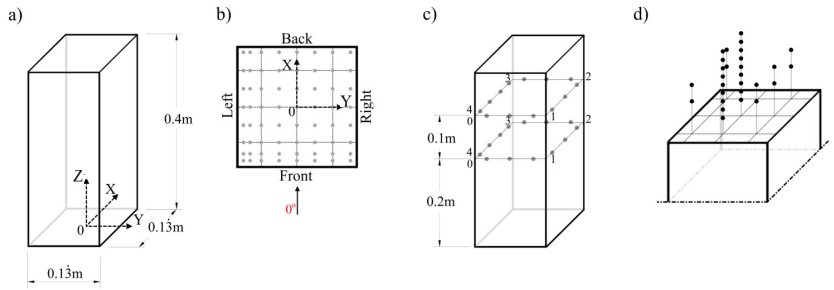


Figure 2: Geometry of the high-rise building with (a) main dimensions and coordinate system; (b) top view with pressure tap locations; (c) facades with pressure tap locations (d) velocity observations measured over the rooftop.

In addition to the velocity measurements, the surface pressure is also sampled at different locations, as shown in Fig. 2 b) and c) marked with light gray circles. Surface pressures are acquired with a sampling frequency of 1000 Hz using a multi-channel simultaneous scanning measurement system. The tubing effects are numerically compensated [37]. More details about the wind tunnel experimentation and the analysis of the flow around high-rise buildings, with a special focus on above the roof, are presented in [19]. In this analysis, experimental data is used to improve the predictive capabilities of two stationary RANS models (standard $\mathcal{K}-\varepsilon$ and $\mathcal{K}-\omega$ SST). Therefore, the time series available for the velocity components and the pressure have been averaged in time.

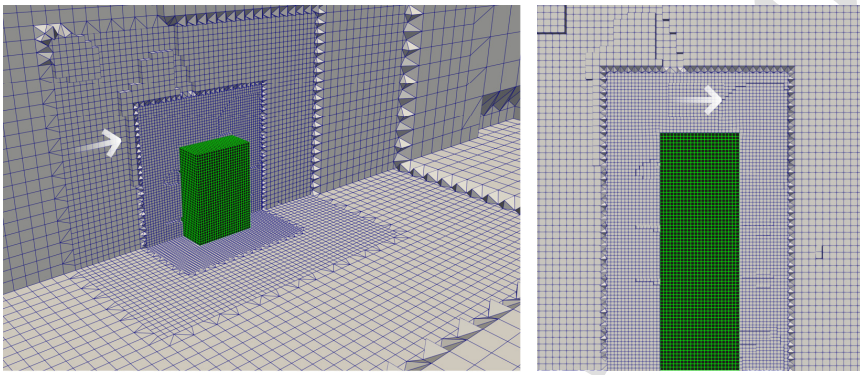


Figure 3: View of the grid used for the RANS calculations. The left shows the central vertical plane and a horizontal plane at $H/2$. The right shows the central vertical plane. The direction of the asymptotic flow is indicated by the white arrow.

2.5. Data Assimilation: Ensemble Kalman Filter

Data Assimilation (DA) [5, 8] is a family of tools allowing to combine several sources of information to obtain an *augmented prediction* exhibiting increased accuracy. Classical applications usually rely on:

- a *model*, which provides a (quasi) continuous representation of the physical phenomenon investigated. Physics-based models such as CFD solvers are an example of *model* for fluid mechanics applications.
- some *observation*, which is usually more accurate than the model, but it is local in space and time. In fluid mechanics, this data may come from high-fidelity numerical simulations or from experiments.

The *augmented prediction* obtained via manipulation of the sources of information can also be actively used to infer an optimized parametric description of the *model*, with the aim to obtain a predictive tool that can provide accurate predictions without having to rely on *observation*. DA has been traditionally used in environmental and weather sciences, but applications in fluid mechanics have seen a rapid rise in recent times [38, 39, 12, 40, 41, 42, 43, 13, 24, 44].

A great variety of methods exists, but two groups can be identified [8, 45]:

Variational methods: methods for which the goal is to minimize a cost function applied for the case studied. This minimum, which is usually reached via parametric optimization of the model, provides an accurate flow state.

Statistical methods: methods that aim to obtain an accurate state estimation of the physical phenomena investigated minimizing the variance of the final solution (i.e., increasing the confidence in the prediction). Statistical methods are mostly sequential [8], even though non-sequential approaches have been developed, such as the Kalman Smoother [46].

Variational methods such as the 4DVar have been extensively used for application in fluid mechanics [47, 38, 23, 43] in particular with steady-state simulations, due to their non-sequential behavior. While statistical sequential tools are supposedly more appropriate for the prediction of non-stationary phenomena, applications to steady flows are reported in the literature [40, 13, 48]. In the present work, we will focus on tools derived from the Kalman Filter, a well-known sequential method.

2.5.1. The Kalman Filter

The Kalman Filter (KF) [6] is a sequential DA method based on the Bayes theorem. It provides a solution to the linear filtering of time-dependent discrete data. The classical formulation for the analysis of a physical quantity $\mathbf{x}$ relies on the combination of results produced via a discrete model $\mathbf{M}$, which is linear in the original KF, and some observation $\mathbf{y}$. Within the framework of KF, both the model and the observation are affected by errors/uncertainties, which are here referred to as v and w, respectively. One of the central hypotheses of the Kalman Filter is that these uncertainties can be accurately described by an unbiased Gaussian distribution i.e. $v = \mathcal{N}(0, \mathbf{Q})$ and $w = \mathcal{N}(0, \mathbf{R})$. $\mathbf{Q}$ and $\mathbf{R}$, which also are a function of time, represent the variance of the model and of the observation, respectively. Considering that these errors can be described by a Gaussian distribution, the solution is completely determined by the first two moments of the state i.e. the physical quantity $\mathbf{x}$ and the error covariance matrix $\mathbf{P} = \mathbb{E}((\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^T)$.

The main drawbacks of the algorithm for complex applications in fluid mechanics are that i) it is designed for linear models $\mathbf{M}$ and ii) the size of $\mathbf{P}$ is directly linked with the number of degrees of freedom of the problem investigated. While the first issue can be bypassed with ad-hoc improvements

of the data-driven strategy, which are included for example in the *extended* KF [8, 45], the second one is more serious. In fact, $\mathbf{P}$ must be advanced in time like the physical variables. In addition, extended manipulation of $\mathbf{P}$ is required, including a matrix inversion. For the number of degrees of freedom used in CFD, which are usually in the range $10^6 - 10^8$, this leads to prohibitive requirements in terms of RAM and computational resources.

2.5.2. The (stochastic) Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) [49, 8] is a popular data-driven strategy based on the KF and Monte Carlo approach which provides an efficient solution to the issues previously discussed. The idea is that the error covariance matrix $\mathbf{P}$ is not advanced in time anymore, but it is approximated via an ensemble of model runs. This strategy allows us to fully account for the non-linearity of the model and it virtually eliminates the computational burdens associated with the manipulation of $\mathbf{P}$. The complete structure of the EnKF, which is summarized in the Alg. 1, is now discussed.

Let us consider the time advancement of the physical system between the instant k and $k+1$. For the latter, observation $\mathbf{y}_{k+1}$ is available. In this case, the data assimilation procedure consists of two phases :

A *forecast* step (superscript f), where the physical state and the error covariance matrix at the time k are advanced in time using the (non-linear) model $\mathcal{M}$:

$$\mathbf{x}^f_{i,k+1} = \mathcal{M}\mathbf{x}^a_{i,k} \tag{18}$$

The EnKF relies on N_e realizations of the model, which is the model ensemble. The realizations can be assembled in a state matrix $\mathbf{X}_S$ of size $[N, N_e]$, where N is the number of degrees of freedom of the physical problem investigated. Therefore, each column of $\mathbf{X}_S$ corresponds to the state $\mathbf{x}^f_{i,k+1}$ of the i^{th} member, where $i \in [1, N_e]$. An approximation of the error covariance matrix $\mathbf{P}_e$ can be obtained by exploiting the hypothesis of statistical independence of the ensemble members:

$$\mathbf{P}^f_e = \mathbf{X}^f(\mathbf{X}^f)^T \tag{19}$$

where $\mathbf{X}^f$ is the anomaly matrix which represents the deviation of all the

values of the state vectors from their ensemble means:

$$\mathbf{X}_{k+1} = \frac{\mathbf{x}_{i,k+1} - \overline{\mathbf{x}}_{k+1}}{\sqrt{N_e - 1}}\,, \qquad \overline{\mathbf{x}}_{k+1} = \frac{1}{N_e}\sum_{i=1}^{N_e} \mathbf{x}_{i,k+1} \tag{20}$$

The sampled observation, which consists of N_o elements, is also expanded to obtain N_e sets of values. To do so, a Gaussian noise based on the covariance matrix of the measurement error $\mathbf{R}_{k+1}$ is added to the observation vector:

$$\mathbf{y}_{i,k+1} = \mathbf{y}_{k+1} + \mathbf{e}_{i,k+1}, \text{ with } \mathbf{e}_{i,k+1} \sim \mathcal{N}(0, \mathbf{R}_{k+1}) \tag{21}$$

The model realizations are projected to the observation space N_e times:

$$\mathbf{s}_{i,k+1} = \mathcal{H}(\mathbf{x}^f_{i,k+1}) \tag{22}$$

where $\mathcal{H}(\mathbf{x}^f_{i,k+1})$ is a non-linear operator mapping the model results to the observation space.

The second step is the *analysis* phase (superscript a), where observation and forecast are combined to obtain the *augmented prediction*. One of the main goals here consists of the determination of the Kalman gain $\mathbf{K}_{k+1}$. This matrix takes into account the correlations between the values of the state vector and the values of the observations, and it is the central element providing the final state estimation of the physical system.

$$\mathbf{S}_{k+1} = \frac{\mathbf{s}_{i,k+1} - \overline{\mathbf{s}}_{k+1}}{\sqrt{N_e - 1}}\,, \qquad \overline{\mathbf{s}}_{k+1} = \frac{1}{N_e}\sum_{i=1}^{N_e} \mathbf{s}_{i,k+1} \tag{23}$$

$$\mathbf{E}_{k+1} = \frac{\mathbf{e}_{i,k+1} - \overline{\mathbf{e}}_{k+1}}{\sqrt{N_e - 1}}\,, \qquad \overline{\mathbf{e}}_{k+1} = \frac{1}{N_e}\sum_{i=1}^{N_e} \mathbf{e}_{i,k+1} \tag{24}$$

$$\mathbf{K}_{k+1} = \mathbf{X}^f_{k+1}(\mathbf{S}_{k+1})^T \left[\mathbf{S}_{k+1}(\mathbf{S}_{k+1})^T + \mathbf{E}_{k+1}(\mathbf{E}_{k+1})^T\right]^{-1} \tag{25}$$

In an infinite ensemble size $\mathbf{E}_{k+1}(\mathbf{E}_{k+1})^T$ tends to the matrix $\mathbf{R}_{k+1}$ of the Kalman filter. In practice the size is limited, thus the product of the perturbations is simplified by the diagonal matrix $\mathbf{R}_{k+1}$ gaining simplification and computational cost [45, 50]. In addition, $\mathbf{P}_e$ can be directly estimated from the ensemble members for each analysis phase, and there is no need for memory storage/time advancement.

Finally, the N_e updates of the state vectors are performed:

$$\mathbf{x}^a_{i,k+1} = \mathbf{x}^f_{i,k+1} + \mathbf{K}_{k+1}(\mathbf{y}_{i,k+1} - \mathbf{s}_{i,k+1}) \tag{26}$$

The EnKF can also be used to optimize the parametric description of the model. The underlying idea is that the parameters are updated at the end of the analysis phase so that the model can provide a more accurate prediction of the physical phenomenon investigated, reducing the difference between the model-predicted state and the final state estimation. Several strategies are proposed in the literature [8] and, among those, one showing efficiency for a relatively small set of parameters (referred to as θ) and easy to implement is the so-called *extended state*. In this strategy, the steps of the EnKF are performed for a state $\mathbf{x}^\star$ which is defined as:

$$\mathbf{x}^\star = \begin{bmatrix} \mathbf{x} \\ \theta \end{bmatrix} \tag{27}$$

That is, the state used for the EnKF includes both the physical state and the parametric description of the model. For this very simple algorithm, the size of the global state is now equal to $N^\star = N + N_\theta$, where N_θ is the number of parameters to be optimized. This modification brings a negligible increase in computational costs if $N_\theta << N$ and it simultaneously provides an updated state estimation and optimized parametric description for the model at the end of the analysis phase.

Algorithm 1: Algorithm for the Ensemble Kalman Filter

Input: $\mathcal{M}$, $\mathcal{H}$, $\mathbf{R}_{k+1}$, and some priors for the state system $\mathbf{x}^a_{i,0}$, where usually $\mathbf{x}^a_{i,0} \sim \mathcal{N}(\mu_N, \sigma^2_N)$

for $k = 0$ *to* $K-1$ **do**

  for $i = 1$ *to* N_e **do**

1 Advancement in time of the state vectors:

$$\mathbf{x}^f_{i,k+1} = \mathcal{M}\mathbf{x}^a_{i,k}$$

2 Creation of an observation matrix from the observation data by introducing errors:

$$\mathbf{y}_{i,k+1} = \mathbf{y}_{k+1} + \mathbf{e}_{i,k+1}, \text{ with } \mathbf{e}_{i,k+1} \sim \mathcal{N}(0, \mathbf{R}_{k+1})$$

3 Calculation of the predicted observation:

$$\mathbf{s}_{i,k+1} = \mathcal{H}(\mathbf{x}^f_{i,k+1})$$

4 Calculation of the ensemble means:

$$\overline{\mathbf{x}}^f_{k+1} = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{x}^f_{i,k+1},\ \overline{\mathbf{s}}_{k+1} = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{s}_{i,k+1},$$
$$\overline{\mathbf{e}}_{k+1} = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{e}_{i,k+1}$$

5 Calculation of the anomaly matrices:

$$\mathbf{X}_{k+1} = \frac{\mathbf{x}_{i,k+1}-\overline{\mathbf{x}}_{k+1}}{\sqrt{N_e-1}},\ \mathbf{S}_{k+1} = \frac{\mathbf{s}_{i,k+1}-\overline{\mathbf{s}}_{k+1}}{\sqrt{N_e-1}},$$
$$\mathbf{E}_{k+1} = \frac{\mathbf{e}_{i,k+1}-\overline{\mathbf{e}}_{k+1}}{\sqrt{N_e-1}}$$

6 Calculation of the Kalman gain:

$$\mathbf{K}_{k+1} = \mathbf{X}^f_{k+1}(\mathbf{S}_{k+1})^T\left[\mathbf{S}_{k+1}(\mathbf{S}_{k+1})^T + \mathbf{R}_{k+1}\right]^{-1}$$

7 Update of the state matrix:

$$\mathbf{x}^a_{i,k+1} = \mathbf{x}^f_{i,k+1} + \mathbf{K}_{k+1}(\mathbf{y}_{i,k+1} - \mathbf{s}_{i,k+1})$$

2.5.3. Inflation

The classical EnKF exhibits a number of shortcomings such as sampling errors due to the limited amount of members available in the ensemble. This is especially true for applications in fluid mechanics and in particular with CFD, where every simulation may need important computational resources and storage space. Therefore, the number of total ensemble members realistically acceptable for three-dimensional runs is around $N_e \in [40, 100]$. As this error is carried over the assimilation steps, one way of reducing this problem is to inflate the error covariance matrix $\mathbf{P}_{k+1}$ by a factor λ^2 [8].

This coefficient $\lambda > 1$ drives the so-called *multiplicative inflation*, which can be applied to the analyzed state matrix. It is responsible for an increased

variability of the state estimation:

$$\mathbf{x}_i^a \longrightarrow \overline{\mathbf{x}^a} + \lambda(\mathbf{x}_i^a - \overline{\mathbf{x}^a}) \tag{28}$$

Clearly, for $\lambda = 1$ the results from the classical EnKF are obtained.

Similarly, the optimization via EnKF of the set of inferred parameters θ can collapse very rapidly towards a local optimum, providing a sub-optimal result. Inflation can be used to mitigate an overly fast collapse of the parametric description of the model, artificially increasing the variability of the parameters and allowing it to target a global optimum solution.

2.5.4. Localization

The classical EnKF establishes a correlation between observation and the degrees of freedom of the model, but this correlation is not affected by the distance between them. In a limited ensemble size like the ones used in CFD, this can lead to spurious effects on the update of the state matrix for large domains. In practice, errors due to the finite ensemble approximations can be significantly larger than the real physical correlation, which naturally decays with distance in continuous systems. Due to the computational limitations to using more members in the ensemble, one way to avoid these spurious effects is to use a corrective multiplicative term to the values of the covariance matrix $\mathbf{P}_{k+1}^f$ that takes into account the physical distance between observation sensors and mesh elements of the state. This strategy is known as *covariance localization*. Just as the inflation, the localization is effective in improving the accuracy of the calculation and reducing the probability of divergence of the EnKF. The principle of covariance localization uses a coefficient-wise multiplication of the covariance matrix $\mathbf{P}_{k+1}^f$ and a corrective matrix that is here called $\mathbf{L}$. This type of operation is known as a Schur product, thus it is also called *Schur localization*. This leads to the expression of the localized Kalman gain Eq. 29.

$$[\mathbf{P}_{k+1}^f]_{i,j}[\mathbf{L}]_{i,j} \longrightarrow \mathbf{K}_{k+1}^{loc} = [\mathbf{L}]_{i,j}[\mathbf{X}_{k+1}^f(\mathbf{S}_{k+1})^T]_{i,j}\left([\mathbf{L}]_{i,j}[\mathbf{S}_{k+1}(\mathbf{S}_{k+1})^T]_{i,j} + \mathbf{R}_{k+1}\right)^{-1} \tag{29}$$

As the matrix $\mathbf{R}_{k+1}$ has a limited impact on the operation, this expression is simplified for convenience in the algorithm. The localized Kalman gain becomes:

$$[\mathbf{P}_{k+1}^f]_{i,j}[\mathbf{L}]_{i,j} \longrightarrow \mathbf{K}_{k+1}^{loc} = [\mathbf{L}]_{i,j}[\mathbf{K}_{k+1}]_{i,j} \tag{30}$$

The structure of the matrix $\mathbf{L}$ must be set by the user, and it should represent the real physical correlation. In continuous systems, the correlation between physical variables decreases fast in space. Therefore, a generally used structure for the localization matrix is an exponential decay form:

$$\mathbf{L}(i,j) = e^{-\Delta_{i,j}^2/\eta} \tag{31}$$

where $\Delta_{i,j}$ is the distance between the given observation sensor and the point of evaluation of the model (center of the mesh element in CFD). η is a decay coefficient that can be tuned accordingly to the characteristics of the test case.

2.6. CONES: Coupling OpenFOAM with Numerical EnvironmentS

CONES (Coupling OpenFOAM with Numerical EnvironmentS) is a C++ library designed to couple the CFD software OpenFOAM with any other kind of open-source code. It is currently employed to carry out sequential DA techniques and, more specifically, advanced data-driven methods based on the EnKF. The communications between the EnKF-based code and OpenFOAM are performed by CWIPI (Coupling With Interpolation Parallel Interface) [51], which is an open-source code coupler for massively parallel multi-physics/multi-components applications and dynamic algorithms.

The main favorable features of CONES in performing DA with OpenFOAM are the following:

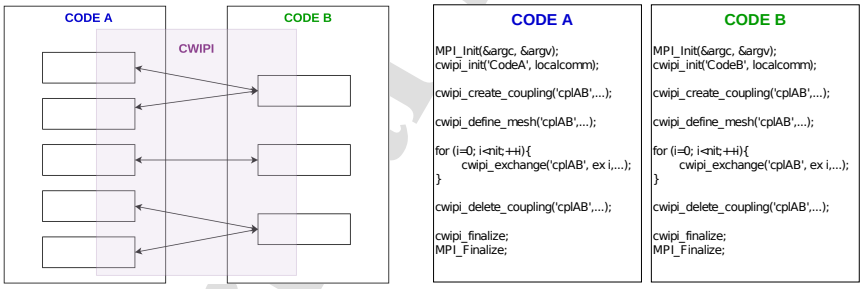
- It is not needed to modify the installation of OpenFOAM but only compile user-made functions.
- The coupling between the CFD code, the observation, and the DA algorithm is performed preserving the original structure of the existing CFD solvers. Every CONES-related function is contained in a Pstream (Part of OpenFOAM) modified library, hence, data exchange is done at the end of the solver loop by calling specific functions, and the calculation loop remains unmodified.
- Simulations and DA are run simultaneously *online*. The coupling is intrusive from the solver's point of view. It is done in such a way that the analysis phases can be performed on the fly by pausing the CFD calculation. Therefore, there is no need to stop and restart CFD calculations, which can require an important amount of computational resources.

- It is very efficient to exchange information about the physical state and the mesh (arrays of millions of elements can be sent and received simultaneously and rapidly).

- Direct HPC communications are established between multiple processors, which handle partitions of the numerical simulations and the DA processes.

The coupler CWIPI developed by ONERA and CERFACS has been chosen due to its powerful management of fully parallel data exchanges based on distributed mesh definition and its ability to interpolate between non-coincident meshes (very useful for some advanced tools based on the EnKF, like the MGEnKF [44]). Most of its uses are related to gas turbine designs [52, 53], but it has also recently been employed in the field of aeroacoustics with OpenFOAM [54].

CWIPI communication protocol is based on the MPI library. Thus, MPI and CWIPI environments must be initialized within the codes. This will allow the use of CWIPI primitives to exchange information between two codes. In Fig. 4 direct communications between two codes through CWIPI and some of the main primitives are illustrated.



(a) CWIPI between two codes (b) Main primitives in CWIPI

Figure 4: Functioning of CWIPI.

In this work, CONES couples the solver SimpleFOAM, which is designed to simulate flows using RANS, with a sequential DA library developed by the team. The structure of a single run is exemplified in Fig. 5. The MPI communications and the coupler CWIPI are initialized in both codes (in OpenFOAM and in the DA library).

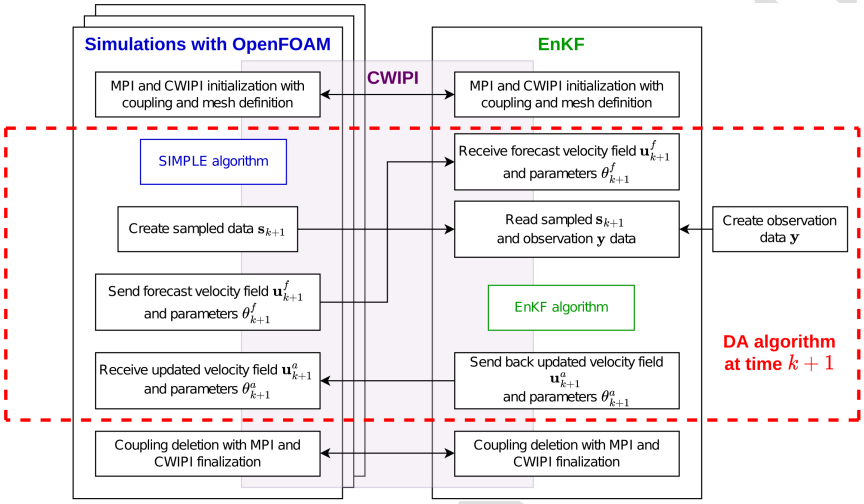


Figure 5: Scheme of CONES for steady simulations.

Despite the fact that applications of EnKF-based tools are tied with the time advancement of the solution, the application to stationary flows is straightforward. An analysis virtual time window is fixed in terms of the number of iterative steps of the code. Once that number of time steps is performed simultaneously by the N_e ensemble members (CFD runs), they send their information to the EnKF code and wait online for the updated flow field / parametric description. Currently, the information exchanged is the velocity field $\mathbf{u}^{f/a}$ and the parameters of the model $\theta^{f/a}$. Hence, the state matrix, composed of as many state vectors as members (CFD simulations) in the ensemble, is the one expressed in Eq. 32 for the DA cycle at time $k+1$.

$$\mathbf{x}^{f/a}_{i,k+1} = \begin{bmatrix} \mathbf{u}^{f/a}_{i,k+1} \\ \theta^{f/a}_{i,k+1} \end{bmatrix} \tag{32}$$

A piece of additional information provided by the simulations is the set of values $\mathbf{s}_{i,k+1}$, i.e., the projection of the model solution on the coordinates of the sensors for each ensemble member. Considering that the sensor placement does not necessarily comply with the center of a mesh element, interpolation of the flow field has to be performed. OpenFOAM possesses several functions

to transform cell-center quantities into particular points. The accuracy of these interpolation methods [55] has been taken into account.

The nature of the observations $\mathbf{y}$ is analyzed in more detail in Sec. 3, but CONES can work with sensors measuring both the pressure p and the velocity field $\mathbf{u}$. In this specific case dealing with stationary simulations, the observation is constant, and it is loaded once, but it could be integrated at each analysis phase in case of analysis of an nonstationary flow. Thus, the DA code receives information from the model and the sensors, producing an updated set of states and parameters ($\mathbf{u}^a$, θ^a), which are sent back to the OpenFOAM simulations. The pressure p is updated for each ensemble member via a Poisson equation, and this complete set of data is used to start a new set of iterative steps. Once the convergence of the model parameters complies with a threshold set by the user, the coupling is deleted, and both MPI and CWIPI environments are finalized.

3. DA experiments

CONES is here used to study the high-rise building flow configuration using the numerical test case presented in Sec. 2.2. In particular, the DA tools are used to optimize the value of some global constants which determine the performance of the turbulence model. These coefficients can be manually set by a user without any structural change of the solver. The present results aim to provide new general recommendations for the usage of turbulence modeling in particular for industrial applications, considering that the physical features of the test case investigated are observed for most flows around three-dimensional bluff bodies. Two main DA investigations are performed. In the first one, the five global coefficients driving the $\mathcal{K}-\varepsilon$ model [30] are optimized with the aim of minimizing the discrepancy between the RANS results and the high-fidelity experimental observation provided. The second analysis shares the same objective, but the model used is $\mathcal{K}-\omega$ SST [2], and in this case, the DA optimization targets the nine models coefficients introduced in Sec. 2.1.

The first key aspect to take into account for the model representation is determining a suitable prior state for the velocity and pressure field, as well as for the parametric description. For the latter, for both the $\mathcal{K}-\varepsilon$ model and the $\mathcal{K}-\omega$ SST model, values found by Margheri et al. [11] using uncertainty propagation of epistemic uncertainties are preferred to the classical values obtained by Launder and Sharma [30]. These baseline values, which are

shown in Tab. 1, are the initial mean of the N_e ensemble simulations. Each value of the parameters for each CFD run is initially determined using a bounded Gaussian distribution $\mathcal{N}(\mu_N, \sigma_N^2)$, where μ_N is the parameter mean value and σ_N is chosen to provide a sufficiently large initial variability of the parametric space based on the work by Margheri et al. [11]. This ensures a sufficiently large initial distribution of the parameters so that the EnKF can successfully target an optimized configuration. Details about the choice for σ_N for each model constant are provided in Appendix A. The number of ensemble members $N_e = 40$ is chosen considering other works in the literature relying on CFD for the model part of the EnKF [56, 57, 44].

$\mathcal{K}-\varepsilon$ **Parameters**	**default values**	**Prior of the EnKF** $\mu_\mathbf{N}$	**Prior of the EnKF** $\sigma_\mathbf{N}$	**Optimized values**
C_μ [-]	0.09	0.1	0.01	0.032
$C_{\varepsilon 1}$ [-]	1.44	1.575	0.1	0.165
$C_{\varepsilon 2}$ [-]	1.92	1.9	0.1	4.080
$\sigma_\mathcal{K}$ [-]	1.0	1.0	0.1	0.476
σ_ε [-]	1.3	1.6	0.1	0.1
$\mathcal{K}-\omega$ SST **Parameters**	**default values**	**Prior of the EnKF** $\mu_\mathbf{N}$	**Prior of the EnKF** $\sigma_\mathbf{N}$	**Optimized values**
$\sigma_{\mathcal{K}1}$ [-]	0.85	0.7	0.05	0.134
$\sigma_{\mathcal{K}2}$ [-]	1.0	1.0	0.05	0.004
$\sigma_{\omega 1}$ [-]	0.5	0.625	0.05	0.812
$\sigma_{\omega 2}$ [-]	0.856	0.856	0.05	0.103
α_1 [-]	0.5556	0.575	0.05	0.013
α_2 [-]	0.44	0.44	0.05	0.069
β_1 [-]	0.075	0.09	0.005	0.45
β_2 [-]	0.0828	0.0828	0.005	0.008
$\beta^\star$ [-]	0.09	0.09	0.005	0.191

Table 1: Global coefficients of RANS turbulence models to be optimized via DA. From left to right column: classical values in the literature, features of the truncated Gaussian distribution used to generate the prior, optimized values.

The observation is obtained from time-averaged data from a total of 118 sensors. Among these, 90 sensors are pressure taps, and 28 sensors are hot wires measuring two components of the velocity field, the streamwise velocity u_x and vertical velocity u_z. This adds up to 146 time-averaged observation

values. The data is loaded at the beginning of the first analysis phase in the following format: $\mathbf{y} = \begin{bmatrix} u_{x1} & \dots & u_{x28} & u_{z1} & \dots & u_{z28} & p_{29} & \dots & p_{118} \end{bmatrix}^T$. As it does not change throughout the calculation, it is stored in RAM and directly used at each analysis phase. One of the main features of the KF-based approaches is that a suitable confidence level in the accuracy of the observation must be provided. The accuracy of present experimental data is estimated to be around 5% for both pressure tap and hot wire measurements. It is also assumed that uncertainty in the experimental measurements is uncorrelated. Therefore, the covariance matrix for the observation R_{k+1} is considered to be constant and expressed as $R = \sigma_m I$, where σ_m is the variance describing the uncertainty in the measurements.

At last, the DA runs are performed using stationary CFD solvers. This implies that a true time evolution is not performed here. Initially, model runs are performed using the different setup descriptions produced by the Gaussian-shaped prior. Then, the DA procedure performs an optimization during the analysis phase. The state augmentation obtained via the analysis, used as initial condition for the forecast, is able to speed up the convergence towards the next analysis phase and stabilize the calculation with a little increase in computational resources. The simulations are then run again for a sufficiently large number of iterative steps so that the solver can efficiently propagate the effects of the new turbulence model and obtain a converged stationary solution. At this point, a DA analysis is performed and the cycle continues until the parametric optimization has reached a suitable degree of convergence. Initial tests indicated that, once the parametric description is changed during the analysis phase, 150 iterations are sufficient to obtain a converged flow field using the updated model.

Tests about the performance of the DA algorithms are reported in Appendix A. In particular, three different implementations of the EnKF have been tested: classical EnKF, EnKF with covariance localization, and EnKF with covariance localization and inflation. Discussions about localization and inflation hyperparameters are provided. The physical fields predicted by the three approaches are almost identical, but the EnKF with localization and without inflation demands lower computational resources. Therefore, the results presented in the following section are obtained with such DA strategy.

4. Results

The analysis is initially focused on the accuracy of the prediction of the velocity field. This prediction is essential for environmental applications. Thus, it is useful for the evaluation of relevant urban indicators such as pedestrian-level wind environment, near-field pollutant dispersion, natural ventilation, urban ventilation and urban wind energy. In this case, despite the shortcomings described in the introduction, RANS is considered a useful analysis tool [58]. The second part of the section is devoted to the analysis of the pressure field. For this physical quantity, which is relevant for structural wind engineering applications focusing on wind loads over the buildings, high-fidelity Large Eddy Simulation is considered to be more appropriate. Nonetheless, it will be shown that data-driven RANS can reach a good level of accuracy for this quantity, depending on the availability and positioning of high-fidelity data used as observation.

Results obtained by the DA runs are compared with available data. The comparisons will include results obtained from the prior simulations (classical RANS $\mathcal{K}-\varepsilon$ and $\mathcal{K}-\omega$ SST models), time-averaged experimental results, a validated reference LES simulation [59], RANS runs using the optimized models obtained by Ben-Ali et al. [20] and Zhao et al. [21]. Ben-Ali et al. actually developed a number of DA strategies of different complexity that encompass the optimization of global constants to the development of correction terms for the dynamic equations of turbulence. In this work, we decided to employ the simplest and least accurate of their models, namely the one targeting the optimization of model constants. This decision has been taken to provide a consistent base of assessment with present results and the findings of Zhao et al. The results obtained using the model by Ben-Ali et al. are almost identical to the prior $\mathcal{K}-\varepsilon$ simulation, therefore we just show the latter in the analysis for the sake of conciseness.

4.1. Velocity field

The analysis of the velocity field is performed first. Velocity is an explicit variable in segregated solvers for incompressible flows. Therefore, the performance of the DA strategies can be assessed by the qualitative improvement obtained for the prediction of this quantity.

Fig. 6 shows the comparison of the streamwise velocity profile u_x and vertical velocity profile u_z for several locations corresponding to the positions of the hot wire, for the DA analyses using the $\mathcal{K}-\varepsilon$ and $\mathcal{K}-\omega$ SST models. The

comparison includes prior and DA-optimized versions of the RANS models, as well as experimental results used as observation. For both models, one can see that the accuracy of the predicted u_x field via DA is sensibly improved for each location. Very minor differences when compared with experiments can be observed, which comply with the level of confidence that was prescribed in the observation. On the other hand, the prediction of the normal velocity u_z is very similar to the prior. This is not surprising, considering that for this variable prior results and experimental data compare well, with differences that are of the same order of magnitude of the confidence in the observation or even less. An interesting result can be observed for Point 20, where the maximum difference between the prior and the experimental data is observed for u_z. In this case, one can see that the DA prediction is getting closer to the experiments, confirming that the EnKF is able to provide a statistically more accurate prediction of the flow, within the confidence limit indicated for the different sources of information. Improvement in the prediction is even more clear for the case of DA optimized $\mathcal{K}-\omega$ SST model. In particular, this is visible in the zone near the roof when considering the flow above Point 20 and Point 36. Also, the accuracy of the prediction of the DA normal velocity u_z is significant. For example, the prior $\mathcal{K}-\omega$ SST model profile above Point 36 shows a strong normal velocity component, whereas DA optimized model and experiments show a flow prediction quasi-parallel to the roof.

The features of the velocity field are further assessed in Fig. 7, where streamlines on a vertical plane $x-z$ at the center of the high-rise building are shown. Here, the prior and the DA runs are compared with a validated LES study, as well as results obtained from the optimized $\mathcal{K}-\varepsilon$ model obtained by Zhao et al. [21]. First of all, one can see a qualitative increase in velocity just above the roof separation for the DA runs when compared with the prior RANS simulations. This result, which is closer to the flow predicted by the reference LES, is associated with the improved prediction of the flow that was seen in Fig. 6. The behavior of the recirculation bubble behind the building is now investigated. In this region, present DA runs are only considering pressure observations placed in two rings, positioned in the upper half of the building, as shown in Fig. 2 c). However, the size of this region is also strongly affected by the features of the flow at the top of the building. A significant reduction of the recirculation bubble behind the building is observed, which is obtained thanks to the combined effect of the velocity / pressure information available at the sensors. One can see that prior RANS simulations overpredict the size of the recirculation bubble, in particular for

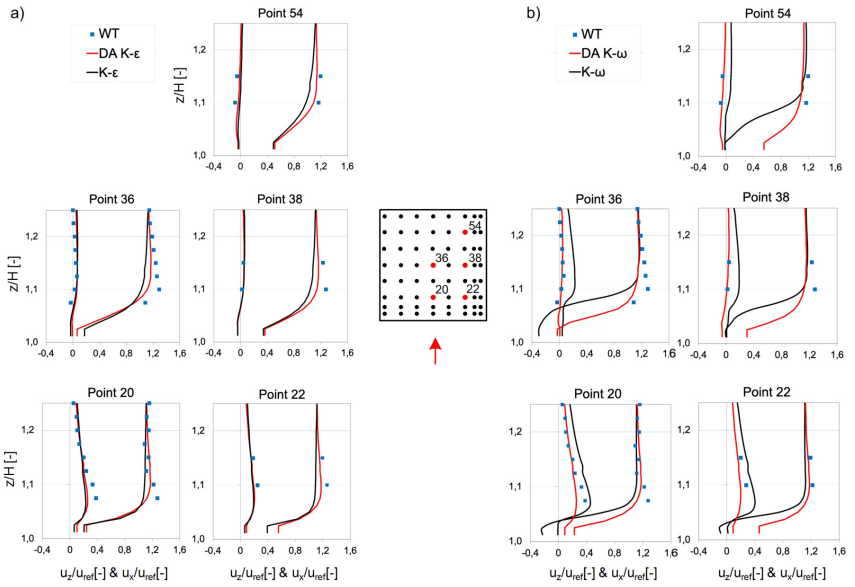


Figure 6: Vertical and streamwise velocity profiles above the marked red locations on the roof (Points: 20, 22, 36, 38, 54): a) comparison between wind tunnel data (WT), $\mathcal{K}-\varepsilon$ model and DA augmented $\mathcal{K}-\varepsilon$ model, b) comparison between wind tunnel data (WT), $\mathcal{K}-\omega$ model and DA augmented $\mathcal{K}-\omega$ model. Here u_{ref} represents the velocity obtained 1 m upstream of the building at height H in every test case ($u_{ref} \approx 13$ m/s in the simulations and $u_{ref} \approx 16$ m/s in the wind tunnel).

the results obtained with the $\mathcal{K}-\omega$ SST model, where no re-attachment of the flow on the roof is observed. The data-driven simulations using heterogeneous observations and the $\mathcal{K}-\varepsilon$ model (present DA run and Zhao et al. model) exhibit in this case a significant reduction, even when compared with the LES. Arguably, the best global representation of the recirculation region is provided by the DA augmented $\mathcal{K}-\omega$ SST model, which exhibits convincing accuracy in particular close to the top of the building. However, its behavior towards the mid-height of the building is not in good agreement with the reference LES. In order to further improve the accuracy of the prediction, one could consider to blend different turbulence models in space, which can be specifically locally tuned. Works in this direction have been recently proposed by Cherroud et al. ([60]) in the framework of machine-learning

reconstruction of turbulence models. This kind of strategy can be integrated into the EnKF in a straightforward manner, even if it obviously requires increased computational resources.

A zoom of the roof area, which is shown in Fig. 8, provides additional information about the prediction of the recirculation region on the roof of the building. One can see that the re-attachment point of the classical RANS $\mathcal{K}-\varepsilon$ model, which is shown by a red arrow, is significantly more upstream than the predicted value by the LES, which is represented by the green arrow. On the other hand, the classical $\mathcal{K}-\omega$ SST model does not provide any reattachment of the boundary layer, which is responsible for the big recirculation bubble, spreading behind the building. The present DA runs significantly improve the prediction of the length of the roof recirculation bubble. On the other hand, the data-driven model by Zhao et al. seems to obtain an almost instantaneous re-attachment of the boundary layer on the roof.

At last, the turbulent kinetic energy field, normalized over the square of the characteristic velocity fluctuation u'_{ref}, is shown in Fig. 9. One can see that both RANS models used for the prior fail to provide the right amount of $\mathcal{K}$ in the big recirculation region behind the building. The $\mathcal{K}-\varepsilon$ model also tends to over-predict the turbulent kinetic energy in front of the building. The present DA-augmented versions of the models provide a significantly better prediction of $\mathcal{K}$. The data-driven model by Zhao et al. also generally improved the prediction for this quantity, but high values for $\mathcal{K}$ are observed in front of the building. This over-prediction is among the factors responsible for the instantaneous reattachment of the recirculation bubble observed on the roof.

In summary, data-driven augmented RANS modeling based on heterogeneous observations proved to provide a significantly better prediction. Although minor differences in results have been observed between the present DA version of the $\mathcal{K}-\varepsilon$ and Zhao et al models, the latter involved a significantly larger number of sensors, which were more homogeneously distributed in the physical domain. Therefore, the present findings highlight the importance of the quality of the information in terms of the location of sensors, and that the quantity of observation may be a secondary factor if sensor placement is efficient.

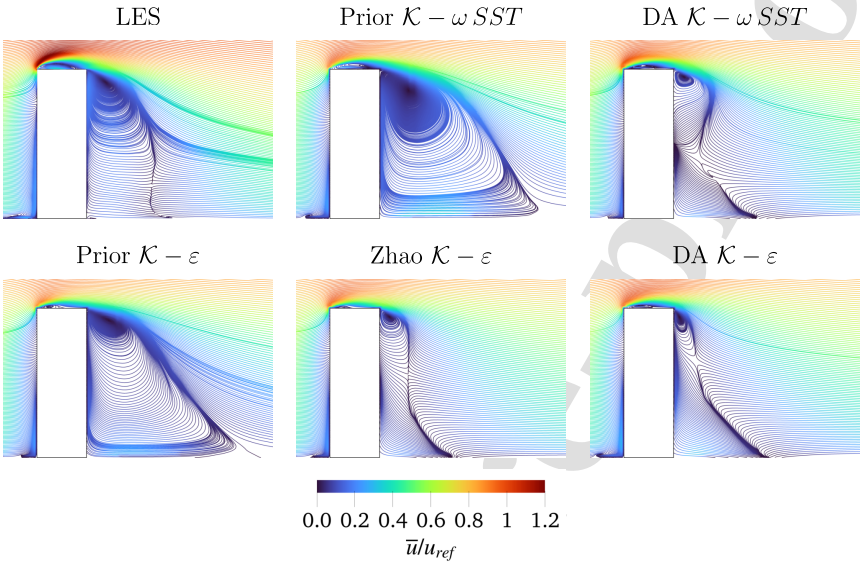


Figure 7: Streamlines colored by the time-averaged velocity field obtained on a $x-z$ plane at the center of the building. Comparisons are performed between a validated Large Eddy Simulation (validated LES) [61], prior and DA augmented $\mathcal{K}-\omega$ SST model, optimized $\mathcal{K}-\varepsilon$ from Zhao et al. [21], prior and DA augmented $\mathcal{K}-\varepsilon$. Here u_{ref} represents the velocity obtained 1 m upstream of the building at height H from the wind tunnel.

4.2. Pressure field

The behavior of the pressure field is now investigated. This physical quantity is significantly more difficult to predict for incompressible numerical simulation because the Poisson equation resolved in the CFD solver uses the pressure as a Lagrangian multiplier. Therefore, the analysis of this quantity is crucial to assess the stability and the precision of the algorithms. The mean pressure coefficient is defined as $C_p\left(C_p = \frac{\bar{p}-p_{ref}}{0.5\rho u_{ref}^2}\right)$, where p_{ref}, ρ and u_{ref} are the free-stream pressure, the air density, and the reference velocity (calculated 1 m upstream of the building at height H in each simulation), respectively. In Fig. 10 the mean pressure coefficient is shown in terms of performance metrics comparison with experimental data. This way, the local difference between each numerical simulation and the experiments is clustered in groups that do not exceed a prescribed error threshold (10%,

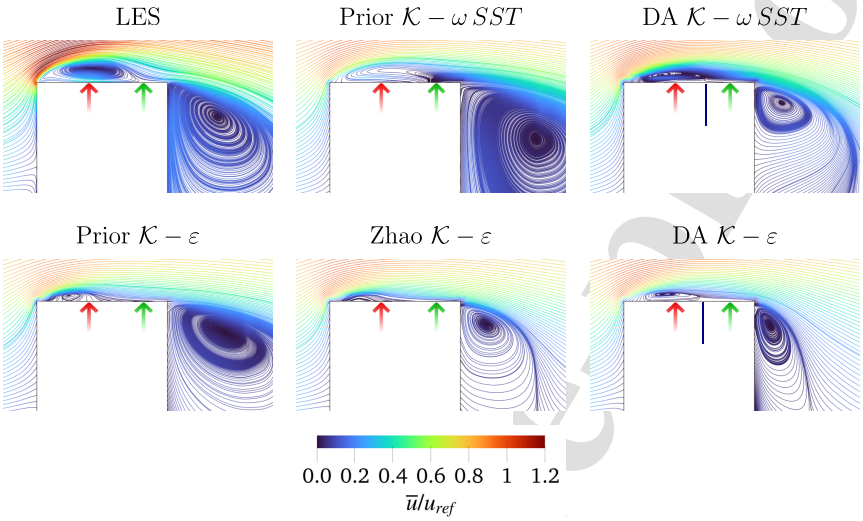


Figure 8: Zoom of the flow streamlines above the rooftop. The red arrow indicates the reattachment position of the recirculation bubble for the classical $\mathcal{K}-\varepsilon$ model, and the green arrow provides the same information for the validated LES simulation. The blue line indicates the reattachment position for the simulation shown in each image. Here u_{ref} represents the velocity obtained 1 m upstream of the building at height H from the wind tunnel.

20%, and 30% in this case). The analysis of this criterion for the error threshold 10% may erroneously lead to the conclusion that prior simulations (for example, 23% of the occurrences for $\mathcal{K}-\omega$ SST model) behave better than the DA runs (16% of the occurrences). This information is misleading, though, as a large number of occurrences for the DA runs are just outside this interval. In fact, as large margins of error are considered, one can see that the DA runs outperform the prior RANS. For a 20% error threshold, an improvement up to 17% occurrences is observed with the use of DA. The improvement increases to around $10\% - 30\%$ when a 30% error threshold is considered. For the latter threshold, the DA augmented $\mathcal{K}-\omega$ SST model reaches more than 70% limit which certifies the global high accuracy of the method. The present results are also plotted in the form of a histogram in Fig. 11. This representation confirms that results obtained by present DA runs tend to cluster towards lower error regions when compared with classical

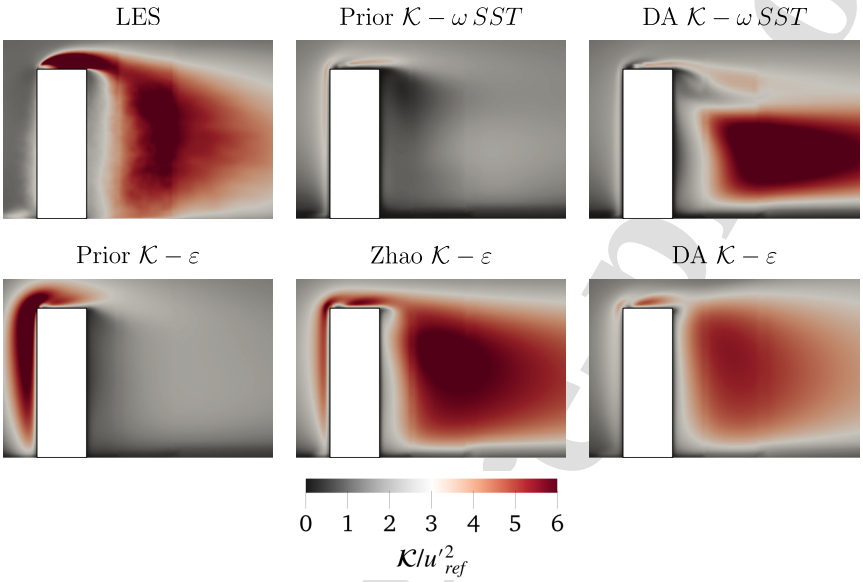


Figure 9: Turbulent kinetic energy $\mathcal{K}$ obtained on a $x-z$ plane at the center of the building. Comparisons are performed between a validated Large Eddy Simulation (validated LES) [61], prior and DA augmented $\mathcal{K}-\omega$ SST model, optimized $\mathcal{K}-\varepsilon$ from Zhao et al. [21], prior and DA augmented $\mathcal{K}-\varepsilon$. Here $u'_{ref}=I_u u_{ref}$ is obtained from the streamwise turbulent intensity $I_u=13\%$ and u_{ref} measured 1 m upstream of the building at height H.

RANS. However, this aspect deserves significantly more investigation because results obtained with the data-driven models obtained by Ben-Ali et al. and Zhao et al., which are not presented here, show a degraded accuracy for the predicted pressure field. This could potentially imply that heterogeneous observation may be useful to improve the global prediction of the numerical solver, probably because of the correlation between physical variables which is driven by non-linear dynamics.

Finally, the comparison of the mean pressure coefficient C_p calculated at pressure taps on the roof in experiments and RANS calculations is shown in Fig. 12. Results for the $\mathcal{K}-\omega$ SST model indicate again a significant improvement of the prediction. The RANS version used as a prior exhibits a quasi-constant value of C_p for the whole length x/B, as the boundary

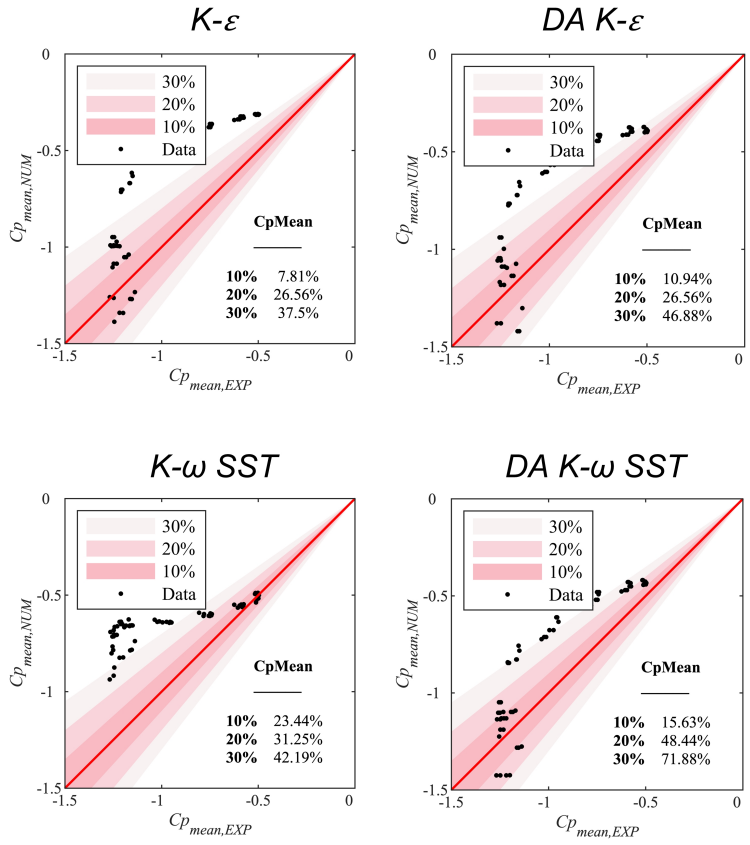


Figure 10: Scatter plot of mean pressure coefficient C_p for numerical simulations with performance metrics-comparison with experimental data. Top row: $\mathcal{K}-\varepsilon$ model, bottom row: $\mathcal{K}-\omega$ SST. Left column: prior RANS model, right column: DA augmented RANS.

layer does not reattach to the roof. On the other hand, the DA augmented $\mathcal{K}-\omega$ SST model exhibits a significantly improved prediction, following much more closely the experimental data. Namely, both of these C_p distributions exhibit a "hump" shape, typical for a flow with a separated region followed by a reattachment [62]. This observation is in line with the flow pattern observed over the rooftop in Fig. 8. For the DA run using the $\mathcal{K}-\varepsilon$ model,

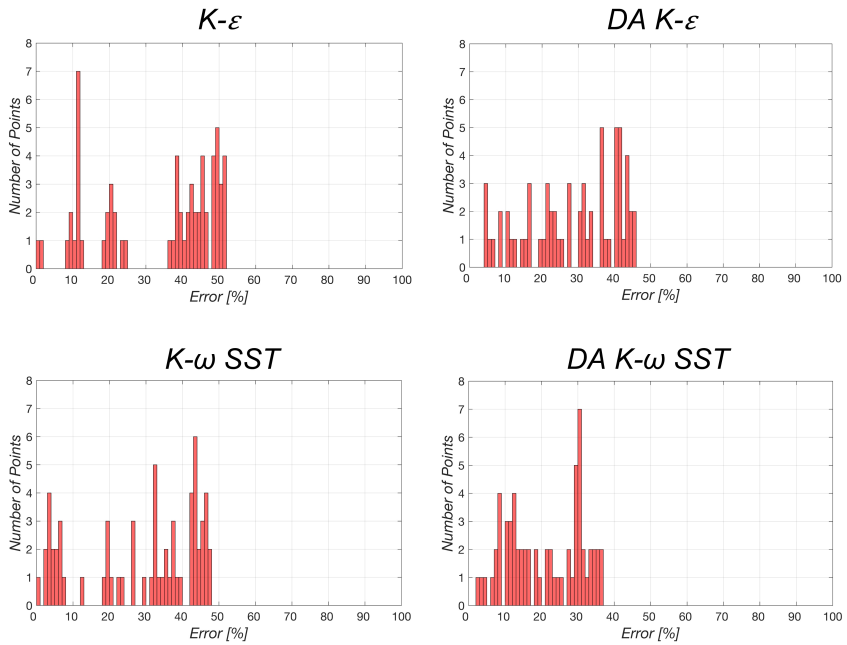


Figure 11: Histograms representing the error in the prediction of the pressure coefficient C_p when compared with experiments. Numerical results are sampled in the locations of the pressure taps. Top row: $\mathcal{K}-\varepsilon$ model, bottom row: $\mathcal{K}-\omega$ SST. Left column: prior RANS model, right column: DA augmented RANS.

improvements are also observed, but they appear to be a minor shift from the prior RANS model. Therefore, the magnitude of such improvements is not as important as for the DA calculation based on the $\mathcal{K}-\omega$ SST model.

5. Conclusions

The newly developed platform CONES has been used to perform a data-driven investigation of the flow around a high-rise building. More precisely, heterogeneous experimental samples, in the form of data from pressure taps and hot wires, have been integrated with RANS CFD runs, performed using the open-source code OpenFOAM. The coupling has been performed using techniques based on the Ensemble Kalman Filter (EnKF), including

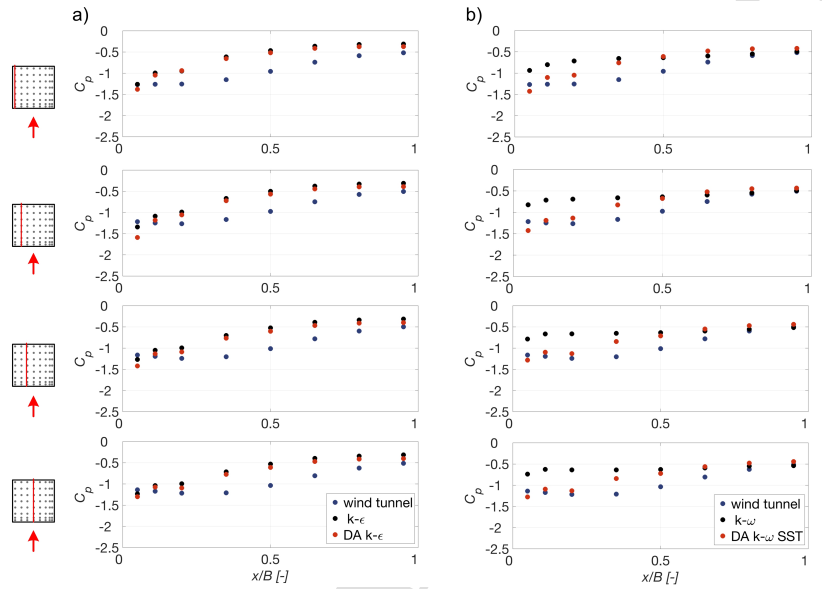


Figure 12: Mean pressure coefficient at pressure tap locations along red lines marked at the roof: a) comparison between wind tunnel data (WT), $\mathcal{K}-\varepsilon$ model and DA augmented $\mathcal{K}-\varepsilon$ model; b) comparison between wind tunnel data (WT), $\mathcal{K}-\omega$ SST and DA augmented $\mathcal{K}-\omega$ SST.

advanced manipulations such as localization and inflation. The augmented state estimation obtained via EnKF has also been employed to improve the predictive features of the model by optimization of the five/nine free global model constant of the $\mathcal{K}-\varepsilon/\mathcal{K}-\omega$ SST turbulence models, respectively, used to close the equations.

The results have shown that a global improvement has been observed for the physical quantities of investigation, and the results obtained with the different DA strategies are equivalent. For this last point, physical and covariance localization, which have been compared for the $\mathcal{K}-\varepsilon$ model, appear to be effective for the study of complex flows. The reduction of degrees of freedom of the DA problem has not affected the quality of the results, while globally reducing the time needed for the data-driven procedures. On the other hand, the usage of inflation has not produced better results, in particular, due to the increase of computational resources required.

The analysis of the velocity field shows that the EnKF allows significantly reducing the error in the streamwise and vertical direction, according to the confidence provided for the observation. The effects of the parametric inference are observed also in the recirculation region behind the building. In this case, the accuracy of the results is affected by the RANS model chosen for the optimization. The physical topology of the flow becomes more similar to the reference LES validated with experimental data, even if the recirculation bubble is overly reduced in size. For the pressure field, improvements are observed as the error in the prediction of the mean pressure coefficient globally reduces. In the roof area, the improvement of the statistical behavior of pressure is tied to the increased accuracy in the estimation of the reattachment of the recirculation bubble on top. The present results have been obtained using tools available for every user of a CFD code, which is a segregated structure and a global description of the coefficients controlling turbulence modeling. Potentially, more sophisticated coupled solvers could provide improved results when used in DA tools using pressure data as observation.

Future investigations include more complex parametric descriptions of the turbulence modeling employed, including coupling between DA tools and machine learning applications. Additionally, in the subsequent stages of this research, the sensor placement problem will be given due consideration. This encourages the vision of developing a highly efficient DA on-the-fly designed tool where fields will be continuously updated as new information becomes available from sensing devices, fostering a more adaptive and responsive approach.

Acknowledgement

Our research activities are supported by the funding of the French Agence Nationale de la Recherche (ANR) through projects PRC 2020 ALEKCIA and JCJC 2021 IWP-IBM-DA. This work was granted access to the HPC resources of GENCI in the framework of allocation A12 for project A0122A01741 on the IRENE supercomputer (TGCC). Florent Duchaine and Miguel Ángel Moratilla-Vega are warmly acknowledged for the help provided during the early stages of the development of CONES. The co-author A.S.G. would like to acknowledge the support of the “Fonds National de la Recherche, Luxembourg” (FNR) for funding the CORE Junior project DATA4WIND - “Data-Driven Approach for Urban Wind Energy Harvesting”, C19/SR/13639741.

References

[1] S. B. Pope, Turbulent flows, Cambridge University Press, 2000.

[2] D. C. Wilcox, Turbulence Modeling for CFD, 3rd Edition, DCW Industries Inc., La Canada CA, 2006.

[3] B. Blocken, Computational fluid dynamics for urban physics: Importance, scales, possibilities, limitations and ten tips and tricks towards accurate and reliable simulations, Building and Environment 91 (2015) 219–245.

[4] D. Simon, Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches, Wiley, 2006.

[5] S. B. Daley, Atmospheric Data Analysis, Cambridge University Press, 1991.

[6] R. E. Kalman, A new approach to linear filtering and prediction problems, Journal of Basic Engineering 82 (1960) 35–45.

[7] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, Springer-Verlag/Berlin/Heildelberg, 2009.

[8] M. Asch, M. Bocquet, M. Nodet, Data Assimilation: Methods, Algorithms, and Applications, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016. doi:10.1137/1.9781611974546.

[9] H. Xiao, P. Cinnella, Quantification of Model Uncertainty in RANS Simulations: A Review, Progress in Aerospace Sciences 108 (2019) 1–31.

[10] C. Gorlé, G. Iaccarino, A framework for epistemic uncertainty of turbulent scalar flux models for Reynolds-averaged Navier-Stokes simulations, Physics of Fluids 25 (2013) 055105.

[11] L. Margheri, M. Meldi, M. V. Salvetti, P. Sagaut, Epistemic uncertainties in RANS model free coefficients, Computers & Fluids 102 (2014) 315–335.

[12] H. Xiao, J. L. Wu, J. X. Wang, R. Sun, C. Roy, Quantifying and reducing model-form uncertainties in Reynolds-averaged Navier–Stokes simulations: A data-driven, physics informed Bayesian approach, Journal of Computational Physics 324 (2016) 115–136.

[13] X. Zhang, H. Xiao, T. Gomez, O. Coutier-Delgosha, Evaluation of ensemble methods for quantifying uncertainties in steady-state CFD applications with small ensemble sizes, Computers & Fluids 203 (2020) 104530.

[14] K. Duraisamy, H. Xiao, G. Iaccarino, Turbulence modeling in the age of data, Annual Review of Fluid Mechanics 51 (2019) –.

[15] M. Schmelzer, R. Dwight, C. P., Discovery of algebraic Reynolds-stress models using sparse symbolic regression, Flow, Turbulence and Combustion 104 (2020) 579–603.

[16] R. D. Sandberg, Y. Zhao, Machine-learning for turbulence and heat-flux model development: A review of challenges associated with distinct physical phenomena and progress to date, International Journal of Heat and Fluid Flow 95 (2022) 108983.

[17] A. Glumac, O. Jadhav, V. Despotović, B. Blocken, S. Bordas, A multi-fidelity wind load assessment via machine learning: a high-rise building case. (2023) 110135.

[18] Y. Tominaga, A. Mochida, R. Yoshie, H. Kataoka, T. Nozu, M. Yoshikawa, T. Shirasawa, Aij guidelines for practical applications of cfd to pedestrian wind environment around buildings, Journal of Wind Engineering and Industrial Aerodynamics 96 (2008) 1749–1761.

[19] H. Hemida, A. Šarkić Glumac, G. Vita, K. K. Vranešević, R. Höffer, On the flow over high-rise building for wind energy harvesting: An experimental investigation of wind speed and surface pressure, Applied Sciences 10 (2020) 5283.

[20] M. Ben Ali, G. Tissot, S. Aguinaga, D. Heitz, E. Mémin, Mean wind flow reconstruction of a high-rise building based on variational data assimilation using sparse pressure measurements, Journal of Wind Engineering and Industrial Aerodynamics 231 (2022) 105204.

[21] R. Zhao, S. Liu, J. Lie, N. Jiang, Q. Chen, Generalizability evaluation of k-epsilon models calibrated by using ensemble kalman filtering for urban airflow and airborne contaminant dispersion, Building and Environment 212 (2022) 108823.

[22] C. Semeraro, M. Lezoche, H. Panetto, M. Dassisti, Digital twin paradigm: A systematic literature review, Computers in Industry 130 (2021) 103469.

[23] V. Mons, J. C. Chassaing, P. Sagaut, Optimal sensor placement for variational data assimilation of unsteady flows past a rotationally oscillating cylinder, Journal of Fluid Mechanics 823 (2017) 230–277.

[24] V. Mons, O. Marquet, Linear and nonlinear sensor placement strategies for mean flow reconstruction via data assimilation, Journal of Fluid Mechanics 923 (2021) A1.

[25] F. Jorgensen, How to measure turbulence with hot-wire anemometers – a practical guide, Tech. rep., Dantec Dynamics (2004).

[26] OpenFOAM - Field Operation And Manipulation, `https://www.openfoam.com`.

[27] J. Ferziger, M. Peric, Computational Methods in Fluid Dynamics, New-York : Springer-Verlag, 1996.

[28] M. Meldi, M. V. Salvetti, P. Sagaut, Quantification of errors in large-eddy simulations of a spatially evolving mixing layer using polynomial chaos, Physics of Fluids 24 (2012) 035101.

[29] E. Constant, J. Favier, M. Meldi, P. Meliga, E. Serre, An immersed boundary method in openfoam : verification and validation, Computers and Fluids 157 (3) (2017) 55 – 72.

[30] B. Launder, B. Sharma, Application of the energy-dissipation model of turbulence to the calculation of flow near a spinning disc, Letters in Heat and Mass Transfer 1 (1974) 131–137.

[31] F. R. Menter, M. Kuntz, R. Langtry, Ten Years of Industrial Experience with the SST Turbulence Model, Turbulence, Heat and Mass Transfer 4 (2003).

[32] F. R. Menter, M. Kuntz, R. Langtry, Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications, AIAA Journal 32 (8) (1993) 1598–1605.

[33] L. Margheri, M. Meldi, M. V. Salvetti, P. Sagaut, Epistemic uncertainties in rans model free coefficients, Computers & Fluids 102 (2014) 315–335. doi:http://dx.doi.org/10.1016/j.compfluid.2014.06.029.

[34] .-.-. EN, Eurocode 1: Actions on structures-Part 1-4: General actions - Wind actions, Eur. Comm. Stand., 2005.

[35] T. Norton, J. Grant, R. Fallon, D.-W. Sun, Optimizing the ventilation configuration of naturally ventilated livestock buildings for improved indoor environmental homogeneity, Build Environ. 45 (2010) 983e95.

[36] R. Ramponi, B. Blocken, Cfd simulation of cross-ventilation for a generic isolated building: impact of computational parameters, Building and Environment 53 (2012) 34–48.

[37] C. Neuhaus, Numerische frequenzabhaengige kalibrierung langer druckmessschlauchsysteme, internal report, wind engineering and flow mechanics. ruhr-universit€at bochum, germany (2010).

[38] D. P. G. Foures, N. Dovetta, D. Sipp, P. J. Schmid, A data-assimilation method for Reynolds-averaged Navier-Stokes-driven mean flow reconstruction, Journal of Fluid Mechanics 759 (2014) 404–431.

[39] M. C. Rochoux, S. Ricci, D. Lucor, B. Cuenot, A. Trouve, Towards predictive data-driven simulations of wildfire spread - Part I: Reduced-cost Ensemble Kalman Filter based on a Polynomial Chaos surrogate model for parameter estimation, Natural Hazards and Earth System Sciences 14 (2015) 2951–2973.

[40] M. Meldi, A. Poux, A reduced order model based on Kalman Filtering for sequential Data Assimilation of turbulent flows, Journal of Computational Physics 347 (2017) 207–234.

[41] M. Meldi, Augmented Prediction of Turbulent Flows via Sequential Estimators: Sensitivity of State Estimation to Density of Time Sampling for Available Observation, Flow, Turbulence and Combustion 101 (2018) 389–412.

[42] J. W. Labahn, H. Wu, B. Coriton, J. H. Frank, M. Ihme, Data assimilation using high-speed measurements and LES to examine local extinction events in turbulent flames, Proceedings of the Combustion Institute 37 (2019) 2259–2266.

[43] P. Chandramouli, E. Memin, D. Heitz, 4D large scale variational data assimilation of a turbulent flow with a dynamics error model, Journal of Computational Physics 412 (2020) 109446.

[44] G. Moldovan, G. Lehnasch, L. Cordier, M. Meldi, A multigrid/ensemble Kalman filter strategy for assimilation of unsteady flows, Journal of Computational Physics 443 (2021) 110481.

[45] A. Carrassi, M. Bocquet, L. Bertino, G. Evensen, Data assimilation in the geosciences: An overview of methods, issues, and perspectives, WIREs Climate Change 9 (2018). doi:10.1002/wcc.535.

[46] G. Evensen, P. J. Van Leeuwen, An Ensemble Kalman Smoother for Nonlinear Dynamics, Monthly Weather Review 128 (6) (Jun. 2000).

[47] G. Artana, A. Camilleri, J. Charlier, E. Memin, Strong and weak constraint variational assimilations for reduced order fluid flow modeling, Journal of Computational Physics 231 (2014) 3264–3288.

[48] X. Zhang, C. Michelin-Ströfer, H. Xiao, Regularized ensemble Kalman methods for inverse problems, Journal of Computational Physics 416 (2020) 109517.

[49] G. Evensen, The ensemble Kalman Filter for combined state and parameter estimation - Monte Carlo techniques for data assimilation in large systems, IEEE Control Systems 29 (2009) 83–104.

[50] I. Hoteit, D.-T. Pham, M. E. Gharamti, X. Luo, Mitigating Observation Perturbation Sampling Errors in the Stochastic EnKF, Monthly Weather Review 143 (7) (2015) 2918–2936. doi:10.1175/MWR-D-14-00088.1.

[51] A. Reflox, B. Courbet, A. Murrone, P. Villedieu, C. Laurent, et al., CEDRE Software, Aerospace Lab 2 (2011) 1–10.
URL https://hal.archives-ouvertes.fr/hal-01182463

[52] F. Duchaine, et al., Analysis of high performance conjugate heat transfer with the OpenPALM coupler, Computational Science & Discovery 8 (2015) 015003. doi:https://doi:10.1088/1749-4699/8/1/015003.

[53] P. Legrenzi, K. V. Karpaga, P. J. Gary, T. H. Indi, Simple and Robust Framework for Coupled Aerothermal Gas Turbine Simulation using Low-Mach and Compressible Industrial CFD Solvers, American Institute of Aeronautics and Astronautics 1640 (2016). doi:https://doi.org/10.2514/6.2016-1640.

[54] M. A. Moratilla-Vega, M. Angelino, H. Xia, G. J. Page, An open-source coupled method for aeroacoustics modelling, Computer Physics Communications 278 (2022) 108420. doi:https://doi.org/10.1016/j.cpc.2022.108420.

[55] E. Leonard, H. Qiao, N. Saleh, A comparison of interpolation methods in fast fluid dynamics, in: 6th International High Performance Buildings Conference, Vol. 341, 2021.
URL https://docs.lib.purdue.edu/ihpbc/

[56] M. Katzfuss, J. R. Stroud, C. K. Wikle, Understanding the ensemble kalman filter, The American Statistician 70 (2016). doi:10.1080/00031305.2016.1141709.

[57] V. Mons, Y. Du, T. Zaki, Ensemble-variational assimilation of statistical data in large-eddy simulation, Physical Review Fluids 6 (2021) 104607.

[58] T. S. T. Potsis, Y. Tominaga, Computational wind engineering: 30 years of research progress in building structures and environment, J. Wind Eng. Ind. Aerodyn 234 (2023) 105346.

[59] K. Kostadinović-Vranešević, G. Vita, S. P. A. Bordas, A. Šarkić Glumac, Furthering knowledge on the flow pattern around high-rise1 buildings: Les investigation of the wind energy potential, J. Wind Eng. Ind. Aerodyn 226 (2022) 105029.

[60] S. Cherroud, X. Merle, P. Cinnella, X. Gloerfelt, Space-dependent aggregation of data-driven turbulence models (2023).

[61] K. K. Vranešević, G. Vita, S. P. Bordas, A. Š. Glumac, Furthering knowledge on the flow pattern around high-rise buildings: LES investigation of the wind energy potential, Journal of Wind Engineering and Industrial Aerodynamics 226 (2022) 105029. doi:10.1016/j.jweia.2022.105029.

[62] F. Haan, The effect of turbulence on the aerodynamics of the long-span bridges, Ph.D. thesis, Department of Aerospace and Mechanical Engineering, University of Notre Dame (2000).

[63] G. Moldovan, A. Mariotti, G. Lehnasch, L. Cordier, M. Salvetti, M. Meldi, Data-driven augmented LES for the analysis of the BARC test case, ArXiv - (2022) 1–26.

Appendix A. DA experiments - formulation of the EnKF

In this appendix, several variations of the implementation of the EnKF are tested. The objective is to assess the robustness of the DA model, as well as to select the best strategy in terms of the ratio between accuracy and costs to be employed in the two DA analyses. This preliminary investigation has been performed for the $\mathcal{K}-\varepsilon$ model only. Three independent DA experiments are performed. The variations do not deal with details of the model or the observation, but they consider different features of the DA procedure. More precisely, the cases analyzed are:

- Case A: classical EnKF.
- Case B: EnKF with covariance localization.
- Case C: EnKF with covariance localization and inflation.

Parameter	$\mathcal{K}-\varepsilon$ **model default values**	**Prior of the EnKF**		
		$\mu_{\mathbf{N}}$	$\sigma_{\mathbf{N}}$ for cases A,B	$\sigma_{\mathbf{N}}$ for case C
C_μ [-]	0.09	0.1	0.01	0.005
$C_{\varepsilon 1}$ [-]	1.44	1.575	0.1	0.05
$C_{\varepsilon 2}$ [-]	1.92	1.9	0.1	0.05
$\sigma_{\mathcal{K}}$ [-]	1.0	1.0	0.1	0.05
σ_ε [-]	1.3	1.6	0.1	0.05

Table A.2: Comparison between conventional constants from RANS $\mathcal{K}-\varepsilon$ model and the initial parameters employed for the EnKF ($N_e = 40$).

Preliminary analyses have been performed to identify a suitable initial configuration for the hyperparameters driving the performance of the DA algorithm. The initial normal distributions reported in Tab. A.2 are bounded between σ_N and $7\sigma_N$, the limit has been empirically set depending on the sensitivity of the coefficients. For example, $C_{\varepsilon 1}$ is bounded by $1.25\sigma_N$ but σ_ε is bounded by $7\sigma_N$. The initial physical state for each ensemble is obtained from a single run using the values of the model constants in Ref. [11].

For case studies A and B, 150 iterative steps are performed between successive analysis phases. For study C the number of iterative steps has been lowered to 100 considering additional analysis phases needed due to inflation.

The number of iterations has been chosen observing results from preliminary analyses, which pointed out how at least 50 iterations were needed to obtain a complete signature of the new parametric setting over the physical quantities and thus reach convergence.

For localization in cases B and C, the domain is also clipped in a volume sufficiently large around the observations, according to the proposals by Moldovan et al. [63] for the BARC geometry. When the number of sensors/observations available is limited, the bottleneck in terms of computational costs for the EnKF is represented by the products involving the anomaly matrix. Considering that the size of the state matrix is tied to the number of degrees of freedom simulated by the model, a physical localization (clipping) can reduce the global computational costs and ensure the stability of the algorithm. The state estimation is applied only in the control volume shown in Fig. A.13, containing 179568 mesh elements, which means 35% of the total number of cells. Also, the coefficients η of the covariance localization (see Eq. 31) are specifically selected for each space direction so that the discontinuity of the velocity field at the boundary of the clipping region is equal or lower to 0.3%, ensuring continuity of the physical solutions. Taking this criterion into account, $\eta_x = \eta_z = 0.0195$ and $\eta_y = 0.0438$.

The history of the optimization of the model coefficients for the three DA runs is now commented. It is important to stress that, despite some important differences observed in the optimized sets of model constants obtained with the three strategies, the prediction of the flow is very similar. This implies that the differences observed are in a region where the sensitivity of the solution to the parameters is low i.e. relatively large parametric variations correspond to small physical changes.

Appendix A.1. Case A: classical EnKF

In this first case, the classical EnKF is used. The run is ended when a suitable convergence of the parameters is reached, which is in this case after 100 analysis phases (i.e. a total of 15000 CFD iterations). The evolution of the mean value of the five parameters of the $\mathcal{K} - \varepsilon$ model is shown in Fig. A.14. One can see that the final results obtained by the EnKF are significantly different than the baseline values and that the speed at which the parameters converge is significantly different. In particular, the evolution of σ_ε deserves some comments. This coefficient controls the magnitude of the turbulent diffusion term in the equation for ε, $D_\varepsilon = \nu_t/\sigma_\varepsilon + \nu$, which is associated with non-homogeneous conditions (see Sec. 2). The optimization

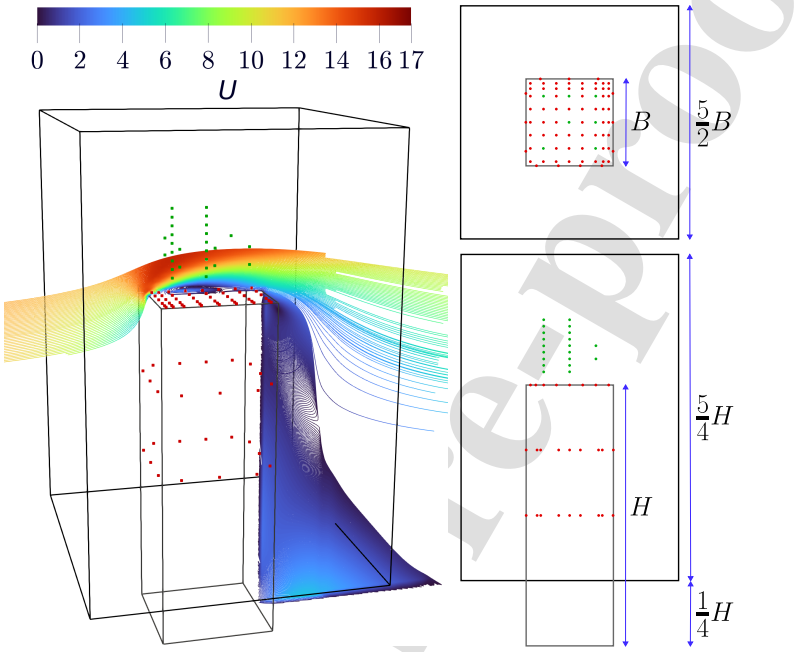


Figure A.13: Clipping box used for localization: pressure sensors are represented in red and velocity sensors are displayed in green.

performed by the EnKF targets very low values for σ_ε during the calculation, increasing the relevance of D_ε in the equation. However, noise propagated by the Kalman gain can turn the value for some ensemble members to be negative, resulting in a divergence of the calculation. Therefore, a constraint has been imposed so that values cannot be lower than a small but positive value prescribed. For the other parameters, one can see that C_μ and $C_{\varepsilon 1}$ converge to a value close to 1/3 of the initial estimate, $\sigma_\mathcal{K}$ does not exhibit large variations and $C_{\varepsilon 2}$ is significantly larger. Also, this last parameter does not seem fully converged. Comparisons between results obtained with the three runs, which are shown in Tab. A.3, indicate a large variability of this coefficient. The analysis of the physical results seems to indicate that, for $C_{\varepsilon 2} > 4$, the solution exhibits very low sensitivity to variations of this parameter.

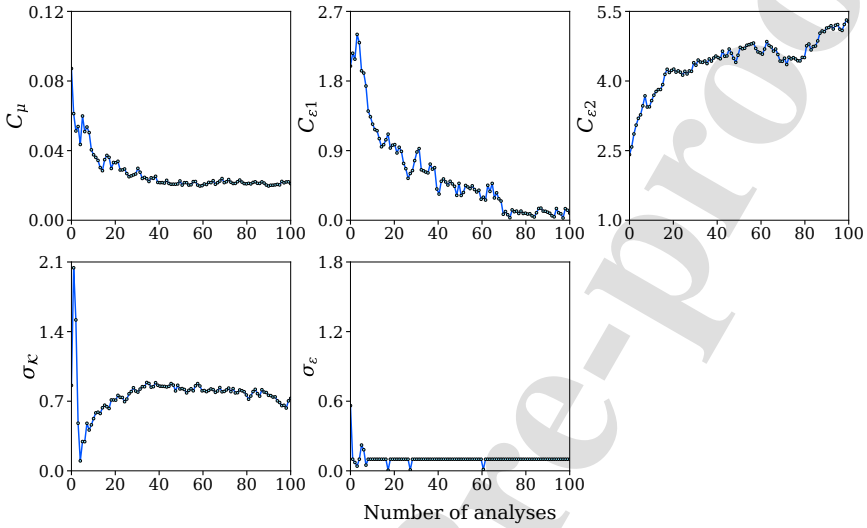


Figure A.14: Evolution of the $\mathcal{K}-\varepsilon$ model coefficients. The DA strategy used is the EnKF without inflation and localization.

Appendix A.2. Case B: EnKF with covariance localization

In this case, the calculation is performed with covariance localization. The evolution of the five coefficients is shown in Fig. A.15. The trend and in particular the evolution of σ_ε are similar to the ones observed for Case A. Remarks provided for case A, in particular for the evolution of σ_ε, are valid as well for this run. One should take into account that covariance and physical localization reduce the computational requirements of the DA analysis, therefore this case is computationally more efficient and it suggests that these techniques may be very efficient for CFD calculations with a very large number of degrees of freedom.

Appendix A.3. Case C: EnKF with both inflation and localization

The DA calculation is performed here relying on deterministic inflation for the model's parameters and covariance localization. This is the most advanced run in terms of the complexity of the DA algorithm. The evolution of the five parameters is shown in Fig. A.16. To ensure the robustness of the simulation during the first time steps, the inflation quantifier λ is gradually increased from 1.05 to 1.3 and, later, removed to obtain the convergence

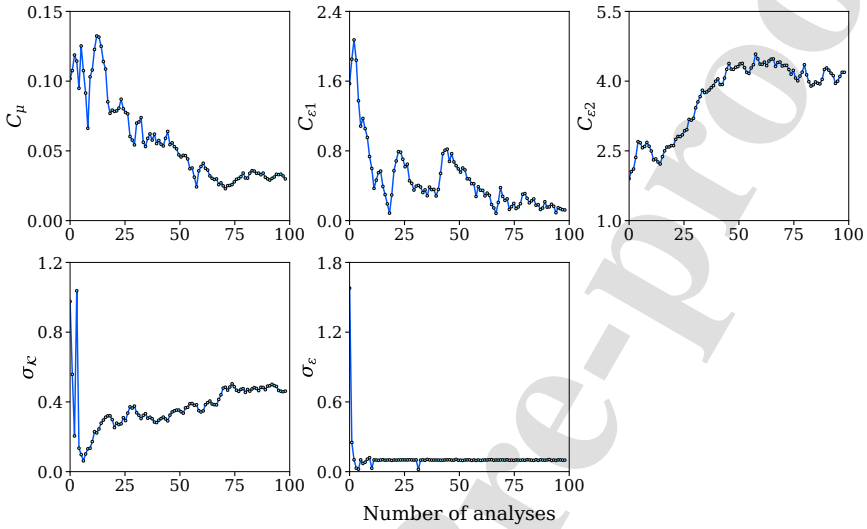


Figure A.15: Evolution of the $\mathcal{K}-\varepsilon$ model coefficients. The DA strategy used is the EnKF without inflation but with localization.

($\lambda = 1.05$ for $k \in [1, 40]$, $\lambda = 1.1$ for $k \in [41, 120]$, $\lambda = 1.2$ for $k \in [121, 160]$, $\lambda = 1.3$ for $k \in [161, 200]$, and $\lambda = 1$ for $k > 200$). Some coefficients such as C_μ and $\sigma_\mathcal{K}$ show a higher sensitivity to changes in the value of λ, highlighting the importance of inflation in identifying a suitable large parametric space for the optimization. For this reason, convergence is reached significantly later in this case. Also, the threshold value for σ_ε is increased here, in order to avoid stability problems that could be easily triggered by the higher variability associated with the parametric inflation.

The impact of the physical prediction of the three different parametric descriptions, which are reported in Tab. A.3, are investigated in Sec. 4.

Some remarks should be made about the performance of the three DA runs. Despite the differences in the techniques used and the apparently different results obtained for the parameter optimization, the prediction of the physical variables is pretty similar. The values of the model parameters have probably converged towards a robust optimum, where the sensitivity of the solution to further parametric variation is very low. This aspect, which needs further investigation, may indicate that robust optimization can be

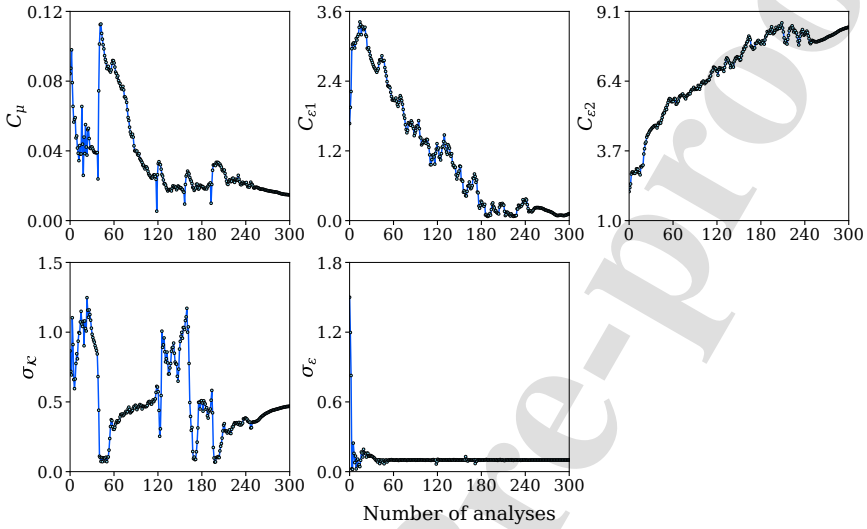


Figure A.16: Evolution of the $\mathcal{K}-\varepsilon$ model coefficients. The DA strategy used is the EnKF with inflation and localization.

Parameter	Optimized values		
	Case A	**Case B**	**Case C**
C_μ [-]	0.021	0.032	0.015
$C_{\varepsilon 1}$ [-]	0.091	0.165	0.152
$C_{\varepsilon 2}$ [-]	5.278	4.080	8.574
$\sigma_\mathcal{K}$ [-]	0.729	0.476	0.477
σ_ε [-]	0.1	0.1	0.1

Table A.3: Optimized $\mathcal{K}-\varepsilon$ model coefficients obtained with different strategies based on the EnKF.

obtained by setting a suitable confidence interval for the observation. In this scenario, the application of localization has proven effective. The reduction of degrees of freedom in the DA process, which significantly decreases the computational resources required for each analysis phase, is not responsible for the degradation of the results. On the other hand, probably because of the features of the parametric optimum region found, the inflation techniques have not improved the results.

Augmented state estimation of urban settings using on-the-fly sequential Data Assimilation

Highlights:

- Sequential data assimilation is used to infer optimized parametric description of the $K-\varepsilon$ and $K-\omega SST$ turbulence closure when applied to the analysis of a flow configuration related to urban wind engineering.
- Heterogeneous experimental data is used for this model calibration, including pressure and velocity measurements.
- A data assimilation library, CONES, is developed to perform online EnKF, significantly reducing the computational costs required.
- Advanced data-driven closures improve the flow prediction, in particular in the case of $K-\omega SST$ turbulence model.

Augmented state estimation of urban settings using on-the-fly sequential Data Assimilation - author statement

- L. Villanueva: Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization
- M. M. Valero: Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization
- A. Šarkić Glumac: Software, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Funding acquisition
- M. Meldi: Conceptualization, Methodology, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding acquisition

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: