




# Ensemble Kalman Filter for Assimilating Experimental Data into Large-Eddy Simulations of Turbulent Flows

Jeffrey W. Labahn<sup>1</sup>  · Hao Wu<sup>1</sup> · Shaun R. Harris<sup>1</sup> · Bruno Coriton<sup>2</sup> · Jonathan H. Frank<sup>2</sup> · Matthias Ihme<sup>1</sup>

Received: 27 March 2019 / Accepted: 9 October 2019 /  
© Springer Nature B.V. 2019

## Abstract

Data assimilation techniques are investigated for integrating high-speed high-resolution experimental data into large-eddy simulations. To this end, an ensemble Kalman filter is employed to assimilate velocity measurements of a turbulent jet at a Reynolds number of 13,500 into simulations. The goal of the current work is to examine the behavior of the assimilation algorithm for state estimation of turbulent flows that are of relevance to engineering applications. This is accomplished by investigating the impact that localization, measurement uncertainties, assimilation frequency, data sparsity and ensemble size have on the estimated state vector. For the flow configuration and computational setup considered in this study an optimal value of the localization radius is identified, which minimizes the error between experimental data and state vector. The impact of experimental uncertainties on the state estimation is demonstrated to provide solution bounds on the assimilation algorithm. It is found that increasing the number of ensembles has a positive impact on the state estimation. In comparison, decreasing the assimilation frequency or reducing the experimental data available for assimilation is found to have a negative impact on the state estimation. These findings demonstrate the viability of assimilating measurements into numerical simulations to improve state estimates, to support parameter evaluations and to guide model assessments.

**Keywords** Data assimilation · High-speed experimental data · Large-eddy simulation

## 1 Introduction

There has been increasing interest in numerically studying time-dependent phenomena, such as transition and separation in turbulent flows [1]. However, the ability of large-eddy simulation (LES) to reproduce experimentally observed time-dependent phenomena is limited for several reasons. First, LES is inherently a stochastic representation of a

---

✉ Jeffrey W. Labahn  
jwllabahn@gmail.com

<sup>1</sup> Department of Mechanical Engineering, Stanford University, Stanford, CA 94305, USA

<sup>2</sup> Combustion Research Facility, Sandia National Laboratories, Livermore, CA, USA

time-dependent process and thus cannot be directly compared to a specific experimentally observed phenomena. Second, specifications of boundary conditions and required closure models introduce sources of uncertainty, which along with the chaotic nature of turbulence, limits the capability of LES to capture an experimentally observed event [2]. Thus, a new approach is required to numerically investigate rare or stochastic events.

One such approach is to assimilate experimental data into simulations so that the resulting predictions are improved over those produced by either method on their own [3]. This is accomplished by considering errors associated with the numerical models and experimental data and determining the state vector which best satisfies the information obtained from the numerical model and experimental data. These methods can be employed to analyze model deficiencies in representing physical processes and for estimating uncertainties or unknown parameters. Furthermore, these approaches can be applied to enrich and complement incomplete experimental data with the goal of improving the state vector to better understand physical processes and precursor events responsible for transient phenomena that evolve in time.

Data assimilation (DA) has been used extensively in the numerical weather prediction (NWP) community to integrate measurements from weather stations, ships, aircraft, satellites and other observations with their numerical models [4, 5]. Data assimilation is utilized to obtain the best initial conditions for weather forecasting [6] and for reanalysis [7, 8]. NWP utilizes a range of DA-techniques such as Nudging [9], 3D-Var and 4D-Var [6] and the ensemble Kalman filter (EnKF) method [10].

Whereas the goal of NWP is to produce accurate forecasts of future weather events, the goal of the current work is to investigate DA for use in a *posteriori* analysis of models and physical processes. This distinctly different goal is a consequence of the different time and length scales that govern flows of engineering relevance. While DA is a developed technique in the atmospheric sciences, it has found limited application to engineering problems. Within the fire and combustion community, DA has been applied for parameter and state estimation. Jahn et al. [11] utilized a cost function to assimilate temperature data to estimate parameters within simple zonal models with the goal of improving predictions of fire growth. Kalman filter algorithms have been employed for parameter estimation for idealized flames [12], to combine particle tracking velocimetry and DNS [13], and for coupling high-speed experimental data and LES to investigate ignition and extinction in a turbulent jet flame [14]. In addition, Edwards et al. [15] has investigated continuous ensemble Kalman filtering for a reacting 2D hydrogen-air shear layer, showing that this approach can be applied to constrain LES so that its statistics evolve more in accordance with a given data set.

Within the wider fluid research community, Kalman filtering has been used to assimilate data into 2D simulations of flow around cylinders, turbulent three-dimensional simulations of a spatially evolving mixing layer and the flow around plates [16], and to estimate the probability distribution of inflow boundary conditions for urban environments [17]. Other assimilation techniques such as Newtonian relaxation have also been applied to improve 2D Reynolds-averaged Navier-Stokes simulations of flows over an airfoil [18] and to couple numerical and experimental data in a data optimization feedback loop where initial and boundary conditions could be obtained using only interior flow field information [19].

In this paper, we employ DA to integrate high-speed, high-resolution experimental data obtained from a turbulent jet [20] into LES. Assimilation of the experimental data are performed using an EnKF-algorithm and the performance of the method is investigated to understand its impact on the state estimation. The overarching goal of the current study, is to better understand the behavior of the EnKF for problems that are relevant to the fluids

community. Our first objective is to evaluate the use of DA as a method for integrating experimental data into simulations for state estimation. This is accomplished by comparing the transient predictions and the instantaneous flow structures obtained from a baseline LES without DA to those obtained via the EnKF algorithm. The second objective is to identify the impact that data localization and other EnKF-model considerations have on the resulting predictions. Following this, we investigate how the DA-method is affected by changes in experimental uncertainty, assimilation frequency and sparsity of experimental data. Finally, improvements that can be obtained by increasing the number of ensembles utilized within the EnKF are examined. With this knowledge, the EnKF can be utilized to its fullest extent to improve the predictability of LES by revealing and correcting for errors from a wide range of sources (from initial and boundary conditions, closure models, model parameters and from those that arise from an incomplete understanding of the physical phenomena).

The remainder of this paper is organized as follows. In Section 2, we present the governing equations. An overview of different DA-techniques is provided in Section 3 followed by details on the EnKF method utilized in the current work. Details of the experimental and computational setup are given in Section 5. Results are discussed in Section 6. The paper ends with conclusions and recommendations about the use of DA for turbulent simulations.

## 2 Governing Equations

In the present study, the finite-volume LES solver CharLES<sup>x</sup> [21] is employed for simulating the turbulent jet and we solve the governing equations for the Favre-averaged compressible conservation equations of mass, momentum and energy taking the following form:

$$\tilde{D}_t \bar{\rho} = -\bar{\rho} \nabla \cdot \tilde{\mathbf{u}}, \tag{1a}$$

$$\bar{\rho} \tilde{D}_t \tilde{\mathbf{u}} = -\nabla \bar{p} + \nabla \cdot (\bar{\boldsymbol{\sigma}} + \boldsymbol{\sigma}_{sgs}), \tag{1b}$$

$$\bar{\rho} \tilde{D}_t \tilde{e} = -\nabla \cdot (\bar{\mathbf{q}} + \mathbf{q}_{sgs}) + \nabla \cdot [(\bar{\boldsymbol{\sigma}} - \bar{p}\mathbf{I}) \cdot \tilde{\mathbf{u}}], \tag{1c}$$

where  $\tilde{D}_t = \partial_t + \tilde{\mathbf{u}} \cdot \nabla$  denotes the substantial derivative,  $\rho$  is the density,  $\mathbf{u} = (u, v, w)^T$  is the velocity vector with corresponding velocity components along the axial and spanwise directions  $\mathbf{x} = (x, y, z)^T$ ,  $p$  is the pressure,  $e$  is the specific total energy,  $\boldsymbol{\sigma}$  is the viscous stress tensor,  $\mathbf{q}$  is the heat-flux vector and the subscript “*sgs*” denotes turbulent subgrid quantities that are modeled. Closure for turbulent subgrid stresses are obtained using the Vreman SGS models [22]. Pressure is obtained by solving the ideal-gas law,

$$\bar{p} = \bar{\rho} R \tilde{T}, \tag{2}$$

where  $T$  is the temperature, which is evaluated from the internal energy [21] and  $R$  is the gas constant. For the system considered in Eq. 1, the state vector, consisting of density, velocity and energy, is given by

$$\boldsymbol{\phi} = [\bar{\rho}, \tilde{\mathbf{u}}, \tilde{e}]^T, \tag{3}$$

and Eq. 1 can be finally written as

$$\boldsymbol{\phi}(t + \delta_t) = \mathcal{M}(\boldsymbol{\phi}(t)), \tag{4}$$

where  $\mathcal{M}$  is the model represented by Eq. 1, which evolves the solution from time,  $t$ , to  $t + \delta_t$ , where  $\delta_t$  is the timestep of the numerical model.

### 3 Data Assimilation

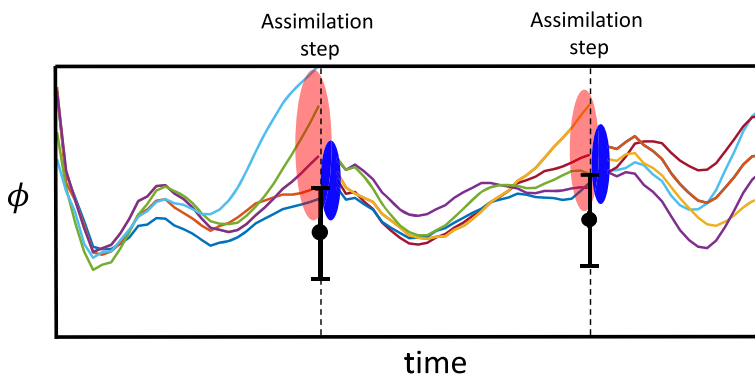
Data assimilation has the potential to be utilized as a tool for state estimation, model evaluation, parameter estimation, or to combine experimental data and numerical simulations to interrogate physical processes that cover transient phenomena. The purpose of data assimilation is to modify the state vector obtained from the numerical model based on specific observations as illustrated in Fig. 1. For the reader not familiar with data assimilation, in the following section we provide a brief review of variational (Newtonian relaxation, 3D-Var and 4D-Var) and statistical (Kalman Filter methods) DA-techniques that have been applied to assimilate experimental observations into simulations.

#### 3.1 Newtonian relaxation

Newtonian relaxation, also known as Nudging, has been applied to assimilate observations within the NWP community. The core idea of Nudging is to relax the prediction towards a given observation. This is accomplished by introducing a source-term into the model equations, Eq. 4, which takes the following form for linear observation operators [3]

$$F = -\frac{1}{\tau} (\mathcal{H}(\phi) - \psi), \quad (5)$$

where the term in brackets is the innovation, which represents the difference between the observation and state vector;  $\psi$  is the observation vector obtained from experiments or other numerical data,  $\mathcal{H}$  is the observation operator which maps the state vector into the vector of measurements and  $\tau$  is the relaxation time or nudging coefficient. This relaxation time controls how strongly the solution must adhere to the observation. For large values ( $\tau \rightarrow \infty$ ), the solution is dominated by the physics represented within the numerical model. Conversely, for  $\tau \rightarrow 0$ , the solution is strongly forced towards the observation and the physics represented by the numerical model has a weak influence on the solution. This method is also known as indiscriminate Nudging [9]. In this method the state vector is directly modified within the numerical model, which has the advantage of minimizing the overhead required to assimilate observations and is similar to the forcing method in DNS [23]. However, this comes with several disadvantages. First, a method for choosing the value of



**Fig. 1** Illustration of the EnKF data assimilation technique. Two data points are assimilated (circles). Colored lines represent state vectors and orange zones highlight their variation prior to the assimilation of data. Blue zones show the updated state vectors, which are pulled towards the experimental data points, improving the state estimation

$\tau$  dynamically without *a priori* knowledge of the timescales governing the physical processes under consideration is required. Second, the state vector is only modified at locations where observations are available. Thus, for sparse data the overall impact on the predicted state vector may be minor. Finally, this method does not directly account for uncertainties present in both the experimental observation and the numerical model. For these reasons this method is not considered in the current study.

### 3.2 Three- and four-dimensional variational data assimilation

More advanced assimilation techniques, such as three-dimensional variational (3D-Var) and four-dimensional variational (4D-Var) methods, overcome many of the limitations present in Newtonian relaxation, but come at the price of increased cost and complexity. In these methods both uncertainties present in the numerical model and experimental observations are directly considered when determining how to update the state vector obtained from the numerical model. This is accomplished by the prior-error covariance matrix  $\mathbf{C}_{\epsilon\epsilon}$  and observation-error covariance matrix  $\mathbf{C}_{\phi\phi}$ , which account for model errors and experimental uncertainties, respectively. Further, by considering the prior-error and observation-error covariance matrices, these methods can also update the state vector at locations where no observations are present through the prior error-covariance matrix, which relates model errors at locations where observations are available to model errors at locations where no observations are present. 4D-Var extends on the principles of 3D-Var by extending the assimilation method to consider the temporal evolution of the model. Within 4D-Var with Gaussian priors, the updated state vector is obtained by minimizing the following cost function [24]

$$\begin{aligned} \mathcal{J}(\boldsymbol{\Phi}^a) = & \frac{1}{2}(\boldsymbol{\Phi}^a - \boldsymbol{\Phi}^p)^\top \mathbf{C}_{\phi\phi}^{-1}(\boldsymbol{\Phi}^a - \boldsymbol{\Phi}^p) \\ & + \frac{1}{2} \sum_{t=0}^{\tau} [\boldsymbol{\psi}_t - \mathcal{H}_t(\boldsymbol{\phi}_t^a)]^\top \mathbf{C}_{\epsilon\epsilon}^{-1} [\boldsymbol{\psi}_t - \mathcal{H}_t(\boldsymbol{\phi}_t^a)] \\ & + \frac{1}{2} \sum_{t=1}^{\tau} [\boldsymbol{\phi}_t^a - \mathcal{M}(\boldsymbol{\phi}_{t-1}^a)]^\top \mathbf{Q}_t^{-1} [\boldsymbol{\phi}_t^a - \mathcal{M}(\boldsymbol{\phi}_{t-1}^a)], \end{aligned} \tag{6}$$

where  $\boldsymbol{\Phi}^p = \{\boldsymbol{\phi}_1^p, \boldsymbol{\phi}_2^p, \dots, \boldsymbol{\phi}_\tau^p\}$  is the prior or first guess state vector at all times during the temporal assimilation window,  $\tau$  is the number of discrete observation times present in the assimilation window,  $\boldsymbol{\Phi}^a$  is the updated state vector at all times ( $\boldsymbol{\Phi}^a = \{\boldsymbol{\phi}_1^a, \boldsymbol{\phi}_2^a \dots \boldsymbol{\phi}_\tau^a\}$ ) and  $\boldsymbol{\psi}_t$  denotes the observations available at time  $t$ . In addition, the three error-covariance matrices for the prior error ( $\mathbf{C}_{\epsilon\epsilon}$ ), the observation error ( $\mathbf{C}_{\phi\phi}$ ) and the model error ( $\mathbf{Q}$ ) require closure through modeling or *a-priori* knowledge. Often in NWP a strongly constrained 4D-Var formulation is employed in which the model is assumed to be perfect ( $\mathbf{Q} = \mathbf{0}$ ).

For 3D-Var with Gaussian priors, the cost function is given by [25],

$$\mathcal{J}(\boldsymbol{\phi}^a) = \frac{1}{2}(\boldsymbol{\phi}^a - \boldsymbol{\phi}^p)^\top \mathbf{C}_{\phi\phi}^{-1}(\boldsymbol{\phi}^a - \boldsymbol{\phi}^p) + [\boldsymbol{\psi} - \mathcal{H}(\boldsymbol{\phi}^a)]^\top \mathbf{C}_{\epsilon\epsilon}^{-1} [\boldsymbol{\psi} - \mathcal{H}(\boldsymbol{\phi}^a)]. \tag{7}$$

Various approaches have been developed to deal with observations, which occur during the assimilation period. In traditional 3D-Var, all observations within the assimilation period are applied at the assimilation time and  $\boldsymbol{\psi} - \mathcal{H}(\boldsymbol{\phi}^a)$  is calculated based on predictions at the analysis time. In comparison, in 3D-FGAT (first-guess at appropriate time) interpolation is performed to have model predictions at the appropriate observation time when calculating  $\boldsymbol{\psi} - \mathcal{H}(\boldsymbol{\phi}^a)$  [26, 27].

### 3.3 Kalman filter

The Kalman Filter is the starting point for a class of statistical and sequential DA-techniques. These sequential techniques estimate the solution state by evolving the prior state through time by considering modeling errors in Eq. 4,

$$\boldsymbol{\phi}^p(t + \delta_t) = \mathcal{M}(\boldsymbol{\phi}^p(t)) + \boldsymbol{\eta}, \tag{8}$$

where  $\boldsymbol{\eta}$  is the error associated with the model. The state vector is updated based on a set of observations in the form of

$$\boldsymbol{\psi} = \boldsymbol{\Psi} + \boldsymbol{\epsilon}, \tag{9}$$

where  $\boldsymbol{\Psi}$  is the true value of the observations in the absence of any errors and  $\boldsymbol{\epsilon}$  is the error associated with the measurements. The values of the errors in Eqs. 8 and 9 are unknown and statistical hypotheses are required before a solution can be obtained. Specifically, within the Kalman filter it is assumed that the error associated with both the prior prediction, obtained via the model  $\mathcal{M}$  and the observation is unbiased. Model and observation errors are uncorrelated and the error covariance matrix of the prior and observation are given by  $\mathbf{C}_{\phi\phi}$  and  $\mathbf{C}_{\epsilon\epsilon}$ , respectively [28].

The goal now is to find the solution  $\boldsymbol{\phi}^a$  that best satisfies Eqs. 8 and 9. Under the assumption of Gaussian errors, this solution can be written in a Bayesian framework as

$$P(\boldsymbol{\phi}^a | \boldsymbol{\psi}) = \frac{P(\boldsymbol{\psi} | \boldsymbol{\phi}^a) P(\boldsymbol{\phi}^a)}{P(\boldsymbol{\psi})}, \tag{10}$$

where  $P(\boldsymbol{\phi}^a | \boldsymbol{\psi})$  is the posterior density of  $\boldsymbol{\phi}^a$  given the measurements  $\boldsymbol{\psi}$ ,  $P(\boldsymbol{\phi}^a)$  is the prior density of  $\boldsymbol{\phi}^a$ ,  $P(\boldsymbol{\psi} | \boldsymbol{\phi}^a)$  is the likelihood function for measurements  $\boldsymbol{\psi}$  and  $P(\boldsymbol{\psi})$  is the evidence. The prior density is assumed to be given as

$$P(\boldsymbol{\phi}^a) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\phi}^a - \boldsymbol{\phi}^p) \mathbf{C}_{\phi\phi}^{-1} (\boldsymbol{\phi}^a - \boldsymbol{\phi}^p)\right), \tag{11}$$

and the likelihood function is defined as

$$P(\boldsymbol{\psi} | \boldsymbol{\phi}^a) \propto \exp\left(-\frac{1}{2} [\boldsymbol{\psi} - \mathcal{H}(\boldsymbol{\phi}^p)] \mathbf{C}_{\epsilon\epsilon}^{-1} [\boldsymbol{\psi} - \mathcal{H}(\boldsymbol{\phi}^p)]\right). \tag{12}$$

With this, Eq. 10 can be rewritten as

$$P(\boldsymbol{\phi}^a | \boldsymbol{\psi}) \propto \exp\left(-\frac{1}{2} \mathcal{J}(\boldsymbol{\phi}^a)\right), \tag{13}$$

where  $\mathcal{J}$  is

$$\mathcal{J}(\boldsymbol{\phi}^a) = (\boldsymbol{\phi}^a - \boldsymbol{\phi}^p) \mathbf{C}_{\phi\phi}^{-1} (\boldsymbol{\phi}^a - \boldsymbol{\phi}^p) + [\mathcal{H}(\boldsymbol{\phi}^a) - \boldsymbol{\psi}] \mathbf{C}_{\epsilon\epsilon}^{-1} [\mathcal{H}(\boldsymbol{\phi}^a) - \boldsymbol{\psi}]. \tag{14}$$

The solution to Eq. 10 is obtained by minimizing Eq. 14 with respect to  $\boldsymbol{\phi}^a$  resulting in the following solution form [28]

$$\boldsymbol{\phi}^a = \boldsymbol{\phi}^p + \mathbf{K} [\boldsymbol{\psi} - \mathcal{H}(\boldsymbol{\phi}^p)]. \tag{15}$$

where  $\mathbf{K}$  is the Kalman gain matrix. The Kalman gain matrix can be expressed as [28]

$$\mathbf{K} = \mathbf{C}_{\phi\phi} \mathbf{H}^T (\mathbf{H} \mathbf{C}_{\phi\phi} \mathbf{H}^T + \mathbf{C}_{\epsilon\epsilon})^{-1}, \tag{16}$$

where  $\mathbf{H} = \partial \mathcal{H} / \partial \boldsymbol{\phi}$  is the Jacobian of the observation operator with respect to the state vector. The presented formulation of the Kalman filter is the starting point for a group of assimilation algorithms such as the extended Kalman filter [29] and the ensemble Kalman

filter. In the present work, we only consider the ensemble Kalman filter and its formulation is presented next.

### 4 Ensemble Kalman Filter

In the current study, the ensemble Kalman filter is chosen due to its capabilities for application to large-scale problems and its robust evaluation of the error covariance which outweigh the limitations of a linear observation operator and assumed Gaussian distribution for the priors. Within the EnKF, the prior error-covariance matrix,  $C_{\phi\phi}$ , is replaced with a sample prior error-covariance matrix,  $P$ , to eliminate the need for storing and evolving the prior error-covariance matrix, which can be computationally expensive. Within the EnKF approach a set of  $N$  independent samples are utilized to build the sample prior error-covariance matrix, which is calculated as

$$P = \frac{1}{N - 1} \sum_{k=1}^N (\phi_k - \langle \phi \rangle) (\phi_k - \langle \phi \rangle)^T, \tag{17}$$

where  $\langle \phi \rangle$  is the mean of all ensembles. In this work, we employ EnKF with a perturbed observation vector [28] to estimate the state vector of the  $k^{\text{th}}$  ensemble member.

The state vector is updated by combining the prior and the perturbed observation for each ensemble resulting in the following expression

$$\phi_k^a = \phi_k^p + K [\psi_k - \mathcal{H}(\phi_k^p)], \tag{18}$$

for  $k = 1, \dots, N$  where  $\psi_k$  is the perturbed observation vector. Following Evensen [30], the perturbed observation vector combines the experimental measurements and observation errors, which are sampled from a normal distribution with expectation  $\mathbf{0}$  and covariance matrix  $C_{\epsilon\epsilon}$ . Equation 18 is then solved for each ensemble independently. The benefit of this approach is that the updated ensembles contain the correct error statistics for the analysis and can be directly integrated forward in time.

The perturbed observation vector combines measurement and observation errors to produce a unique set of measurements for each ensemble. In the current work, the observation operator consists of the weight factors obtained from interpolating the solution on the computational mesh to the experimental data points obtained from Delaunay triangulation and barycentric interpolation.

#### 4.1 Localization of the EnKF

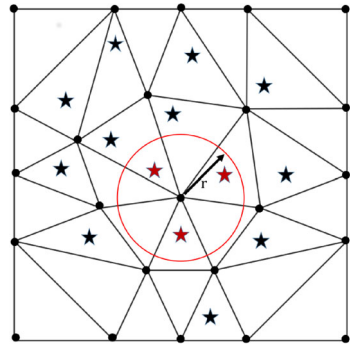
Within the EnKF algorithm, localization of the  $C_{\epsilon\epsilon}$  matrix is performed to reduce long-range spurious correlations due to sampling errors [31], which can lead to unphysical simulation results. Localization has the added benefit of decreasing the size of both the  $C_{\epsilon\epsilon}$  and  $P$  matrices. This limits the observations utilized within the EnKF to those within a given radius of the cell center. In the current work, a unique gain matrix,  $K_i$ , is constructed for the state vector at each grid point  $x_i$  and localization is obtained using a Schur product of the observation and observation covariance matrices. With this, the gain matrix is calculated as

$$K_i = PH^T (HPH^T + \Gamma \circ C_{\epsilon\epsilon})^{-1}, \tag{19}$$

where  $\Gamma \circ C_{\epsilon\epsilon}$  is the Schur product of  $\Gamma$  and  $C_{\epsilon\epsilon}$ , which can be computed using

$$(\Gamma \circ C_{\epsilon\epsilon})_{ij} = \Gamma_{ij} C_{\epsilon\epsilon,ij}. \tag{20}$$

**Fig. 2** Schematic of the localization process for a single cell center (black dots) with localization radius,  $r$  (red circle). Stars represent available experimental data. Only observations within the circle are utilized when constructing the EnKF for the cell center (red stars). The remaining observations (black stars) are not included in the EnKF



In the present work,  $\Gamma$  is defined as

$$\Gamma(\mathbf{x}, \mathbf{y}) = H(r - |\mathbf{y} - \mathbf{x}|), \tag{21}$$

where  $H$  is a Heaviside function,  $\mathbf{y}$  is the measurement location,  $\mathbf{x}$  is the cell center location and  $r$  is defined as the localization radius. This localization operator assumes homogeneity and more complex operators can be employed to account for the underlying flow topology [32]. Once  $\mathbf{C}_{\epsilon\epsilon}$  is localized,  $\mathbf{P}$  is constructed such that it contains information only from cells that have a non-zero entry in  $\mathcal{H}(\phi_k^p)$  for observations within  $\mathbf{C}_{\epsilon\epsilon}$ .

### 4.2 EnKF algorithm

In the current algorithm, the connectivity between the state vector and experimental measurement locations is pre-computed along with weights required for the observation operator. The connectivity and interpolation weights are calculated using the built-in `deLaunayTriangulation`, `ConnectivityList` and `pointLocation` functions from MATLAB. A schematic of the localization process is presented in Fig. 2 and the EnKF-procedure is summarized in Algorithm 1.

**Algorithm 1** Ensemble Kalman filter with localization.

- 1: Define ensemble size  $N$  and localization radius  $r$
- 2: Pre-compute connectivity and interpolation weights required for observation operator
- 3: Define assimilation frequency and calculate corresponding time interval  $\delta_t$
- 4: Initialize ensembles based on random uncorrelated simulation times
- 5: **while**  $t < t_{\text{end}}$  **do**
- 6:     Advance each ensemble  $k$  independently in time for  $t \rightarrow t + \delta_t$
- 7:     Calculate ensemble mean  $\langle \phi \rangle$ .
- 8:     Calculate residual for each ensemble  $(\phi_k - \langle \phi \rangle)$ .
- 9:     **for each** Control volume  $V_i$  **do**
- 10:         Calculate  $\Gamma \circ \mathbf{C}_{\epsilon\epsilon}$ , retaining only non-zero entries
- 11:         Calculate perturbed observation vector  $(\psi_k)$  for each ensemble
- 12:         Calculate innovation  $(\psi_k - \mathcal{H}(\phi_k^p))$  for each ensemble
- 13:         Calculate  $\mathbf{P}$  Eq. 17 for observations within localization radius  $r$
- 14:         Calculate gain matrix  $\mathbf{K}_i$  Eq. 19 and compute  $\phi^* = \mathbf{K}_i [\psi_k - \mathcal{H}(\phi_k^p)]$
- 15:         Compute new state vector Eq. 18 for each ensemble
- 16:     **end for**
- 17: **end while**



In the current work, each ensemble is advanced sequentially with the simulations themselves utilizing the parallel LES solver, CharLES<sup>x</sup>. Within the DA-algorithm, the code has been parallelized and the calculation of the EnKF for each control volume is performed in parallel. In terms of CPU time requirements, the main bottleneck of the algorithm is the sequential advancement of the ensemble calculation. Thus, the total CPU time scales linearly with respect to the number of ensembles. In comparison, for the parametric studies presented below, almost no change in the total CPU time is observed when changing the localization or assimilation frequency.

## 5 Experimental Configuration and Computational Setup

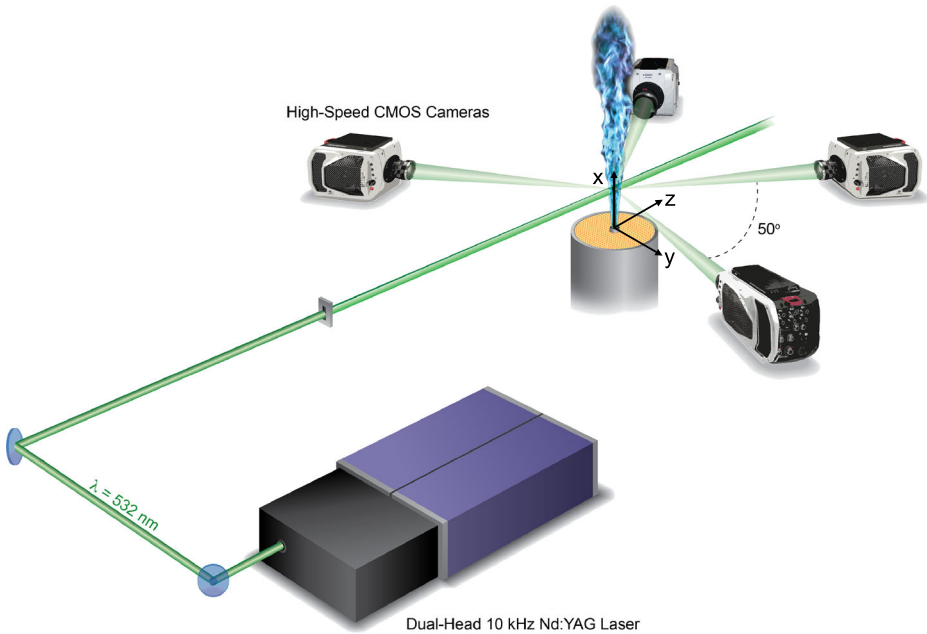
### 5.1 Experimental setup and measurement methods

A turbulent inert jet that was experimentally investigated by Coriton and Frank [20] is selected for the current study. This experimental data set was chosen as it provides both high spatial and temporal resolution for the three-component velocity field in the shear layer of the jet. The inert jet is based on the Sandia flame series [33] with the nozzle diameter enlarged from  $D = 7.2$  mm to  $D = 7.45$  mm. The bulk velocity of the air jet is  $U_b = 27.5$  m/s, resulting in a jet Reynolds number of 13,500. The annular pilot had no flow and the entire burner was surrounded by an air coflow with an exit velocity of 0.9 m/s.

Measurements of the three velocity components in the downstream region of the jet were obtained using Tomographic Particle Image Velocimetry (TPIV). The system consists of a high speed diode-pumped dual-head Nd:YAG laser and a set of four high speed CMOS cameras. A schematic of the experimental configuration is given in Fig. 3. Measurements were performed at a repetition rate of 10 kHz using a laser energy of 5 mJ/pulse and a pulse delay between laser heads of 10  $\mu$ s. A vertical slit was positioned in the beam path to select the most uniform portion of the beam, which was determined using a beam profiling camera. A pair of cameras was positioned on each side of the laser beam with two camera lenses positioned at 20° with respect to the  $y$ -axis and the other two camera lenses at 30°. The cameras were operated at 20 kHz in a frame straddling mode using a  $896 \times 800$  px<sup>2</sup> region of the detector. The jet and coflow were both seeded with 0.3  $\mu$ m aluminum oxide particles. Light scattering from the alumina particles was imaged onto each camera using identical camera lenses and Scheimpflug mounts to compensate for the displacement of the imaging plane. Computation of the velocity vectors consisted of the following steps:

1. Applying smoothing ( $3 \times 3$  px<sup>2</sup> Gaussian filter) and correcting the raw particle images for non-uniformity of particle scattering signals.
2. Reconstructing the particle volume distribution using a Multiplicative Algebraic Reconstruction Tomography (MART) algorithm [34].
3. Calculating the velocity vectors by iterative volume cross-correlation for a final volume interrogation size of  $24 \times 24 \times 24$  voxels<sup>3</sup> ( $393 \times 393 \times 393 \mu\text{m}^3$ ) with 75% overlap.
4. Removing/replacing spurious vectors identified via an outlier detection method and reducing the measurement noise with a spatial filter based on a penalized least-square method [20, 35].

The probe volume size was  $12.3 \times 2.5 \times 16.5$  mm<sup>3</sup> (in axial, spanwise and transversal direction) and contained  $126 \times 26 \times 169$  vectors with 98.4  $\mu$ m spacing representing an over-sampled representation of the measurement volume. A detailed description of the TPIV system including velocity uncertainty estimates is available in Refs. [20, 35].



**Fig. 3** Experimental configuration for high-repetition rate TPIV measurements

In the present work, a 50 ms time sequence of the TPIV measurements was selected at a location centered in the jet shear layer at  $z/D = 1.15$  and an axial height of  $x/D = 15$ . This measurement location was selected by requirements for resolving all turbulent scales [20, 35]. In the current study, approximately 280 million measurements are available to assess the value of DA of high-speed experimental measurements for LES.

## 5.2 Estimation of experimental errors

The observation covariance-error matrix,  $C_{\epsilon\epsilon}$  is assumed to be known and can be calculated as

$$C_{\epsilon\epsilon} = f \left( C_{\epsilon\epsilon}^G, C_{\epsilon\epsilon}^M, C_{\epsilon\epsilon}^R, C_{\epsilon\epsilon}^H \right), \quad (22)$$

where  $C_{\epsilon\epsilon}^G$  is the gross error due to the incorrect collection of data (missing data, incorrect units, data which has been mislabeled (i.e. velocity vectors are attributed to the wrong directions, data entry errors, etc.)),  $C_{\epsilon\epsilon}^M$  is the error due to measurement noise,  $C_{\epsilon\epsilon}^R$  is the representative error which is caused by converting *in situ* observations to values of interest and  $C_{\epsilon\epsilon}^H$  is the observation operator error which is introduced by transforming the experimental observations to values of interest. In the present work, gross errors ( $C_{\epsilon\epsilon}^G$ ) are expected to be negligible as the measurements was obtained from carefully calibrated and controlled laboratory experiments. This have further verified by checking to ensure the data are physically feasible, consistent and has no missing vectors. Errors associated with measurement noise ( $C_{\epsilon\epsilon}^M$ ) and volume reconstruction errors ( $C_{\epsilon\epsilon}^R$ ) have been estimated based on comparable experimental measurements at approximately 5% [20]. Additional representative errors ( $C_{\epsilon\epsilon}^R$ ) occur due to the inherent spatial and temporal averaging of TPIV and the apparent transport of ghost particles. These errors are lumped together to obtain a representative error

of approximately 5%, which has been estimated in collaboration with experimentalists and the work of Wieneke [36].

An estimation of  $C_{\epsilon\epsilon}^H$  can be obtained from an analysis of the observation operator. In the present study, the observation operator is a function that estimates the predicted velocity at each measurement location via linear interpolation. This error will be a function of the grid spacing and the local velocity field, which will vary with time. However, the error contribution associated with this linear interpolation is assumed to be negligible compared to  $C_{\epsilon\epsilon}^M$  and  $C_{\epsilon\epsilon}^R$ . Before calculating the observation covariance-error matrix several additional statistical assumptions are required. First, the off-diagonal terms in  $C_{\epsilon\epsilon}^M$  and  $C_{\epsilon\epsilon}^R$  are assumed to be zero. Second,  $C_{\epsilon\epsilon}^M$  and  $C_{\epsilon\epsilon}^R$  are assumed to be uncorrelated and the total uncertainty associated with the observations is obtained via

$$C_{\epsilon\epsilon} = \sqrt{\|C_{\epsilon\epsilon}^M\|_2^2 + \|C_{\epsilon\epsilon}^R\|_2^2} \tag{23}$$

Based on this information, the overall uncertainty for the velocity components is estimated at 7% and  $C_{\epsilon\epsilon}$  is a diagonal matrix,  $C_{\epsilon\epsilon} = \epsilon \times \psi \times \delta(y_i - y_j)$ , containing the uncertainty associated with each measurement. Here,  $\delta(y_i - y_j)$  denotes the Dirac delta function.

### 5.3 Preprocessing of experimental data

The velocity fields obtained from the experimental data are preprocessed to ensure they are divergence-free. Utilizing a divergence-free velocity field eliminates the potential for additional pressure and density fluctuations to be introduced into the flow-field simulation during the assimilation step. The use of a divergence-free velocity field is justified by considering the expected density fluctuations within the measurement window. The density fluctuations are estimated based on the axial velocity [37] as

$$\frac{\rho'}{\bar{\rho}} = (\gamma - 1)M^2 \left( \frac{u'}{\bar{u}} \right), \tag{24}$$

where  $\rho'$  is the density root-mean square (rms),  $\gamma$  is the specific heat ratio,  $M$  is the Mach number and  $u'$  is the axial velocity rms. Along the centerline at  $x/D = 15$ ,  $\rho = 1.21 \text{ kg/m}^3$ ,  $\gamma = 1.4$ ,  $\bar{u} = 14.5 \text{ m/s}$  and  $u' = 2.5 \text{ m/s}$ . With these properties,  $\rho'$  is estimated at  $10^{-4}\bar{\rho}$ . Thus, for the purpose of the current study, the air jet is considered to be incompressible and the density is considered to remain constant.

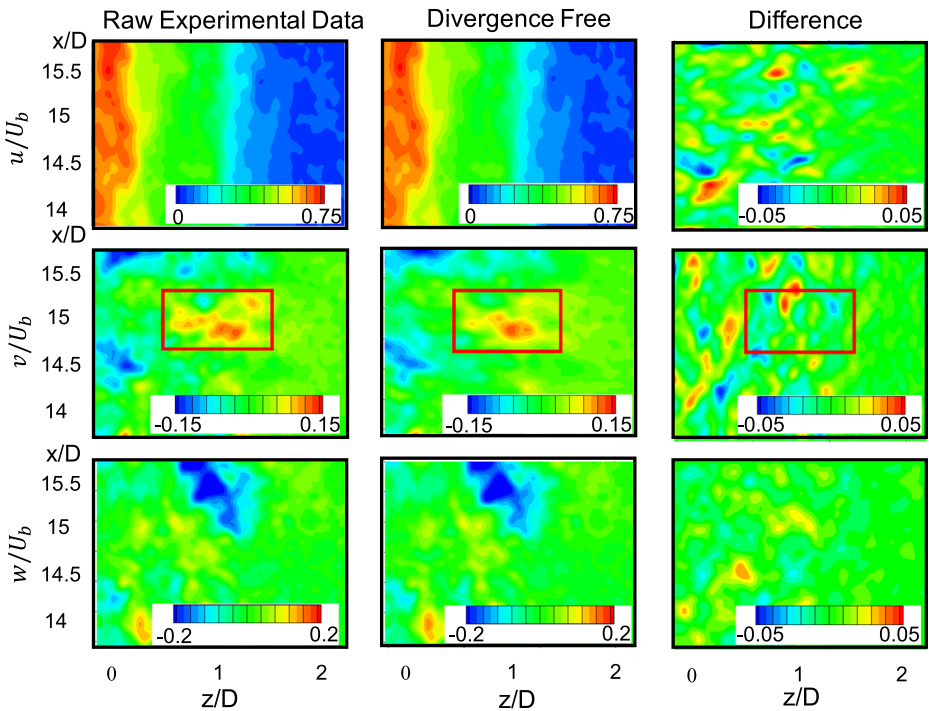
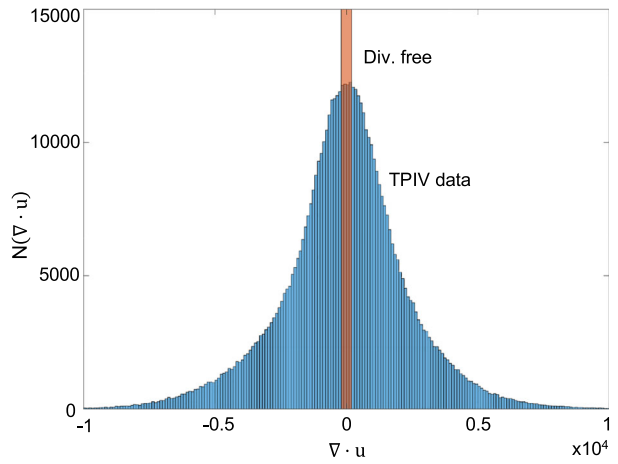
The divergence-corrective scheme of de Silva et al. [38] is applied to obtain the divergence-free velocity field from the experimental data. For each experimental time the divergence-free velocity field is obtained by solving the following minimization problem:

$$\min_{s.t. \nabla \cdot \mathbf{u}=0} \|\mathbf{u} - \boldsymbol{\psi}\|, \tag{25}$$

where  $\mathbf{u}$  is the resulting divergence-free velocity field and  $\boldsymbol{\psi}$  is the raw experimental data. The solution to Eq. 25 is obtained using the FMINCON function within the MATLAB optimization toolbox. A comparison of the apparent divergence of the velocity field obtained before and after the minimization process is presented in Fig. 4.

As can be seen in Fig. 4, the raw experimental data contains large divergences, also observed in [35], which would introduce perturbations into the simulations. In comparison, the divergence-free velocity field obtained via Eq. 25, reduces the divergence of the flow field by several orders of magnitude. Figure 5 shows the three-component velocity fields before and after the application of the divergence corrective scheme. The resulting divergence-free velocity field contains the same large-scale velocity structures, which appear along the top and centerline for the  $v$  and  $w$  velocity components along with the

**Fig. 4** Histogram of apparent divergence for the raw TPIV experimental data ( $\psi$ ) and computed divergence-free velocity field ( $\mathbf{u}$ )



**Fig. 5** Raw experimental velocity field from TPIV (left), divergence-free velocity field obtained from Eq. 25 (middle) and the difference between the two velocity fields (right)

positive velocity observed at the bottom and center of the experimental window. However, the small-scale flow structures observed for the divergence-free velocity field differ significantly compared to the raw experimental data, producing a smoother velocity field (red box). Although the overall flow structures observed are similar, the divergence-free velocity field is utilized for the data assimilation simulations as it eliminates a potential source of errors from modeling and measurements.

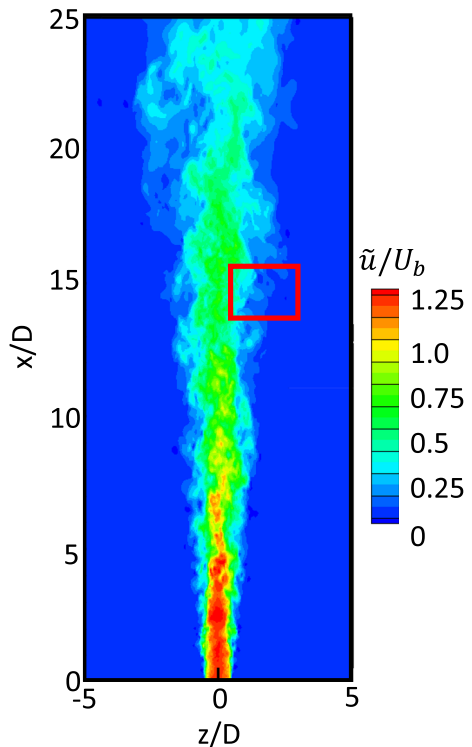
## 5.4 Computational details

A cylindrical computational domain is utilized with dimensions  $77D \times 12.8D \times 2\pi$ , in axial, radial and azimuthal directions, respectively. The governing equations are solved in a Cartesian coordinate system with  $x$  denoting the streamwise direction and  $y$  and  $z$  the corresponding orthogonal spanwise directions, respectively (see Fig. 3 for coordinate system).

An unstructured mesh with approximately 19 million control volumes is used with refinement in three areas: the near-nozzle region, the shear layer and the location of the experimental data with a minimum grid spacing of approximately 0.1 mm. An illustration of the instantaneous flow field and location of the measurement area is presented in Fig. 6.

Within the experimental measurement area the grid has an average resolution, defined based on the cube-root volume  $(\Delta x \Delta y \Delta z)^{1/3}$ , of approximately  $200 \mu\text{m}$ . Within the assimilation window the minimum grid spacing is approximately  $140 \mu\text{m}$  near the centerline and increases to  $350 \mu\text{m}$  near the outer edge of the assimilation window. Thus, each

**Fig. 6** Illustration of the instantaneous flow field and experimental measurement area (red box)



computational cell contains several experimental data points that can be used for data assimilation.

The jet inflow velocity profile is obtained from a DNS pipe flow calculation. Following the experimental setup, a laminar coflow of 0.9 m/s is prescribed. A characteristic pressure boundary condition is applied at the outlet and adiabatic non-slip boundary conditions are applied at the walls.

As a first step, the simulation is advanced in time, without any data assimilation, until a statistically stationary flow is observed. Initial conditions for each ensemble,  $\phi_k$ , are then selected from uncorrelated time instances from this simulation. This ensemble of solutions is then used as the starting point for all future simulations discussed in Section 6. In addition to the simulations containing data assimilation, additional reference cases (starting at  $\phi_{k=1\dots N}$  and without DA) are performed to assess the impact of the DA-method.

## 5.5 Baseline simulations

First, statistical results obtained from a baseline LES are analyzed to ensure that the LES is able to provide an adequate statistical representation of the flow. This is accomplished by comparing predictions for the mean and rms for the three velocity components against available experimental data. In the present study, the LES is evaluated based on the mean and rms centerline profiles and the radial profile at  $x/D = 15$ , corresponding to the center of the measurement window. Contour plots of the mean and rms for the axial velocity within the experimental measurement window are provided in Fig. 7. In the location where the experimental observations are available, strong velocity fluctuations of approximately 2 m/s are observed and occur over the majority of the measurement window. Further, the experiment and LES clearly show the presence of the shear layer where mixing between the jet and air is occurring. Comparing measurements and computed contour plots for mean axial velocity and axial velocity rms shows that the LES is able to capture the main features of the mean axial velocity.

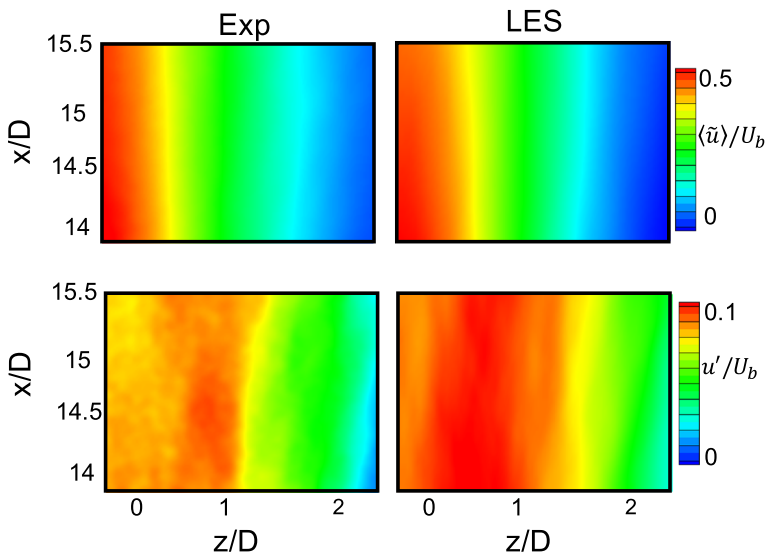


Fig. 7 Comparison of mean and rms axial velocity contours from baseline LES with experimental data

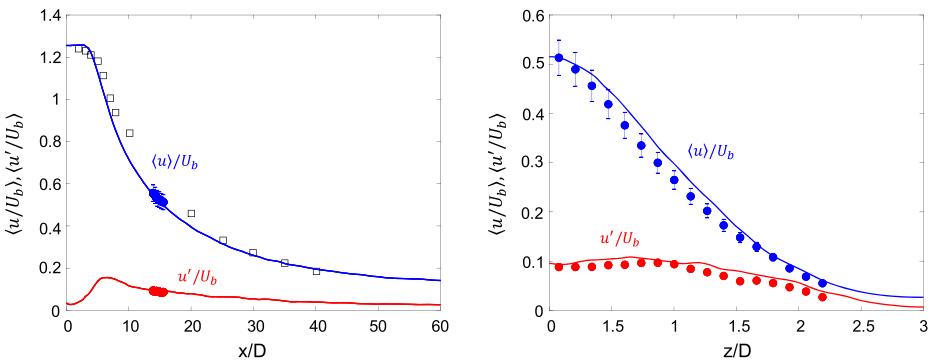
Statistics collected along the jet centerline are compared to experimental data in Fig. 8. The mean axial velocity obtained from the LES shows good agreement with experimental data and reported experimental measurements for a jet of Reynolds number 16,000 [39]. Within the measurement window, both the mean and rms axial velocity are in good agreement with the experimental data. Radial comparisons at  $x/D = 15$  are presented in Fig. 8 (right) showing good agreement between statistics and measurements. Due to the limited experimental data available to assess the accuracy of the downstream velocity and inlet conditions, the present velocity field is deemed adequate to test the impact of data assimilation.

### 5.6 Predictability of LES

As stated in Section 1, due to the chaotic nature of turbulence, the simulation will diverge from the true solution due to exponential growth of small perturbations that can be introduced by the numerical methods or physical models [2]. In the present study, the Lyapunov exponent is calculated to give an indication of the predictability of the LES within the DA-window following the method outlined in [2]. This metric provides a measure of the time horizon over which small perturbations will grow and impact the solution. This value can be compared with the temporal resolution of the experimental data to determine the frequency at which to perform data assimilation. To this end, two simulations, a reference and a perturbed case (with a relative initial perturbation of  $\zeta = 10^{-8}$ ), are advanced in time and the difference between the two simulations is calculated. The initial conditions of the perturbed simulation are obtained from

$$\phi^*(t_0) = \phi(t_0) + \zeta \|\phi(t_0)\| \cdot \hat{e}, \tag{26}$$

where  $\hat{e}$  is a vector which defines the variables to be perturbed,  $\|\cdot\| \equiv \left(\frac{1}{V} \int_V |\cdot| dV\right)$  is the  $L_1$ -norm and  $V$  is the volume of the domain. In the current study, only the axial velocity is perturbed when calculating the Lyapunov exponent,  $\lambda$ . When calculating the Lyapunov exponent the full LES domain is utilized, as the area of highest turbulence occurs near the nozzle. This domain is selected as it provides conservative estimates of the Lyapunov



**Fig. 8** Mean and rms centerline velocity profiles obtained from baseline LES compared to experimental data (circles) and measurements for a jet at Reynolds number 16,000 (squares) [39] (left) and radial velocity profiles from the baseline LES compared to experimental data at  $x/D = 15$  (right)

exponent and predictability time. The resulting separation,  $\|\delta\phi\| = \|\phi^*(t) - \phi(t)\|$ , is then used to determine the Lyapunov exponent following

$$\|\delta\phi(t)\| = \|\delta\phi(t_0)\|e^{\lambda t}, \quad (27)$$

where  $t_0$  is the initial time. The predictability time,  $t_p$ , is related to the Lyapunov exponent via [2]

$$t_p = \lambda^{-1}. \quad (28)$$

Based on the analysis of the global separation the Lyapunov exponent for this flow is approximately  $\lambda = 3500 \text{ s}^{-1}$  resulting in a predictability time of  $t_p = 0.29 \text{ ms}$  or  $1.07\tau_{conv}$ , where  $\tau_{conv}$  is the convective timescale, defined as  $\tau_{conv} = D/U_b$ . In this context, it is noted that the predictability time can be related to a corresponding predictability horizon,  $l_p \simeq t_p U_b$ , over which the flow-field remains correlated. This predictability horizon provides an estimate of the domain-of-influence that is affected by the upstream assimilation window for convection-dominated flows.

In the present study, DA in the form of EnKF is performed to assimilate experimental data into LES. As the predictability of the LES is greater than the measurement frequency (measurements obtained every 0.1 ms), we expect that this bounds the LES to a small realizable region around the experimental trajectory, allowing for a direct comparison with the transient experimental results in the observation window and allows for a numerical investigation of the transient phenomena present in the current flow.

## 6 Results

In this section, several aspects of the DA-algorithm are investigated to determine their impact on the resulting state-vector estimation. For the analysis, the experimental data were down-sampled by a factor of two to reduce the computational resources required during the EnKF, resulting in the assimilation of approximately 280,000 velocity vectors at each analysis step. First, the number of ensembles utilized is varied to determine the impact on the numerical predictions and its effect on potential error reduction when calculating the prior covariance matrix. Second, the degree of localization, in which experimental measurements are included in the calculation of the EnKF, is varied to determine if an optimal localization exists. In this section, the localization is parameterized as  $r/\ell$ , where  $\ell$  is the integral length-scale. Third, the experimental uncertainty is varied to understand its impact on the state-vector estimation. Fourth, the frequency of assimilation is reduced to evaluate the link between assimilation frequency, the predictability time and the solution accuracy. Finally, the number of observations available for assimilation is deliberately reduced and its impact on the accuracy of the predictions is assessed. A summary of cases considered for each section is provided in Table 1.

### 6.1 Number of ensembles

Within the EnKF algorithm, the accuracy of calculating the  $\mathbf{P}$  matrix plays a crucial role in determining the quality of the analysis. To increase the accuracy of the reconstructed  $\mathbf{P}$  matrix, additional stochastic samples (in the form of additional ensembles) are required. Ideally, we would increase the number of ensembles such that the reconstruction of  $\mathbf{P}$  is independent of the number of ensembles. This would also remove the need for localization



**Table 1** Summary of cases considered for each study

Study	$r/\ell$	$N$	Available data	Assimilation interval	$\epsilon$
Ensembles size (Section 6.1)	0.2	{6,12,18}	280,000	0.1 ms	7%
Localization (Section 6.2)	{0.05, 0.075, 0.1, 0.2}	6	280,000	0.1 ms	7%
Data uncertainty (Section 6.3)	0.075	6	280,000	0.1 ms	{0, 7%, $\infty$ }
Frequency (Section 6.4)	0.075	6	280,000	{0.1 ms, 0.3 ms}	7%
Available data (Section 6.5)	{0.05, 0.075, 0.1, 0.2,0.4,0.8}	6	{10,000; 70,000; 280,000}	0.1 ms	7%

as it is required to eliminate long range spurious correlations introduced by under-sampling caused by using a small number of ensembles. However, in practical problems, the number of ensembles is often constrained by available computational resources and run-time considerations. Thus, it is important to understand how increasing the number of ensembles impacts the reconstructed  $\mathbf{P}$  matrix and its impact on reducing the error. This is accomplished by considering the assimilation of 10,000 velocity vectors with a localization of  $r/\ell = 0.2$ , with the localization chosen as a fraction of the integral length scale defined as [40]

$$\ell \approx 0.226r_{1/2}^u, \tag{29}$$

where  $r_{1/2}^u$  is the jet half width at  $x/D = 15$ , resulting in an integral length scale of  $\ell = 1.6$  mm. Starting with a small number of ensembles ( $N = 6$ ), chosen heuristically to produce a baseline, the number of ensembles is increased from 6 to 12 and 18, while keeping the remaining properties constant to isolate the impact of increasing the number of ensemble members. In the current work, each ensemble is initialized from solutions of a separate simulation at sufficiently decorrelated time instances. Each ensemble member is advanced independently for a duration of 0.1 ms, after which time the experimental data are assimilated using the EnKF algorithm. The deviation of the quantity of interest ( $u$ ,  $v$  or  $w$ ) from the experimental data is calculated to assess the impact of EnKF. The normalized reduction in the mean error for a quantity  $\phi$  is defined as

$$E_\phi(t) = 1 - \frac{\|\boldsymbol{\psi}(t) - \mathcal{H}(\boldsymbol{\phi}^a(t))\|_2}{\|\boldsymbol{\psi}(t_0) - \mathcal{H}(\boldsymbol{\phi}^a(t_0))\|_2}, \tag{30}$$

where  $\|\cdot\|_2$  is the  $L_2$ -norm, and  $t_0$  is the time at the beginning of the assimilation sequence, prior to the assimilation of the experimental data. Thus, any reduction in the mean normalized errors is solely caused by increasing the number of ensemble members utilized to estimate the  $\mathbf{P}$  matrix. Table 2 presents the reduction in mean normalized errors for the three cases. When the number of ensembles is doubled from six to twelve, the reduction in the total mean normalized error increases from 64 to 85%, representing approximately a 58% decrease in the average error when compared to the six ensemble case. A further

**Table 2** Reduction in the mean normalized errors for different numbers of ensemble members after the assimilation of ten sets of experimental observations

$N$	$E_u$	$E_v$	$E_w$	$E_u$
6	0.38	0.76	0.82	0.64
12	0.78	0.89	0.89	0.85
18	0.80	0.90	0.90	0.86

increase to eighteen ensembles only marginally improves the state estimate but increases the computational cost by 50%. Thus, for this set of observations, experimental uncertainty and localization, it can be concluded that a maximum of twelve ensembles is required to adequately reconstruct the prior-error covariance. As both six and twelve ensembles provide a significant reduction in the mean normalized errors, computational time requirements should be considered when determining how many ensemble members to use. This can be accomplished by determining if the improvements in prediction quality are worth the increase in computational resources. For the current case, increasing the number of ensemble members from six to twelve requires twice the computational resources. The optimal number of ensembles (between six and twelve) can be determined based on available computational resources. During this work it was also observed that the impact of assimilating the experimental data was not limited to the probe volume, with a noticeable impact on the flow field downstream. However, due to the lack of experimental data downstream and the limited time elapsed for the simulation, it was not possible to quantify the extent of the influence. It is believed that the full spatial impact could be determined by comparing statistics collected over the same time period as those presented in Section 5.5, which was computationally not feasible in the current study. Based on the observations presented above, six ensembles are utilized for the remainder of the paper to allow for a range of parametric studies.

## 6.2 Localization

### 6.2.1 Need for localization

An important question to consider when assimilating experimental data into simulations is, which observations to include for each computational cell when calculating the gain matrix. Ideally, we would like to be able to utilize the EnKF without any localization. However, in practical terms, localization is often required for several reasons. First, as the EnKF is a Monte Carlo method, the  $\mathbf{P}$  matrix is approximated by statistically sampling the ensembles with the convergence proportional to  $N^{-1/2}$ . When utilizing a finite ensemble size, sampling errors are introduced which appear as spurious correlations over long spatial distances or between variables that are known to be uncorrelated [41]. The presence of spurious correlations can also lead to filter divergence where the EnKF analysis produces results that diverge from the true state [42]. One method for reducing spurious correlations is increasing the ensemble size [41, 42]. Thus, the number of ensembles required to accurately reproduce the correct covariance matrix increases with larger  $\mathbf{P}$  matrices (which spans long spatial distances or contains variables that are uncorrelated or weakly correlated) thereby, increasing the computational resources required. As the number of ensembles is often determined based on available resources or run-time requirements, a finite number of ensembles, often on the order of tens to hundreds, can be used to reconstruct the  $\mathbf{P}$  matrix and a different method to reduce spurious correlations is required. In the present work, the chosen method

is the localization of the experimental data, which reduces the impact of long-range spurious correlations, improving the rank of  $\mathbf{P}$  and increasing the local degrees of freedom for the analysis [41, 42].

Although localization is primarily utilized to enable EnKF to run with a small finite ensemble size, it has a secondary benefit of further reducing the computational cost associated with the assimilation process. Without any degree of localization, the size of  $\mathbf{P}$ ,  $\mathbf{HPH}^T$  and  $\mathbf{C}_{\epsilon\epsilon}$  can be prohibitively large. Specifically, the size of each of these matrices scale as

$$\mathbf{P} \in \mathbb{R}^{N_c N_s \times N_c N_s}, \mathbf{H} \in \mathbb{R}^{N_o \times N_c N_s}, \mathbf{HPH}^T \in \mathbb{R}^{N_o \times N_o}, \mathbf{C}_{\epsilon\epsilon} \in \mathbb{R}^{N_o \times N_o}, \quad (31)$$

where  $N_c$  is the number of grid points,  $N_s$  is the number of variables in the state vector and  $N_o$  the number of observations. For example, in the current study based on the assimilation of the three velocity components and computational mesh,  $N_c$  is of  $\mathcal{O}(10^7)$ ,  $N_s$  is of  $\mathcal{O}(10)$  and  $N_o$  is of  $\mathcal{O}(10^6)$ . Thus,  $\mathbf{P}$  would consist of a square matrix of size  $\mathcal{O}(10^{15})$  and require approximately 20 petabytes of storage if stored as a dense matrix. With the inclusion of localization, the storage requirements are drastically reduced.

### 6.2.2 Degree of localization

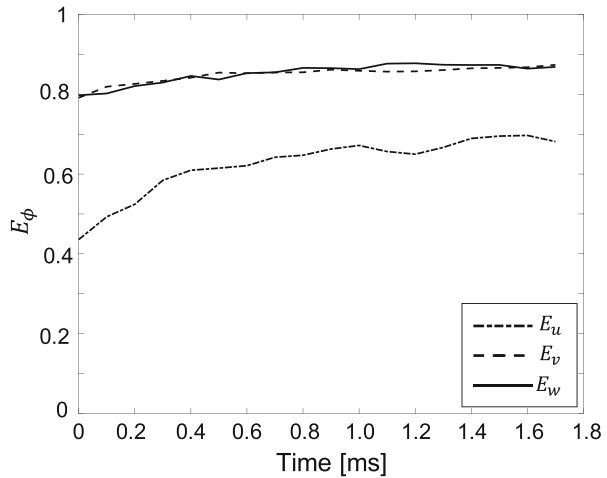
The degree of localization is a model parameter [41–43] and its impact on the overall quality of the state estimation is an important consideration in understanding how EnKF can be used for LES with high-resolution experimental data. Four sets of ensemble simulations with different localization values are considered to test the impact of localization on the accuracy of the EnKF predictions. The reduction in the mean normalized errors is presented in Table 3 for four values of the localization radius,  $r/\ell = \{0.05, 0.075, 0.1, 0.2\}$ .

For all four simulations a reduction of the error between 45% and 87% is observed through the use of the EnKF algorithm. Compared to the axial velocity, a larger reduction in error is observed for the other two velocity components. This characteristic is the result of two main factors. First, although the relative uncertainty in the experimental measurements for each velocity component is held constant, the absolute uncertainty is higher for the axial velocity due to its larger magnitude. Second, the relative error, as measured by the innovation, associated with the other two velocity components is larger than that of the axial velocity. Both of these characteristics result in a larger change in the update step for these two velocity components and a subsequent larger reduction in the normalized mean error. Also observable from Table 3 is the existence of a local minimum, which occurs for a localization of  $r/\ell = 0.075$ . This represents the optimal localization for the given flow configuration, number of ensembles and available experimental data. The temporal evolution of the mean normalized errors for  $r/\ell = 0.075$  is presented in Fig. 9 to demonstrate how EnKF reduces the normalized errors as the simulation progresses.

**Table 3** Ensemble mean reduction in the mean normalized errors for four localization radii

$r/\ell$	$E_u$	$E_v$	$E_w$	$E_u$
0.05	0.65	0.78	0.81	0.74
0.075	0.68	0.83	0.83	0.78
0.1	0.62	0.87	0.87	0.77
0.2	0.45	0.74	0.83	0.65

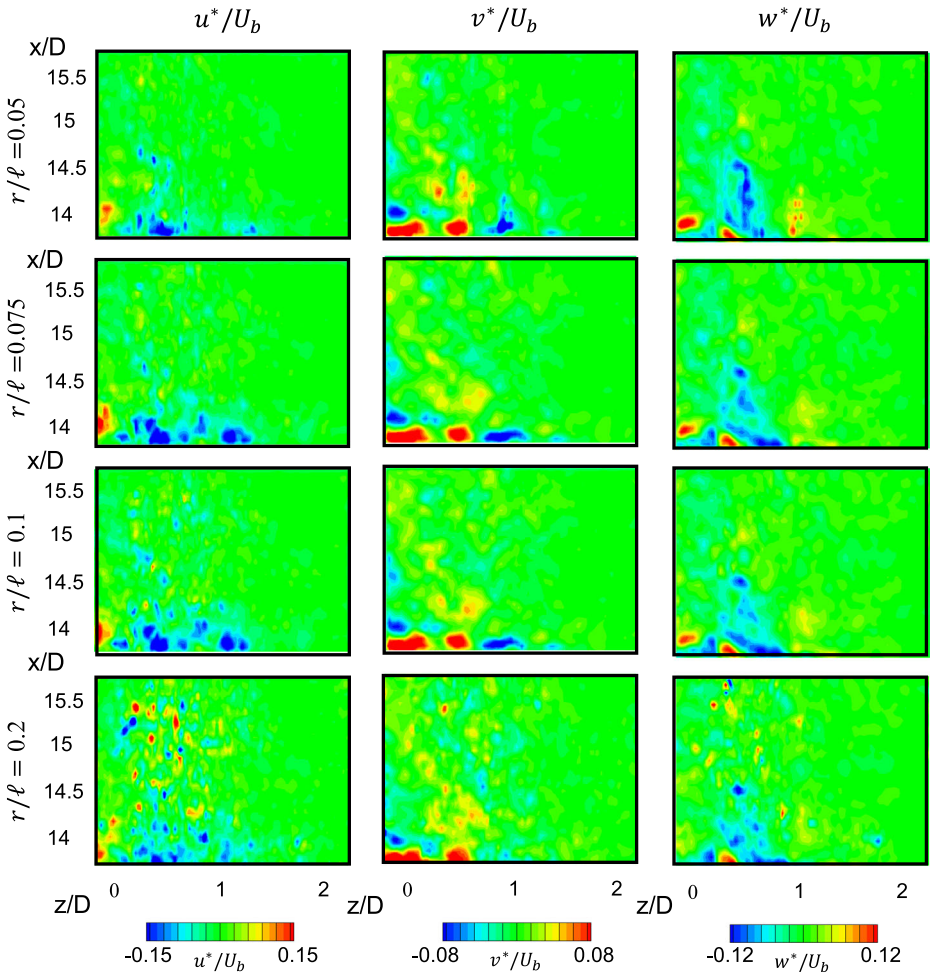
**Fig. 9** Temporal evolution of the ensemble mean reduction in the mean normalized errors with a localization of  $r/\ell = 0.075$  for each velocity component. Data are assimilated at 0.1 ms



As can be seen from Fig. 9, the largest error reduction occurs during the first assimilation step where a mean error reduction of approximately 40% and 80% is observed for the axial and transverse velocity components, respectively. After the initial DA-step, a further reduction for the axial velocity is observed over the next seven assimilation steps. The reduction in the mean axial error plateaus. In comparison, the transverse velocity components reach a steady state behavior after the first assimilation step, further demonstrating how EnKF impacts each velocity component differently. Similar characteristics are observed for other localization values.

Next, we investigate how the localization radius impacts the gain matrix, the update to the velocity components and the resulting velocity field. By considering the local  $\mathbf{K}$  matrix for a single grid point we are able to study the factors that impact the determination of the optimal localization. Within the Kalman gain matrix, the condition number of  $\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{C}_{\epsilon\epsilon}$ , can give an indication of whether the local EnKF algorithm is ill-conditioned [43]. As the condition number increases, the solution to Eq. 19 becomes more susceptible to large numerical errors. An analysis indicates that the condition number increases by several orders of magnitude as the localization is increased from 5% to 10% of the integral length scale. Thus, based on the current conditions, we expect larger numerical errors to be present when less localization is applied. This is in agreement with the observations of Neger [43] who attributed certain filter divergence to the large condition number of the  $\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{C}_{\epsilon\epsilon}$  matrix. For very large localization (small values of  $r$ ), the number of computational cells that only have a few measurements within the localization radius increases, limiting the information that can be assimilated from the experimental data. This limited transfer of information from the measurements results in an incomplete picture of the local flow field and has a detrimental impact on the analysis. This can be observed by investigating contours of the update step and resulting velocity fields, presented in Figs. 10 and 11, respectively.

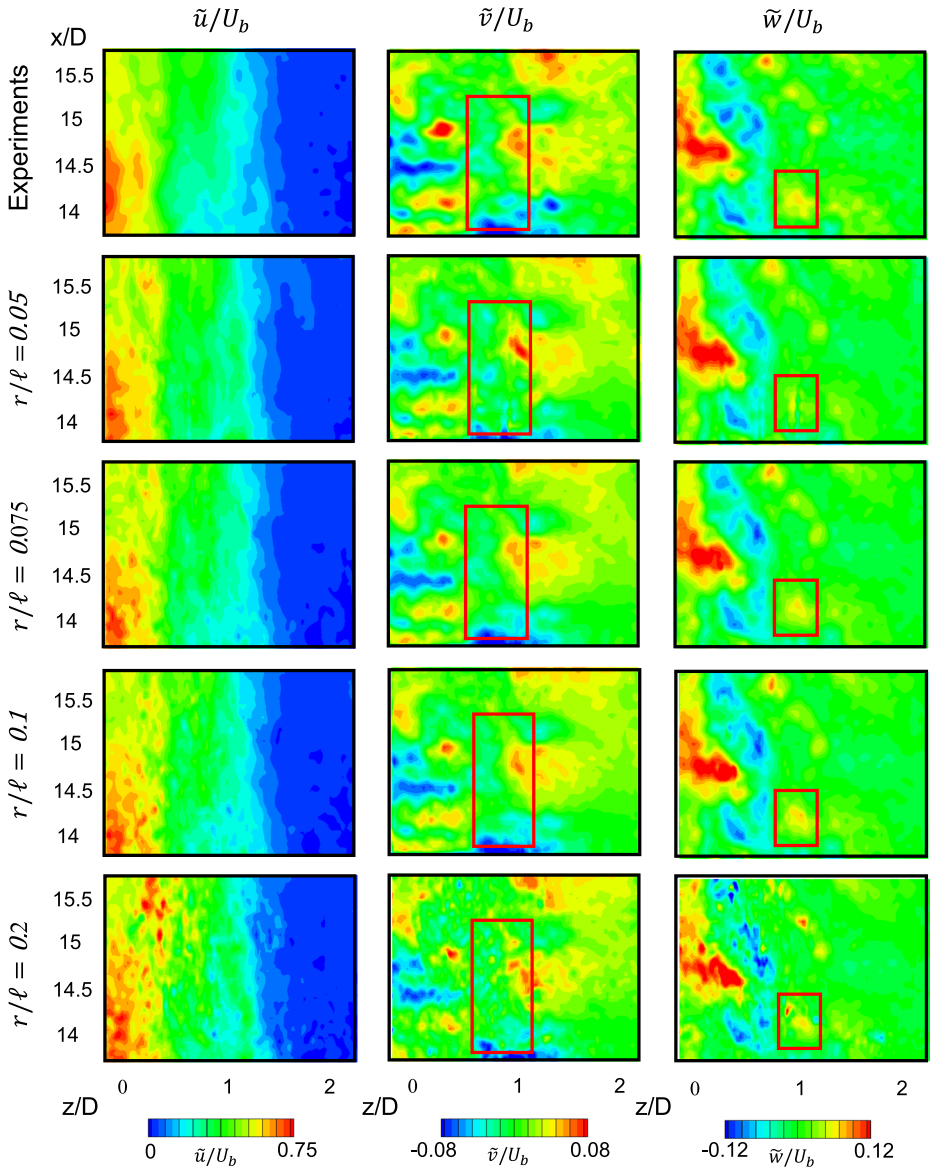
The impact of increased numerical errors occurring with larger values of  $r$  can be observed by comparing the update,  $\phi^* = \mathbf{K} [\psi_k - \mathcal{H}(\phi_k^p)]$ , that occurs during the assimilation step for each velocity component as the value of  $r$  is increased as shown in Fig. 10. Comparing the update step for  $r/\ell = 0.075$ , which corresponds to optimal localization and



**Fig. 10** Contours of the velocity update obtained from the EnKF algorithm for different localization parameters

$r/\ell = 0.2$ , which corresponds to a sub-optimal localization, several important observations can be made. First, major features of the update for the  $\tilde{v}$  and  $\tilde{w}$  velocity components at the bottom of the domain are very similar, indicating that the algorithm is behaving very similarly at this location for these two values of  $r$ . It is noted that farther downstream and for the entire domain for the axial velocity, the update step becomes increasingly noisy. For  $r/\ell = 0.2$ , the small pockets of positive or negative updates that occur very close to each other negatively impact the solution. In comparison, for  $r/\ell = 0.075$  a smoother update field is observed downstream with minimal noise introduced into the update.

With the largest localization applied, corresponding to the smallest localization radius ( $r/\ell = 0.05$ ), regions exist within the DA-window that only contain one or two



**Fig. 11** Velocity contours obtained from the EnKF algorithm for different localization parameters compared to the experimental data

measurements due to the sparsity of the experimental data within the localization radius of the cell center. This can clearly be seen in Fig. 11 where streaks form in the updated velocity field, due to incomplete information of the local flow field provided to the EnKF algorithm resulting in a poor update (red boxes). As the simulation progresses, errors from these cells convect with the flow field preventing the relative error from decreasing further.

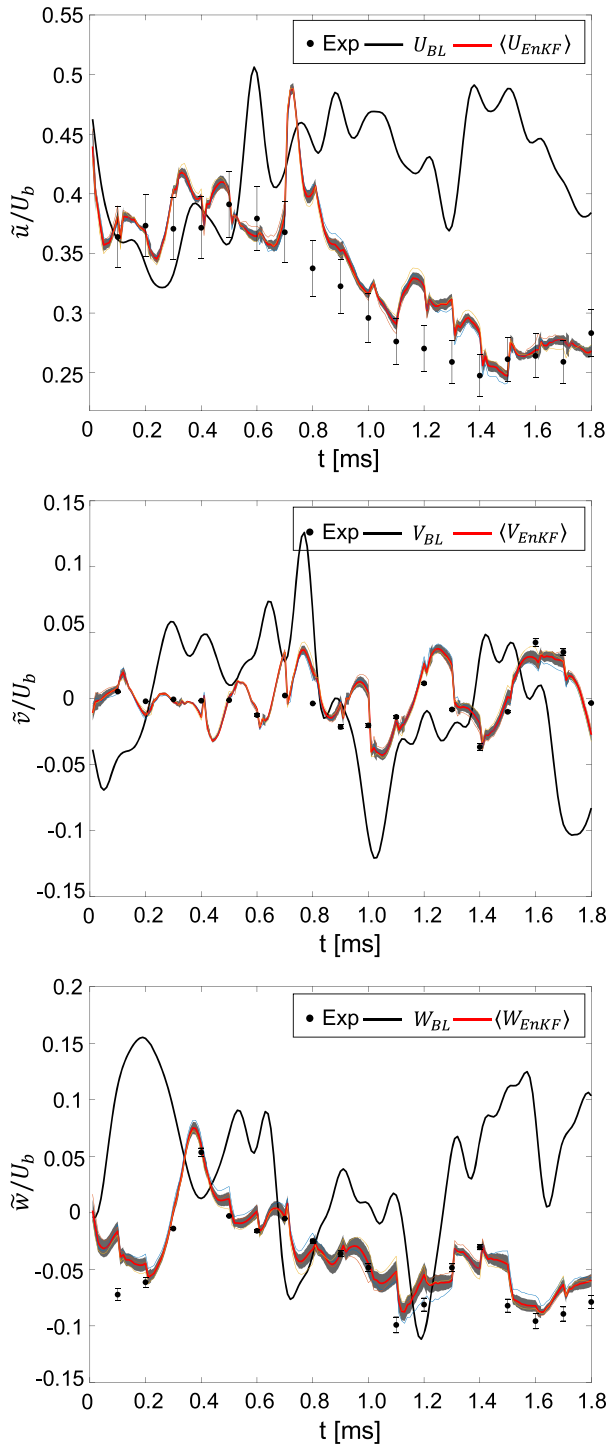
In comparison, the remaining three localization radii are larger than the experimental resolution and the update step calculated for each grid cell has a better representation of the local flow field. With a localization of  $r/\ell = 0.075$ , which is slightly larger than the experimental resolution, a noticeable improvement in the resulting velocity fields is observed, whereas with a small localization a smearing of the flow features is observed and a loss of the finer flow-field structures can be seen. Thus, the optimal choice of localization is a compromise between maximizing the number of observations available at each control volume to provide the best representation of the local flow field and reducing the condition number of  $\mathbf{HPH}^T + \mathbf{C}_{\epsilon\epsilon}$  to reduce the numerical errors introduced. As it has been demonstrated that the EnKF is able to better reproduce the transient behavior of the jet, next we investigate how the experimental uncertainty impacts the ability of EnKF to recover the correct transient behavior.

Transient predictions of the three velocity components are presented in Fig. 12 for the baseline LES and the EnKF with a localization of  $r/\ell = 0.075$  to demonstrate how the EnKF algorithm modifies the transient velocity profiles. The transient velocity profile obtained from the baseline LES contains significant deviations from the experimental data over 1.8 ms of simulation time. For the axial velocity, the baseline LES consistently predicts higher velocities at this location compared to the measurements and DA-predictions. Although the baseline LES is not expected to capture the exact transient velocity profiles observed experimentally as it is a stochastic representation of the flow field, the large deviations highlight the limitations of using LES to investigate experimentally observed transient phenomena. At the beginning of the assimilation sequence,  $t_0 = 0$ , a noticeable improvement in the transient velocity profile is observed. However, between 0.6 and 0.8 ms, the algorithm is unable to completely recover the measurement data due to the large deviation from the experimental mean and low spread in the ensembles, represented by the gray band. In comparison, the axial velocity fields obtained from the assimilation sequences better reflect the measurements after 0.9 ms. Also observable from Fig. 12 is that the relative uncertainty of DA, shown by the gray band, is not constant over the assimilation window, with higher uncertainty in the simulations predicted between 0.8 and 1.4 ms. Similar trends are also seen for the  $\tilde{v}$  and  $\tilde{w}$  velocity components as shown in Fig. 12.

The baseline LES predicts higher magnitudes for the  $\tilde{v}$  velocity components, with two periods of higher and lower velocities that are not observed experimentally. It is shown that with the assimilation of TPIV, DA is able to improve the predictions of the spatio-temporal evolution of velocity field compared to the baseline LES. For the  $\tilde{w}$  velocity component, a similar disparity is observed between the baseline LES and experiments. However, at all times DA is in better agreement with the measurements, demonstrating the ability of EnKF to locally correct the velocity field to better match the conditions observed experimentally. This demonstrates that DA can be used to evaluate existing models, by comparing the magnitude of the update step in the EnKF algorithm. This information can be used to determine areas that require model improvements. It should also be noted that DA is not always within the experimental uncertainty of the measurements, as EnKF considers both the experimental uncertainty and sample prior error-covariance, estimated through the  $\mathbf{P}$  matrix, when calculating the updated velocity.

Finally, the Reynolds stresses are calculated within the assimilation window and compared with the baseline LES and experimental data. The probability density function (PDF) of the Reynolds stresses is chosen for comparison to understand how the distribution of the Reynolds stresses changes with the assimilation of the experimental data and is shown

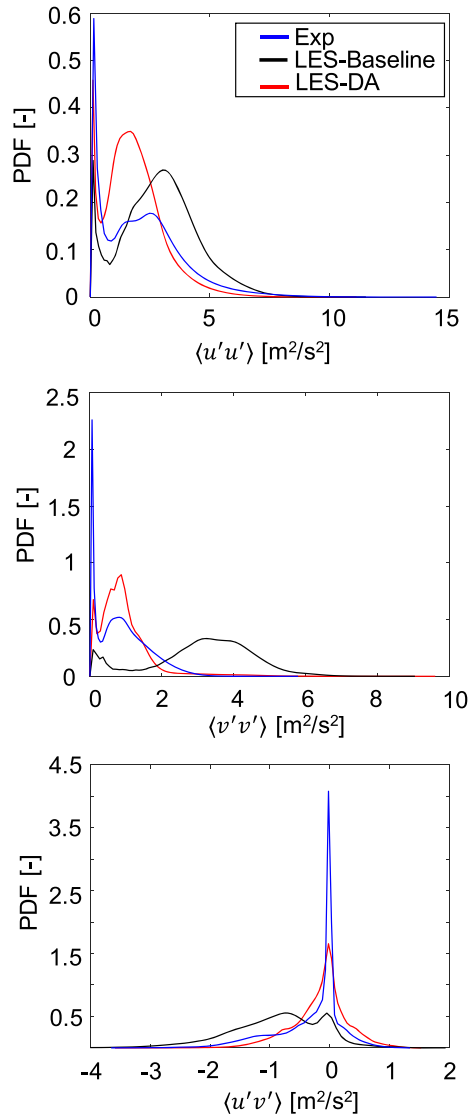
**Fig. 12** Temporal evolution of the three velocity components for the baseline simulation (thick black line) and the mean obtained from the ensemble members (red line) compared to the experimental measurements at  $(x/D = 14.6, y/D = 0.037, z/D = 0.70)$ . The shaded region in gray represents the rms of the ensemble members and three ensemble trajectories are shown by thin lines. Top  $\tilde{u}$ , middle  $\tilde{v}$ , bottom  $\tilde{w}$





in Fig. 13. For the  $\langle u'u' \rangle$  stress component, the LES-DA results better reproduce the peak observed at zero and the tail of the PDF. Compared to the baseline simulation, the LES-DA produces a narrower PDF, however both the baseline and LES-DA simulations fail to capture the location of the second peak in the PDF. For both the  $\langle v'v' \rangle$  and  $\langle u'v' \rangle$  stress components, the assimilation of the TPIV data significantly improves the shape of the calculated PDF. In both cases, the LES-DA simulation better matches the peaks observed in the experimental data and the tails of the PDF. These results further demonstrate that assimilation of the experimental data has a beneficial impact on the LES-DA simulation's ability to reproduce the experimentally observed conditions.

**Fig. 13** PDF of the Reynolds stresses for  $\langle u'u' \rangle$  (top),  $\langle v'v' \rangle$  (middle) and  $\langle u'v' \rangle$  (bottom), obtained from a baseline LES (black line) and LES-DA simulation (red line) compared to the experimental data (blue line)



### 6.3 Experimental uncertainty

When obtaining experimental data from different sources, different levels of uncertainties in the observations may be present. In this section, we demonstrate how EnKF is able to reject highly uncertain measurements. This is accomplished by considering three simulations (using six ensembles and a localization of  $r/\ell = 0.075$ ) with different observation uncertainties. The first simulation utilizes “perfect” experimental data where the uncertainty approaches zero. The second simulation consists of results from Section 6.2, containing uncertainties of approximately 7%, whereas for the final simulation, poor experimental data are demonstrated with  $\epsilon \rightarrow \infty$ . Following the analysis performed in Section 6.2, both the mean error reduction,  $E$ , for each velocity component as well as the impact on the transient predictions of the axial velocity are presented to assess the impact of the uncertainty on the recovered solution (Table 4).

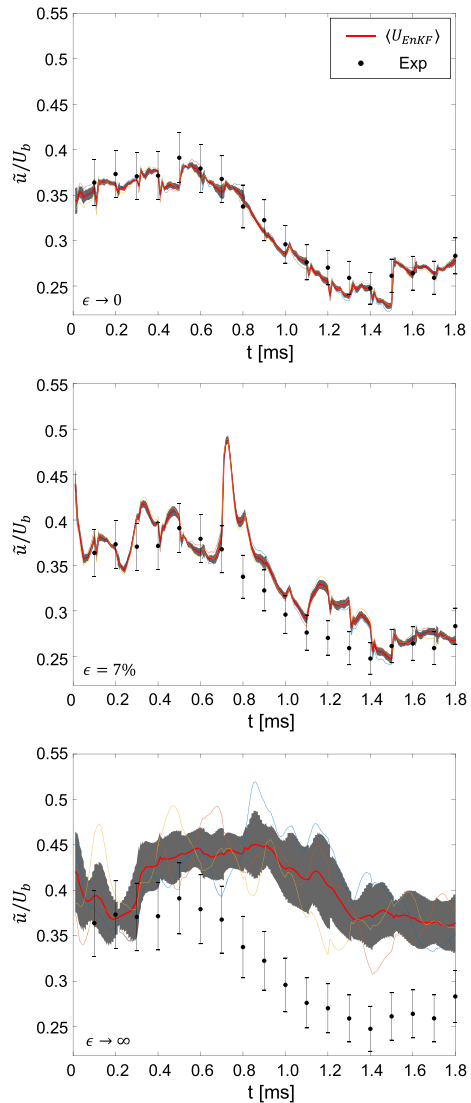
When “perfect” experimental data are assimilated a reduction in error of approximately 90% is observed for each of the velocity components. A further reduction of the error is not possible as each state vector is updated based on several experimental measurements, defined based on the localization radius. Thus, even with “perfect” experimental data, some error will propagate into the simulation through the update step and prevent the simulation from reaching zero error. By increasing the uncertainty in the experimental data, EnKF puts less weight on the experimental observation. When the uncertainty approaches infinity the resulting error reduces by 27% compared to the baseline, defined as the time step before the assimilation begins. However, as shown shortly this reduction is a by-product of utilizing the ensemble mean and the time varying nature of the experimental data and simulations. In comparison, the error reduction experienced with an uncertainty of 7% falls within these two extremes. Thus, the results presented in this section can be thought of as a comparison of the limiting behavior one can expect when applying the EnKF, demonstrating that this algorithm is able to reject bad experimental data as demonstrated in Fig. 14.

Next, the transient axial velocity predictions obtained for the three experimental uncertainties are compared in Fig. 14 to provide a quantitative representation of the predictions. For this comparison the baseline simulation is removed to improve clarity. When  $\epsilon \rightarrow \infty$  is specified the transient velocity predictions do not closely follow the experimental data and the variation within the ensemble members remains constant over the simulation. Further, by comparing the individual ensemble members, it can be seen that EnKF does not significantly alter the velocity trajectories and that each ensemble member oscillates around the ensemble mean, further demonstrating that EnKF has rejected the experimental measurements in the analysis. In comparison, when the uncertainty in the experimental data approaches zero the predictions are much closer to the experimental data and the variation within the ensembles is very small. At an uncertainty of 7%, the recovered solution has larger deviations from the measurements than for the case with  $\epsilon \rightarrow 0$  but correctly reproduces major transient flow features observed experimentally. These results show that EnKF can correctly account for different levels of experimental uncertainties through the gain matrix and is able to reject uncertain observations automatically.

**Table 4** Reduction in the mean normalized errors for three experimental uncertainties

Experimental uncertainty	$E_u$	$E_v$	$E_w$	$E_u$
$\epsilon \rightarrow 0$	0.87	0.87	0.87	0.87
$\epsilon = 7\%$	0.68	0.83	0.83	0.78
$\epsilon \rightarrow \infty$	0.14	0.41	0.31	0.27

**Fig. 14** Temporal evolution of axial velocity for the different levels of experimental uncertainty compared to the experimental measurements at  $(x/D = 14.6, y/D = 0.037, z/D = 0.70)$ ; Top: “perfect” experimental observation ( $\epsilon \rightarrow 0$ ); middle:  $\epsilon = 7\%$  experimental uncertainty; bottom: uncertain observations ( $\epsilon \rightarrow \infty$ ). The shaded region in gray represents the rms of the ensemble members and three ensemble trajectories are shown by thin lines. Error bars represent 7% uncertainty in measurement values



### 6.4 Assimilation frequency

From Section 5.6 it was found that the predictability time of the LES is approximately three times larger than the time interval between experimental measurements. This suggests that it should be possible to assimilate experimental data less frequently and still recover a similar accuracy within the simulation. This observation is tested by rerunning the optimal localization found in Section 6.2, using six ensembles, a localization of  $r/\ell = 0.075$  and only assimilating every third set of measurements. As a first step in analyzing the results, the reduction in the normalized error,  $E$ , for the axial velocity is compared for two normalized errors and two assimilation time intervals in Table 5.

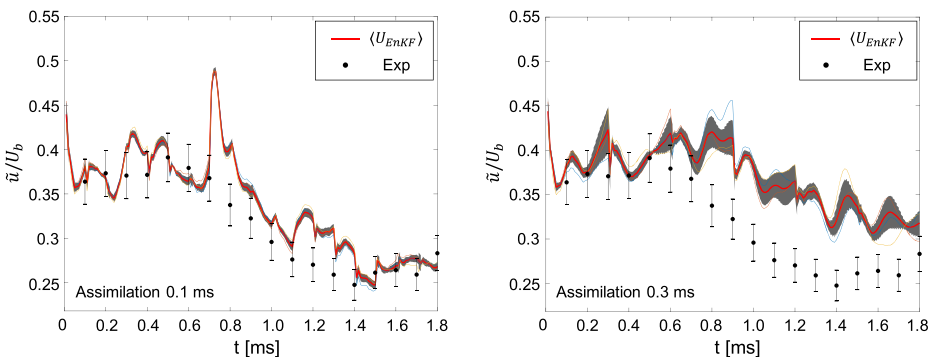
The first error is the normalized error reduction in the prior,  $E_u^p$ , at the end of the assimilation window and the second is the normalized error reduction of the updated solution,

**Table 5** Reduction in the mean normalized errors for different assimilation time intervals

Assimilation interval	$E_u^P$	$E_u^a$
0.1 ms	0.6	0.68
0.3 ms	0.5	0.65

$E_u^a$ . For the updated solution, the normalized error reduction achieved from the simulations where the data are assimilated every 0.1 ms and every 0.3 ms are almost identical throughout the simulation. This demonstrates that the overall accuracy of the analysis after the assimilation step is not adversely impacted by lowering the assimilation frequency. However, an important observation can be made when comparing the reduction in the forecast error. With an assimilation interval of 0.1 ms, the forecast error reduction reaches a steady-state value of approximately 60% after 1 ms of simulation time. In comparison, with the observations assimilated every 0.3 ms the forecast error reduction reaches approximately 50%. By reducing the assimilation frequency, the simulation has more time for errors to grow between assimilation steps. The growth of these errors causes each of the simulations within the ensemble to diverge further from the “true” solution. This can also be observed by comparing the transient axial velocity predictions obtained from these two sets of simulations as shown in Fig. 15. When the assimilation is performed every 0.3 ms, the variance between the ensembles grows continuously after the assimilation of the experimental observations. Following this growth, a large update is observed during the next assimilation step which moves the predictions closer to the experimental data. In comparison, when assimilation is done every 0.1 ms, smaller innovation steps are observed for the ensemble members and smaller update steps are obtained over the majority of the assimilation interval. However, between 0.7 and 0.8 ms a significant deviation of the axial velocity can be seen and a decrease in ensemble rms is observed. As the deviation occurs between assimilation steps this behavior is due to the LES model and local flow dynamics. During subsequent assimilation steps the axial velocity recovers towards the experimental data, demonstrating that the algorithm helps bound the solution.

Overall, these results demonstrate that more frequent assimilation of the experimental data has a positive impact on the analysis. This behavior may be amplified in the current



**Fig. 15** Temporal evolution of axial velocity for the different assimilation time intervals compared to the experimental measurements at  $(x/D = 14.6, y/D = 0.037, z/D = 0.70)$ . Left assimilation every 0.1 ms, Right assimilation every 0.3 ms. The shaded region in gray represents the rms of the ensemble members and three ensemble trajectories are shown by thin lines

study due to the small window in which experimental data are available and the small localization radius. In cases where experimental data are available over the entire computational domain, this behavior may be less severe as the entire flow field can be updated at every assimilation time, potentially reducing or eliminating some sources of error.

### 6.5 Data sparsity

In the previous sections, high-repetition rate experimental data was assimilated into the LES. However, the behavior of EnKF when sparse measurements are available is also of interest. In the present study, the divergence-free velocity field calculated in Section 5.3 is down-sampled to simulate the situation where sparse experimental data are available resulting in three experimental data sets with between 10,000 and 280,000 velocity vectors available at each assimilation step. For each observation resolution, a sequence of simulations with different amounts of localization is performed to determine the localization that maximizes error reduction. Six localization radii are utilized ranging from 5–80% of the integral length scale to ensure that the complete range of localization is considered. Several important characteristics can be observed from the relative reduction in error presented in Fig. 16. First, the localization radius that produces the largest reduction in error increases as the available observations decrease. This behavior is due to coarsening the experimental resolution, which requires a larger localization radius to ensure each control volume has sufficient observations to assimilate. Second, the reduction of error is largest with the highest number of observations and decreases when lower resolutions are utilized. This is attributed to the reduction of available experimental data, which results in less information on the local flow field being assimilated and the larger localization radius, which includes data points further away from the computational cell. Accurately reproducing the covariance for these points is difficult due to the limited number of ensembles and the weaker correlation between distant points. As errors in the covariance matrix directly propagate into the analysis, the overall reduction for larger localization radii is directly related to the accuracy of the constructed  $\mathbf{P}$

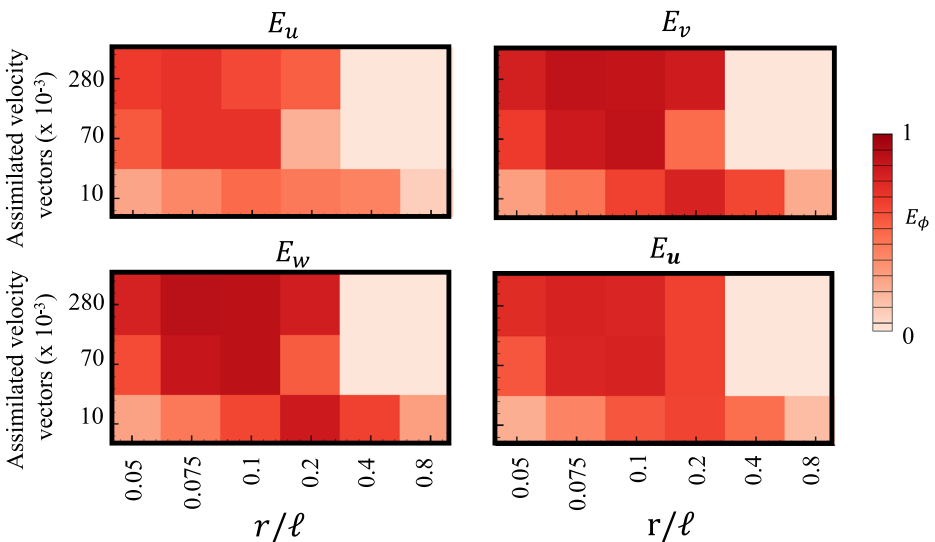


Fig. 16  $E_\phi$  for different experimental resolution and localization values

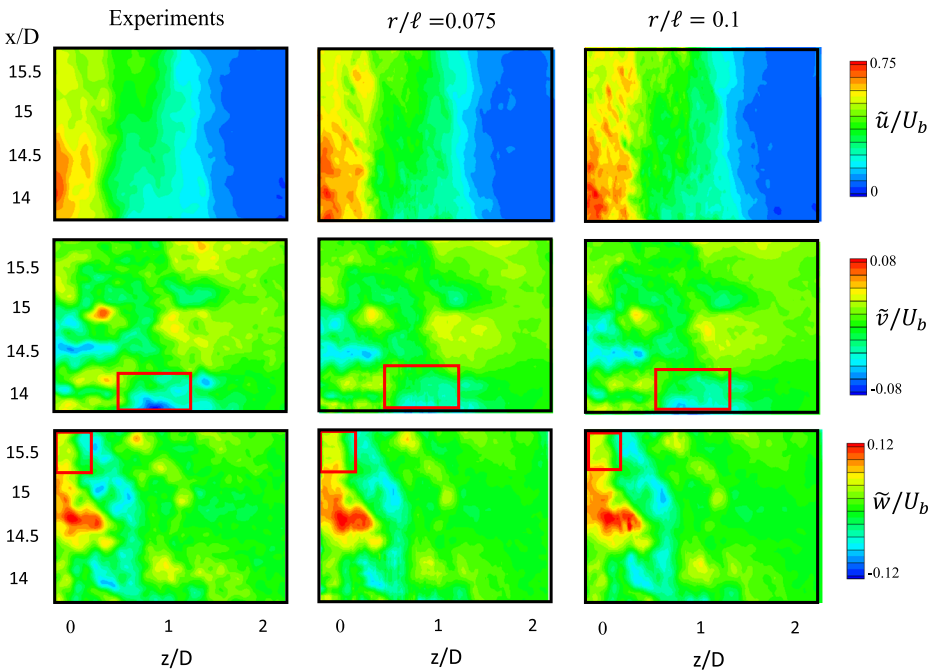
**Table 6** Reduction in mean normalized errors for different localizations when assimilating 70,000 velocity vectors

$r/\ell$	$E_u$	$E_v$	$E_w$	$E_{\mathbf{u}}$
0.05	0.53	0.67	0.54	0.57
0.075	<b>0.68</b>	0.83	0.83	0.77
0.1	0.66	<b>0.86</b>	<b>0.86</b>	<b>0.78</b>
0.2	0.44	0.84	0.76	0.65

Maximum reduction for each value of  $E_{\phi}$  shown in bold

matrix. It should also be noted that the localization radius that minimizes the overall error and error for each of the velocity components is not necessarily the same, as can be seen for the middle data set shown in Table 6.

The resulting velocity fields for two localization radii,  $r/\ell = 0.075$  and  $r/\ell = 0.1$ , are shown in Fig. 17, along with the experimental data. Overall, the error reduction in the axial velocity for these two localization radii are within 2%. This is perhaps due to the stochastic nature of the EnKF method chosen for the current study. In comparison, the flow field obtained with a localization of  $r/\ell = 0.1$  is on average in better agreement with the experimental measurements for the  $\tilde{v}$  and  $\tilde{w}$  velocity fields, demonstrating that the behavior of the axial velocity within the EnKF differs. We attribute this different behavior to the ratio between errors associated with the numerical model, assessed from the  $\mathbf{P}$  matrix and experimental uncertainties from the  $\mathbf{C}_{\epsilon\epsilon}$  matrix being closer to unity for the axial velocity.



**Fig. 17** Velocity contours for two localization values compared to experimental measurements. Red boxes highlight areas where a localization radius of  $r/\ell = 0.01$  is in better agreement with the experimental data

## 7 Conclusions

Three-dimensional velocity fields obtained from high-speed and high-resolution TPIV were assimilated into LES using an ensemble Kalman filter for state estimation of a turbulent inert jet. A baseline LES, without DA, demonstrated that the current simulation accurately reproduced the statistical behavior of the flow-field in the experimental measurement window. The predictability of the LES was evaluated using the Lyapunov exponent and found to be approximately 0.3 ms, three times longer than the experimental measurement interval (0.1 ms). Thus, the current LES and experimental data set represent a viable test case for assessing the performance of the ensemble Kalman filter and for understanding its impact on the state estimation. The impact of different aspects of the DA-algorithm, such as data localization, uncertainty in measurement data, assimilation frequency, sparsity of experimental data and number of ensembles employed were investigated.

A series of simulations were performed with different localization parameters. For the configuration under consideration with the prior error-covariance matrix estimated from six ensembles, a localization radius of  $r/\ell = 0.075$  was found to produce the best results. Smaller localization radii resulted in the assimilation of an incomplete representation of the flow field, whereas larger localization radii reduce the quality of the results due to the deterioration of the quality of the sample prior error-covariance matrix. Using this optimal localization radius, the measurement uncertainty utilized within the assimilation algorithm was varied to quantify the range of possible solutions that could be obtained from the algorithm. It was found that the reduction in the mean error between experimental data and the solution obtained from EnKF strongly depends on the assumed experimental error. However, even with “perfect” experimental data some error remained as the numerical models were imperfect. Thus, improved error quantification is required to provide a better estimate of the experimental errors.

A reduction in the assimilation frequency was found to lead to a reduction in the qualitative prediction of the state estimate. However, it was shown that the use of an assimilation frequency on the order of the characteristic predictability time as determined by the Lyapunov exponent enables assimilation with little degradation in the updated state vector. An increase in the number of ensembles improved the estimation of the prior error-covariance matrix. Increasing the number of ensembles from six to eighteen was found to reduce the error in the recovered solution by between 10–58% for the different velocity components but increased the computational cost by a factor of three. Further, it was shown that the optimal localization radius increases as the data available for assimilation decreases. In addition, the quality of the recovered solution was found to deteriorate as less experimental data were available for assimilation.

The present study demonstrates that the EnKF algorithm provides a robust method for assimilating measurements into simulations, thereby enabling the direct utilization of multidimensional high-speed experiments for state estimation. The flexibility of the algorithm allows for accommodating different levels of data sparsity, measurement frequencies, localization and computational resources. With this, model evaluations and comparisons can be performed by assessing the degree of deviation from experimental data and the size of the updated region during the assimilation step.

The current work demonstrates that available computational resources and high repetition rate experimental capabilities are at a point where model evaluation through the use of data assimilation is becoming feasible. The application of the ensemble Kalman filter and other assimilation techniques for parameter estimation applied to LES is also of interest, but a full evaluation of its potential is left for future work.

**Acknowledgments** The authors gratefully acknowledge financial support through NASA with award NNX15AV04A and the Stanford-Ford Alliance. J.H. Frank and B. Coriton gratefully acknowledge the support of the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

## Compliance with Ethical Standards

**Conflict of interests** The authors declare that they have no conflict of interest.

## References

1. Tennekes, H., Lumley, J.L.: *A First Course in Turbulence*. MIT Press, Cambridge (1972)
2. Nastac, G., Labahn, J.W., Magri, L., Ihme, M.: Lyapunov exponent as a metric for assessing the dynamic content and predictability of large-eddy simulations. *Phys. Rev. Fluids* **2**(9), 094606 (2017)
3. Asch, M., Bocquet, M., Nodet, M.: *Data assimilation: methods, algorithms, and applications*. SIAM (2016)
4. Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, I., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. Royal Meteorol. Soc.* **137**(656), 553–597 (2011)
5. ECMWF: Annual report 2016, <https://www.ecmwf.int/en/annual-report-2016/contents> (2016)
6. Navon, I.M.: Data assimilation for numerical weather prediction: a review. In: *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, pp. 21–65. Springer (2009)
7. Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al.: The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**(3), 437–471 (1996)
8. Uppala, S.M., Kållberg, P.W., Simmons, A.J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, A.C.M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., Isaksen, I., Janssen, P.A.E.M., Jenne, R., McNally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J.: The era-40 re-analysis. *Q. J. Royal Meteorol. Soc.* **131**(612), 2961–3012 (2005)
9. Omrani, H., Drobninski, P., Dubos, T.: Using nudging to improve global–regional dynamic consistency in limited-area climate modeling: what should we nudge? *Clim. Dyn.* **44**(5–6), 1627–1644 (2015)
10. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**(C5), 10143–10162 (1994)
11. Jahn, W., Rein, G., Torero, J.L.: Forecasting fire growth using an inverse zone modelling approach. *Fire Safety J.* **46**, 81–88 (2011)
12. Gao, X., Wang, Y., Overton, N., Zupanski, M., Tu, X.: Data-assimilated computational fluid dynamics modeling of convection-diffusion-reaction problems. *J. Comput. Sci.* **21**, 38–59 (2017)
13. Suzuki, T.: Reduced-order Kalman-filtered hybrid simulation combining particle tracking velocimetry and direct numerical simulation. *J. Fluid Mech.* **709**, 249–288 (2012)
14. Labahn, J.W., Wu, H., Coriton, B., Frank, J.H., Ihme, M.: Data assimilation using high-speed measurements and LES to examine local extinction events in turbulent flames. *Proc. Combust. Inst.* **37**, 2259–2266 (2019)
15. Edwards, J.R., Patton, C.H., Mirgolbabaei, H., Wignall, T.J., Echekeki, T.: 4D data assimilation for large eddy simulation of high speed turbulent combustion. *AIAA Paper*, 2015–3836 (2015)
16. Meldi, M., Poux, A.: A reduced order model based on Kalman filtering for sequential data assimilation of turbulent flows. *J. Comput. Phys.* **347**, 207–234 (2017)



17. Sousa, J., García-Sánchez, C., Gorlé, C.: Improving urban flow predictions through data assimilation. *Build Environ.* **132**, 282–290 (2018)
18. Symon, S., Dovetta, N., McKeon, B.J., Sipp, D., Schmid, P.J.: Data assimilation of mean velocity from 2D PIV measurements of flow over an idealized airfoil. *Exp. Fluids* **58**(5), 61 (2017)
19. Rossmann, T., Knight, D.D., Jaluria, Y.: Data assimilation optimization for the evaluation of inverse mixing and convection flows. *Fluid Dyn. Res.* **47**(5), 051405 (2015)
20. Coriton, B., Frank, J.H.: High-speed tomographic PIV measurements of strain rate intermittency and clustering in turbulent partially-premixed jet flames. *Proc. Combust. Inst.* **35**(2), 1243–1250 (2015)
21. Ma, P.C., Lv, Y., Ihme, M.: An entropy-stable hybrid scheme for simulations of transcritical real-fluid flows. *J. Comput. Phys.* **340**, 330–357 (2017)
22. Vreman, A.W.: An eddy-viscosity subgrid-scale model for turbulent shear flow: algebraic theory and applications. *Phys. Fluids* **16**(10), 3670–3681 (2004)
23. Eswaran, V., Pope, S.B.: An examination of forcing in direct numerical simulations of turbulence. *Comput. Fluids.* **16**(3), 257–278 (1988)
24. Bannister, R.N.: A review of operational methods of variational and ensemble-variational data assimilation. *Q. J. Royal Meteorol. Soc.* **143**(703), 607–633 (2017)
25. Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F., Fisher, M., Pailleux, J.: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). i: formulation. *Q. J. Royal Meteorol. Soc.* **124**(550), 1783–1807 (1998)
26. Fisher, M., Andersson, E.: *Developments in 4D-Var and Kalman filtering*, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading (2001)
27. Lorenc, A.C., Rawlins, F.: Why does 4D-Var beat 3D-Var. *Q. J. Royal Meteorol. Soc.* **131**(613), 3247–3257 (2005)
28. Evensen, G.: *Data Assimilation: the Ensemble Kalman Filter*. Springer, Berlin (2009)
29. Welch, G., Bishop, G.: *An introduction to the Kalman filter*. Technical report TR 95-041, University of North Carolina at Chapel Hill (2001)
30. Evensen, G.: The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynam.* **53**(4), 343–367 (2003)
31. Schwartz, C.S., Romine, G.S., Sobash, R.A., Fossell, K.R., Weisman, M.L.: NCAR'S experimental real-time convection-allowing ensemble prediction system. *Weather Forecast.* **30**(6), 1645–1654 (2015)
32. Gaspari, G., Cohn, S.E.: Construction of correlation functions in two and three dimensions. *Q. J. Royal Meteorol. Soc.* **125**(554), 723–757 (1999)
33. Barlow, R.S., Frank, J.H.: Effects of turbulence on species mass fractions in methane/air jet flames. *Symp. (Int.) Combust.* **27**(1), 1087–1095 (1998)
34. Elsinga, G.E., Scarano, F., Wieneke, B., van Oudheusden, B.W.: Tomographic particle image velocimetry. *Exp. Fluids* **41**(6), 933–947 (2006)
35. Coriton, B., Steinberg, A.M., Frank, J.H.: High-speed tomographic PIV and OH PLIf measurements in turbulent reactive flows. *Exp. Fluids* **55**(6), 1743 (2014)
36. Wieneke, B.: PIV uncertainty quantification from correlation statistics. *Meas. Sci. Technol.* **26**(7), 074002 (2015)
37. Smits, A.J., Dussauge, J.P.: *Turbulent shear layers in supersonic flow*. Springer, Berlin (2006)
38. de Silva, C.M., Philip, J., Marusic, I.: Minimization of divergence error in volumetric velocity measurements and implications for turbulence statistics. *Exp. Fluids* **54**(7), 1557 (2013)
39. Mi, J., Nobes, D.S., Nathan, G.J.: Influence of jet exit conditions on the passive scalar field of an axisymmetric free jet. *J. Fluid Mech.* **432**, 91–125 (2001)
40. Kothnur, P.S., Tsurikov, M.S., Clemens, N.T., Donbar, J.M., Carter, C.D.: Planar imaging of CH, OH, and velocity in turbulent non-premixed jet flames. *Proc. Combust. Inst.* **29**, 1921–1927 (2002)
41. Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Syst. Mag.* **29**(3), 83–104 (2009)
42. Houtekamer, P.L., Mitchell, H.L.: Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **126**(3), 796–811 (1998)
43. Nerger, L.: On serial observation processing in localized ensemble Kalman filters. *Mon. Weather Rev.* **143**(5), 1554–1567 (2015)