Reducing the Cost of Ensemble-Based Data Assimilation in Multiple-Query Scenarios through Covariance Augmentation

Andrew F. Ilersich*
University of Toronto, Toronto, Ontario, M3H 5T6, Canada

Kyle A. Schau†, Joseph C. Oefelein‡, Adam M. Steinberg§
Georgia Institute of Technology, Atlanta, Georgia, 30332, USA

Masayuki Yano¶
University of Toronto, Toronto, Ontario, M3H 5T6, Canada

We present and assess a method to reduce the computational cost of performing data assimilation (DA) for reacting flow in multiple-query scenarios, where we consider several scenarios with similar underlying dynamics. We focus on ensemble-based DA, in particular the ensemble Kalman filter (EnKF). The accuracy of the EnKF, which depends on the quality of the sample covariance, improves with the ensemble size, but so does its computational cost. To reduce the ensemble size while maintaining accurate covariance, we propose a data-driven approach to augment the covariance based on the statistical behavior learned from model evaluations. We assess our augmentation method using a one-dimensional model problem and a synthetic reacting flow case. We show in both bases that ensemble size, and thus computational cost, may be reduced by a factor of three to four while maintaining accuracy.

I. Nomenclature

Variables

n_{en}	=	ensemble size
$n_{\text{en}B}$	=	background ensemble size
n_B	=	number of background distributions
n_t	=	number of time steps
$\mathbf{u}$	=	$n_u \times 1$ state vector
$\mathbf{U}$	=	$n_u \times n_{\text{en}}$ ensemble matrix: each row is a state variable and each column is an ensemble member
$\overline{\mathbf{U}}$	=	$n_u \times n_{\text{en}}$ mean over ensemble members, repeated over n_{en} columns
$\tilde{\mathbf{U}}$	=	$n_u \times n_{\text{en}}$ deviation from ensemble mean, i.e., $\mathbf{U} - \overline{\mathbf{U}}$
$\mathbf{C}$	=	$n_u \times n_u$ state sample covariance, i.e., $\frac{1}{n_{\text{en}}-1}\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T$
$\mathbf{v}$	=	$n_v \times 1$ observation vector calculated from state $\mathbf{u}$
$\mathbf{V}$	=	$n_v \times n_{\text{en}}$ observation ensemble matrix: each column is an observation calculated from an ensemble member
$\overline{\mathbf{V}}$	=	$n_v \times n_{\text{en}}$ mean over observation ensemble members, repeated over n_{en} columns
$\tilde{\mathbf{V}}$	=	$n_v \times n_{\text{en}}$ deviation from observation ensemble mean, i.e., $\mathbf{V} - \overline{\mathbf{V}}$
$\mathbf{D}$	=	$n_v \times n_v$ observation sample covariance, i.e., $\frac{1}{n_{\text{en}}-1}\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$
$\mathbf{v}^{\text{obs}}$	=	$n_v \times 1$ observation data vector, externally provided
$\mathbf{R}$	=	$n_v \times n_v$ observation noise covariance
$\mathbf{K}$	=	$n_u \times n_v$ Kalman gain
d	=	Wasserstein distance

*Graduate Student Research Assistant, Institute for Aerospace Studies, AIAA Student Member
†Graduate Student Research Assistant, Guggenheim School of Aerospace Engineering, AIAA Student Member
‡Professor, Guggenheim School of Aerospace Engineering, AIAA Associate Fellow
§Associate Professor, Guggenheim School of Aerospace Engineering, AIAA Associate Fellow
¶Assistant Professor, Institute for Aerospace Studies

$\mathbf{W}$ = $n_u \times n_{\mathrm{en}B}$ background ensemble matrix
$\mathbf{B}$ = $n_u \times n_u$ background covariance

Functions

$G(\cdot)$ = Nonlinear process model
$H(\cdot)$ = Nonlinear observation model

Subscripts

i = state variable $i \in [1, n_u]$
j = ensemble member $j \in [1, n_{\mathrm{en}}]$
k = observation time step $k \in [1, n_t]$
$k|k-1$ = forecast estimate at time step k, i.e., estimate given observation data up to time step $k-1$
$k|k$ = analysis estimate at time step k, i.e., estimate given observation data up to time step k
ℓ = background distribution index $\ell \in [1, n_B]$

II. Background and Introduction

Large eddy simulation (LES) has emerged as the preferred paradigm for modeling complicated turbulent combustion systems due to its superior ability to represent the geometry-dependent physics of large-scale fluid motion compared to Reynolds averaged Navier-Stokes (RANS) techniques. Nevertheless, the ability of LES to accurately represent a physical system is impeded by the need to model terms representing interactions between the resolved scales and subfilter scales [1]. Parallel to the development of LES has been the advent of laser diagnostic techniques with sufficient repetition-rate to acquire time-correlated quantitative data on key state variables in an experimental system, though spatial resolution remains limited [2]. While such time-resolved data have been useful to gain physical understanding, the dynamical information contained in the measurements has, thus far, not been effectively leveraged to directly improve LES models and/or the ability of LES to mimic complicated behaviors in experiments.

One potential method of more directly coupling experiments and LES is through data assimilation (DA). While DA has been extensively used in meteorology [3], its application to turbulent combustion is still in its infancy [4, 5]. Within the myriad of DA methods, the ensemble Kalman filter (EnKF) is attractive for turbulent combustion simulations due to its ability to treat highly nonlinear problems. It operates by forward propagating an ensemble of LES solutions through a 'forecast' step and periodically adjusting the solution based on experimental data in an 'analysis' step.

The EnKF requires estimating the uncertainty in the LES model through the covariance of the ensemble members. An accurate covariance estimate is essential to the success of the analysis step, i.e. the incorporation of experimental data. However, the covariance estimated from a small ensemble (compared to the number of significant eigenmodes of the system state) will inherently be underestimated. While methods have been developed to more accurately estimate the covariance in the context of meteorology simulations [6], these require user-selected parameters [6, 7] and are likely insufficient to describe the complexity of a turbulent reacting flow. These parameters are derived from prior understanding of the system, limiting their applicability to different problems.

Here, we present a novel data-driven method for estimating the covariance in the context of ensemble-based DA for turbulent combustion, which we term covariance augmentation. Information from statistically-resolved runs is retained and drawn upon in subsequent underresolved runs, characterizing the system and reducing the computational cost. This is especially useful for multiple-query problems, where several distinct simulations of similar systems are performed. Such a situation may occur, for example, when studying trends in a gas turbine combustor as a function of operating conditions. [8].

We assess the method using two test problems. The first is the KS equation, a 1-D model of a laminar flame front. The second is a simulation of extinction events in a non-premixed laminar flame. We first evaluate a *process model* starting from an initial state $\mathbf{u}_0$ and propagating it forward in time to produce the "ground truth" solution $\mathbf{u}_k$, where the subscript $k \in [1, n_t]$ denotes the observation time index. We then produce the corresponding synthetic observation data $\mathbf{v}_k^{\mathrm{obs}},\ k \in [1, n_t]$, by applying an *observation model* to the ground truth. We then perform EnKF DA on an ensemble of solutions using $\mathbf{v}_k^{\mathrm{obs}}$ as the measurements of the system state. In the forecast and analysis steps of the EnKF, we use the same process model and observation model used to produce the ground truth and observation data, so there is no

inherent model deficiency in the filter. We show that the proposed covariance augmentation method allows a significant reduction in ensemble size, and hence cost, without reducing accuracy.

III. Ensemble Kalman Filter

The ensemble Kalman filter (EnKF) is a DA technique that estimates a system state and the associated uncertainty from a known process model, observation model, and observation data. As mentioned in Section II, for our synthetic test problems, the EnKF uses process and observation models that are identical to those used to produce the ground truth and observation data. The process model used to propagate the true state $\mathbf{u}_{k-1}$ to observation time step k takes the form

$$\mathbf{u}_k = G\left(\mathbf{u}_{k-1}\right), \tag{1}$$

for a nonlinear operator $G\left(\cdot\right)$. In our case, $G\left(\cdot\right)$ represents one iteration of the KS equation or many successive time-steps of the fluid conservation equations (i.e. between observation times), depending on the problem of interest. The observation model used to produce observation data $\mathbf{v}_k^{\text{obs}}$ from true state $\mathbf{u}_k$ takes the form

$$\mathbf{v}_k^{\text{obs}} = H\left(\mathbf{u}_k\right) + \mathbf{r}_k, \tag{2}$$

for a nonlinear operator $H\left(\cdot\right)$ and a Gaussian noise term $\mathbf{r}_k \sim \mathcal{N}\left(\mathbf{0}, \mathbf{R}\right)$. The nonlinear operator $H\left(\cdot\right)$ models, for instance, the acquisition of PIV data associated with the ground-truth LES state.

The EnKF represents the state estimate and associated uncertainty as a Gaussian probability distribution in a Monte-Carlo fashion with an ensemble of states $\{\mathbf{u}_j\}_{j=1}^{n_{\text{en}}}$. The Gaussian probability distribution is represented fully with the mean and covariance, which the EnKF estimates with the sample mean and sample covariance of the ensemble. We initialize the ensemble based on our prior knowledge of the state distribution. For instance, in our synthetic examples considered in Sections V and VI, the ensemble members are drawn randomly from the history of the ground truth

$$\{\mathbf{u}_{k_1},\ \mathbf{u}_{k_2},\ \ldots,\ \mathbf{u}_{k_{n_{\text{en}}}}\}, \qquad k_1, \ldots, k_{n_{\text{en}}} \sim \mathcal{U}(1, n_t) \tag{3}$$

where $\mathcal{U}$ is the discrete uniform distribution. This ensures that the ensemble members are both physically plausible and very poorly converged. The task of the filter therefore is to converge the ensemble around the ground truth.

For each time step k, the filter produces a state estimate in a two-step process. We calculate a forecast ensemble by directly applying (1) to each ensemble member

$$\mathbf{u}_{j\,k|k-1} = G\left(\mathbf{u}_{j\,k-1|k-1}\right), \qquad j = 1, \ldots, n_{\text{en}}. \tag{4}$$

We define the forecast ensemble matrix $\mathbf{U}_{k|k-1} = \begin{bmatrix} \mathbf{u}_{1\,k|k-1} & \ldots & \mathbf{u}_{n_{\text{en}}\,k|k-1} \end{bmatrix}$ and the ensemble mean matrix $\overline{\mathbf{U}}_{k|k-1} = \begin{bmatrix} \overline{\mathbf{u}}_{k|k-1} & \ldots & \overline{\mathbf{u}}_{k|k-1} \end{bmatrix}$. We then find the sample covariance

$$\mathbf{C}_{k|k-1} = \frac{1}{n_{\text{en}}-1}(\mathbf{U}_{k|k-1} - \overline{\mathbf{U}}_{k|k-1})(\mathbf{U}_{k|k-1} - \overline{\mathbf{U}}_{k|k-1})^T = \frac{1}{n_{\text{en}}-1}\tilde{\mathbf{U}}_{k|k-1}\tilde{\mathbf{U}}_{k|k-1}^T, \tag{5}$$

where we define the ensemble deviation matrix $\tilde{\mathbf{U}}_{k|k-1} = \mathbf{U}_{k|k-1} - \overline{\mathbf{U}}_{k|k-1}$. We also introduce the observation ensemble $\mathbf{V}_{k|k-1} = \begin{bmatrix} \mathbf{v}_{1\,k|k-1} & \ldots & \mathbf{v}_{n_{\text{en}}\,k|k-1} \end{bmatrix} = \begin{bmatrix} H\left(\mathbf{u}_{1\,k|k-1}\right) & \ldots & H\left(\mathbf{u}_{n_{\text{en}}\,k|k-1}\right) \end{bmatrix}$ calculated from the forecast ensemble. We then find the observation sample covariance

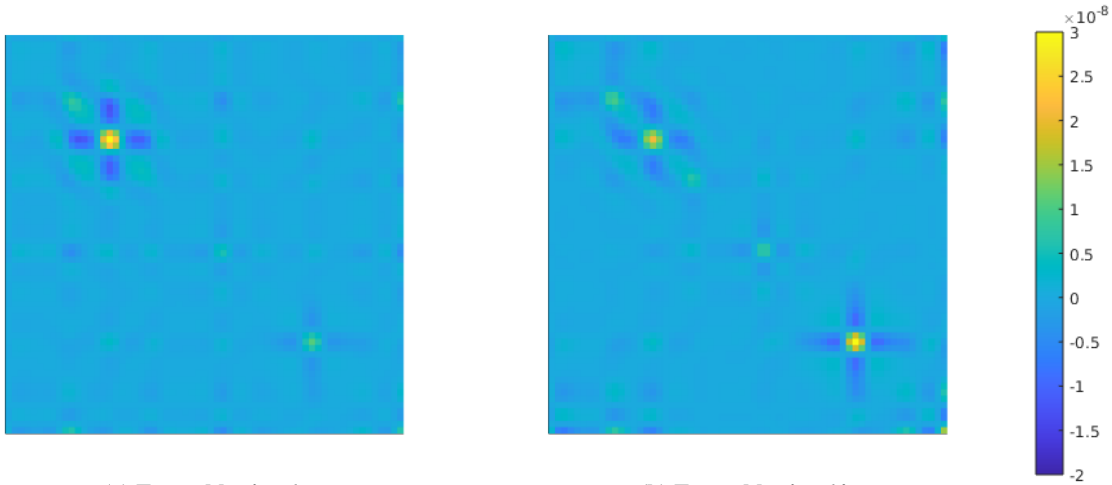
$$\mathbf{D}_{k|k-1} = \frac{1}{n_{\text{en}}-1}\left(\mathbf{V}_{k|k-1} - \overline{\mathbf{V}}_{k|k-1}\right)\left(\mathbf{V}_{k|k-1} - \overline{\mathbf{V}}_{k|k-1}\right)^T = \frac{1}{n_{\text{en}}-1}\tilde{\mathbf{V}}_{k|k-1}\tilde{\mathbf{V}}_{k|k-1}^T. \tag{6}$$

In the analysis step, we incorporate the observation data $\mathbf{v}_k^{\text{obs}}$ given by (2) to compute the analysis estimate

$$\mathbf{u}_{j\,k|k} = \mathbf{u}_{j\,k|k-1} + \mathbf{K}_k\left(\mathbf{v}_k^{\text{obs}} - \mathbf{v}_{j\,k|k-1}\right), \qquad j = 1, \ldots, n_{\text{en}}, \tag{7}$$

where the *Kalman gain* is given by

$$\mathbf{K}_k = \frac{1}{n_{\text{en}}-1}\tilde{\mathbf{U}}_{k|k-1}\tilde{\mathbf{V}}_{k|k-1}^T\left(\mathbf{D}_{k|k-1} + \mathbf{R}\right)^{-1}, \tag{8}$$



(a) Ensemble size 6 **(b) Ensemble size 64**

Fig. 1 Examples of 64×64 sample covariance matrices from EnKF applied to the KS equation. (a) shows a rank-deficient covariance estimate while (b) shows one that is well-resolved.

and $\mathbf{R}$ is the covariance of the observation noise term in (2). For a linear observation model $H(\cdot) = \mathbf{H}$, the Kalman gain is

$$\mathbf{K}_k = \mathbf{C}_{k|k-1}\mathbf{H}^T \left(\mathbf{H}\mathbf{C}_{k|k-1}\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{9}$$

which makes explicit the dependence on forecast covariance $\mathbf{C}_{k|k-1}$.

Because the forecast covariance $\mathbf{C}_{k|k-1}$ is used to calculate the Kalman gain $\mathbf{K}_k$, accurately resolving the covariance is critical to the filter's performance. This is determined by the ensemble size. An underlying "true" covariance with many significant eigenmodes will require a larger ensemble size to properly represent, as this requires a covariance matrix of higher rank.

An example is presented in Figure 1 of a well-resolved and an under-resolved sample covariance. These are obtained from applying the EnKF to the KS equation, the model problem considered in Section V. The rank-limited approximation in Figure 1a contains spurious cross-correlation terms and underestimates the magnitude of most variance terms along the diagonal. In practice, this leads to an underestimated covariance, i.e., the uncertainty in the state estimate computed from the ensemble is much smaller than the true error in the state estimate. The state then becomes resistant to correction in the analysis step of the filter.

One should note from the analysis step (7) and the definition of the Kalman gain (8) that the linear update ultimately lies in the column space of the ensemble deviation matrix $\tilde{\mathbf{U}}_{k|k-1}$. An undersized ensemble leads to a heavily restricted vector space, increasing the likelihood that the desired update is outside the column space of $\tilde{\mathbf{U}}_{k|k-1}$ and thus cannot be applied. As we demonstrate, this has the greater effect on the performance of an undersized ensemble than underestiamted covariance.

An artificial resolving technique attempts to correct the consequences of under-resolved covariance without the extra process model evaluations that come with increased ensemble size. These methods have seen significant attention in the literature [6, 7, 9], as they promise improved performance for negligible increase to computational cost. Two popular approaches are *inflation* and *localization*.

Inflation is an artificial resolving technique that prevents an undersized ensemble from converging toward the mean by artificially spreading the ensemble members from the mean by an inflation factor [6, 9, 10]. While inflation does increase the ensemble spread, it is a rank-preserving operation. If the performance is limited primarily by the rank-deficient covariance, then inflation is not a useful technique. It also may be the case, as we see in Section V, that undersized ensembles do not necessarily underestimate the covariance, making inflation irrelevant.

Localization is an artificial resolving technique that aims to suppress spurious long-distance correlation terms in the forecast covariance by enforcing a spatial correlation length. In practice, the performance of localization is extremely dependent on the choice of this length scale. We therefore require an accurate understanding of the dynamical system to avoid compromising performance with a poorly-chosen length scale. As DA is most commonly applied in meteorology, length scale estimation for localization has been extensively studied in that context [11, 12]. This requirement that we

have prior understanding of the system makes it difficult to apply localization to novel problems. The assumption of one dominant length scale is also unlikely to be true in the context of reacting flow, where different structures within the flow may have associated length scales that span several orders of magnitude [13]. Any choice of length scale for localization will either result in suppressing genuine larger-scale correlation terms or permitting spurious smaller-scale terms. Initial tests of inflation and localization indicate that these techniques are not well-suited to the test problems considered in Sections V and VI.

IV. Covariance Augmentation

Here we propose a data-driven *covariance augmentation* approach to artificially resolving the forecast covariance in a two-phase process.

1) A *generating run* performs the EnKF with a sufficiently large ensemble size to resolve the sample covariance. Information from this generating run is retained in a set of distributions, each of which is represented by an ensemble.
2) Subsequent *augmented runs* perform the EnKF with an undersized ensemble, perhaps extremely so, which we then augment by supplementing the deficient natural ensemble with artificial ensemble members. These artificial members are drawn from an appropriate distribution selected from the set produced in the generating run.

These artificial members are so named because they are not produced by evaluating the process model with every forecast step, unlike the "natural" ensemble members. They are only used to artificially resolve the forecast covariance and improve the quality of the analysis update.

This method is designed for multiple-query scenarios, where similar simulations are performed across a wide parameter space. Under normal circumstances, the computational cost scales linearly with n_f, the number of scenarios considered. If we assume that the simulation model is much more costly than the DA, then the runtime complexity is $O\left(n_{\text{en}} n_f\right)$ for n_f EnKF runs of a fixed n_{en}.

When we apply our filter to many similar scenarios, each generating run further "trains" our statistical model of the system. The more resolved this statistical model, the more that augmented runs may rely on the artificial members generated and reduce the natural ensemble size n_{en}. For a given level of accuracy, the complexity of a multiple-query problem is now $O(n_{\text{en}}{}' n_f)$, where $n_{\text{en}}{}' < n_{\text{en}}$ is the size of the "natural" ensemble. It can be made smaller than the original ensemble size n_{en} thanks to covariance augmentation.

A. Generating Runs

The generating run is an unaugmented run of he EnKF that produces a set of n_B background covariance matrices $B = \{\mathbf{B}_1,\ \mathbf{B}_2,\ \ldots,\ \mathbf{B}_{n_B}\}$ that will be used to augment the true ensemble for the purposes of covariance estimation in future augmented runs. The EnKF must use an ensemble size large enough to ensure that the sample covariance is statistically converged.

To introduce the idea of background set in the simplest setting, consider the case of a background set with a single covariance matrix $\mathbf{B}$. The forecast deviation ensemble $\tilde{\mathbf{U}}_{k|k-1}$ from each time step is retained, forming the background ensemble

$$\tilde{\mathbf{W}} = \left[\begin{array}{cccc} \tilde{\mathbf{U}}_{1|0} & \ldots & \tilde{\mathbf{U}}_{k|k-1} & \ldots \quad \tilde{\mathbf{U}}_{n_t|n_t-1} \end{array} \right], \tag{10}$$

which is then used to estimate the background covariance

$$\mathbf{B} = \frac{1}{n_{\text{en}B} - 1} \tilde{\mathbf{W}} \tilde{\mathbf{W}}^T \tag{11}$$

for $n_{\text{en}B} = n_t n_{\text{en}}$ background ensemble members. This produces a "long-exposure" background distribution which retains and emphasizes steady behaviour.

In most physical systems however, the statistical behaviour is highly unsteady and therefore contextual. So rather than use only one background, our augmentation method generates and draws from the set of n_B backgrounds. The structure of a generating run with $n_B = 3$ backgrounds is illustrated in Figure 2. This poses a regime classification problem. In a generating run each new forecast ensemble must either be matched to the most appropriate background or made into its own new background. To sort forecast ensembles into n_B background ensembles, we use k-means clustering with the Wasserstein metric as the measure of similarity between distributions. The Wasserstein metric is chosen for its stability with low-rank distributions, as it does not involve a matrix inverse. It has also been used successfully in the past in

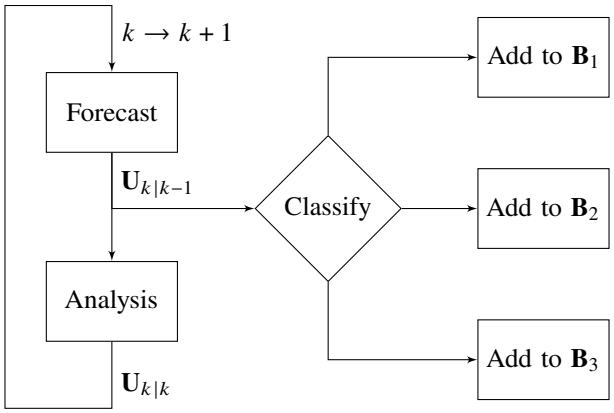


Fig. 2 Structure of a generating run using a library $B = \{\mathbf{B}_1,\ \mathbf{B}_2,\ \mathbf{B}_3\}$

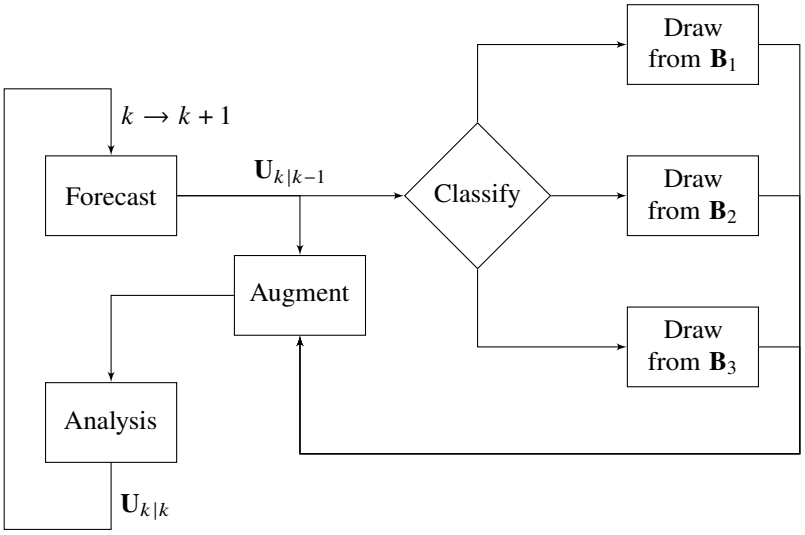


Fig. 3 Structure of an augmented run using a library $B = \{\mathbf{B}_1,\ \mathbf{B}_2,\ \mathbf{B}_3\}$

reacting flow regime classification [14]. When applied to two zero-mean Gaussian distributions with covariances $\mathbf{B}_\ell$ and $\mathbf{C}_{k|k-1}$, the Wasserstein distance d^2 is

$$d^2 = \mathrm{tr}\left(\mathbf{C}_{k|k-1} + \mathbf{B}_\ell - 2\left(\mathbf{C}_{k|k-1}^{1/2}\mathbf{B}_\ell\mathbf{C}_{k|k-1}^{1/2}\right)^{1/2}\right). \tag{12}$$

If the background set contains fewer than n_B distributions, the ensemble deviation matrix $\tilde{\mathbf{U}}_{k|k-1}$ with covariance $\mathbf{C}_{k|k-1}$ becomes its own background ensemble $\tilde{\mathbf{W}}_\ell$ with covariance $\mathbf{B}_\ell$. If the background set contains n_B distributions, we first match $\mathbf{C}_{k|k-1}$ to the background $\mathbf{B}_\ell$ with the lowest Wasserstein distance, then we append $\tilde{\mathbf{U}}_{k|k-1}$ to the corresponding background ensemble $\tilde{\mathbf{W}}_\ell$.

B. Augmented Runs

The augmented run is a run of the EnKF where the ensemble size may be significantly smaller than that required to statistically resolve the forecast covariance. The forecast ensemble is augmented with artificial members $\mathbf{U}_{\mathrm{art}_k}$ drawn from a steady distribution

$$\mathbf{U}_{\mathrm{art}_k} \sim \mathcal{N}\left(\overline{\mathbf{U}}_{k|k-1}, \mathbf{B}_\ell\right) \tag{13}$$

where the background covariance $\mathbf{B}_\ell$ is selected from a set $B = \{\mathbf{B}_\ell\}_{\ell=1}^{n_B}$ using the classification algorithm based on the Wasserstein metric. We then form an augmented forecast ensemble matrix from the natural and artificial members,

$$\mathbf{U}_{k|k-1} \leftarrow \left[\begin{array}{cc} \mathbf{U}_{k|k-1} & \mathbf{U}_{\mathrm{art}_k} \end{array} \right]. \tag{14}$$

This augmented ensemble is used in place of the natural ensemble in (5), providing an augmented sample covariance. This in turn provides an augmented Kalman gain in (8), which we apply in the analysis step. The structure of an augmented run with $n_B = 3$ backgrounds is illustrated in Figure 3.

The forecast ensembles from augmented runs are not used to further resolve the background distributions. Because the background distributions are used to form the augmented forecast ensemble and this, in turn, affects the convergence of natural ensemble members, any new information is not independent of the information already present in the background set. This would introduce a potential source of bias in our statistical model. Generating runs and augmented runs are therefore mutually exclusive.

V. Model Problem Results

In this section, we apply the EnKF with and without augmentation to a 1-D model problem, the KS equation. We follow the synthetic test problem formulation described in Section II.

There are two types of plots that will be used to demonstrate the filter performance: error and uncertainty. In the continuous case, the error over time ϵ_k is calculated by taking the L^2 norm over the domain Ω of the error between the ensemble mean and the reference solution, and the uncertainty σ_k is defined analogously in terms of the variance in the distribution

$$\epsilon_k = \sqrt{\frac{\int_\Omega \left(u_k - \overline{u}_{k|k}\right)^2 dx}{\int_\Omega u_k^2 dx}}, \qquad \sigma_k = \sqrt{\frac{\int_\Omega \mathrm{var}\left(u_{k|k}\right) dx}{\int_\Omega u_k^2 dx}}. \tag{15}$$

When the ensemble size is sufficiently large, we can consider the EnKF to be “well-resolved". We identify this as the ensemble size where the error and uncertainty have reached their respective asymptotic limits.

There are two aspects to the performance of an EnKF. First is its ability to converge from the initial ensemble toward the underlying trajectory. Second is its ability to “track" the trajectory once converged. For the initial convergence, the ensemble mean is not reflective of the true solution and the covariance is very large, so the filter relies on the observation and the cross-correlation terms in the covariance to converge the ensemble toward the true solution. Once converged, the filter needs to continually correct the state estimate if it begins to drift away from the true solution. This drift will occur as a consequence of using a chaotic model problem; by definition, slightly different initial conditions can develop into completely different states over a long enough time period. The ensemble must capture this drift over time in its members, which would increase the covariance. This then leads to greater correction from observation in the analysis step.

We take a moment now to discuss the plots used to present the results .We wish to assess the performance of the filter as a function of the ensemble size n_{en}. However, the error ϵ_{n_t} and the uncertainty σ_{n_t} at the final time are random variables because the EnKF state estimate depends on the initial ensemble, which is chosen randomly. Given this, we plot the probability density of the error ϵ_{n_t} and uncertainty σ_{n_t} as a function of n_{en} (e.g., in Figure 5). For each ensemble size, the EnKF is run repeatedly until empirical distributions of error and uncertainty are resolved. The trace in each plot represents the mean of the distribution. The shaded colourbars represent the probability density function, as found by kernel density estimation. Note also that the x-axis of these plots is scaled by the square root of the ensemble size. This is because, from the central limit theorem, the convergence rate of mean and standard deviation for a Monte Carlo approximation of a normal distribution is $O\left(1/\sqrt{n_{\mathrm{en}}}\right)$.

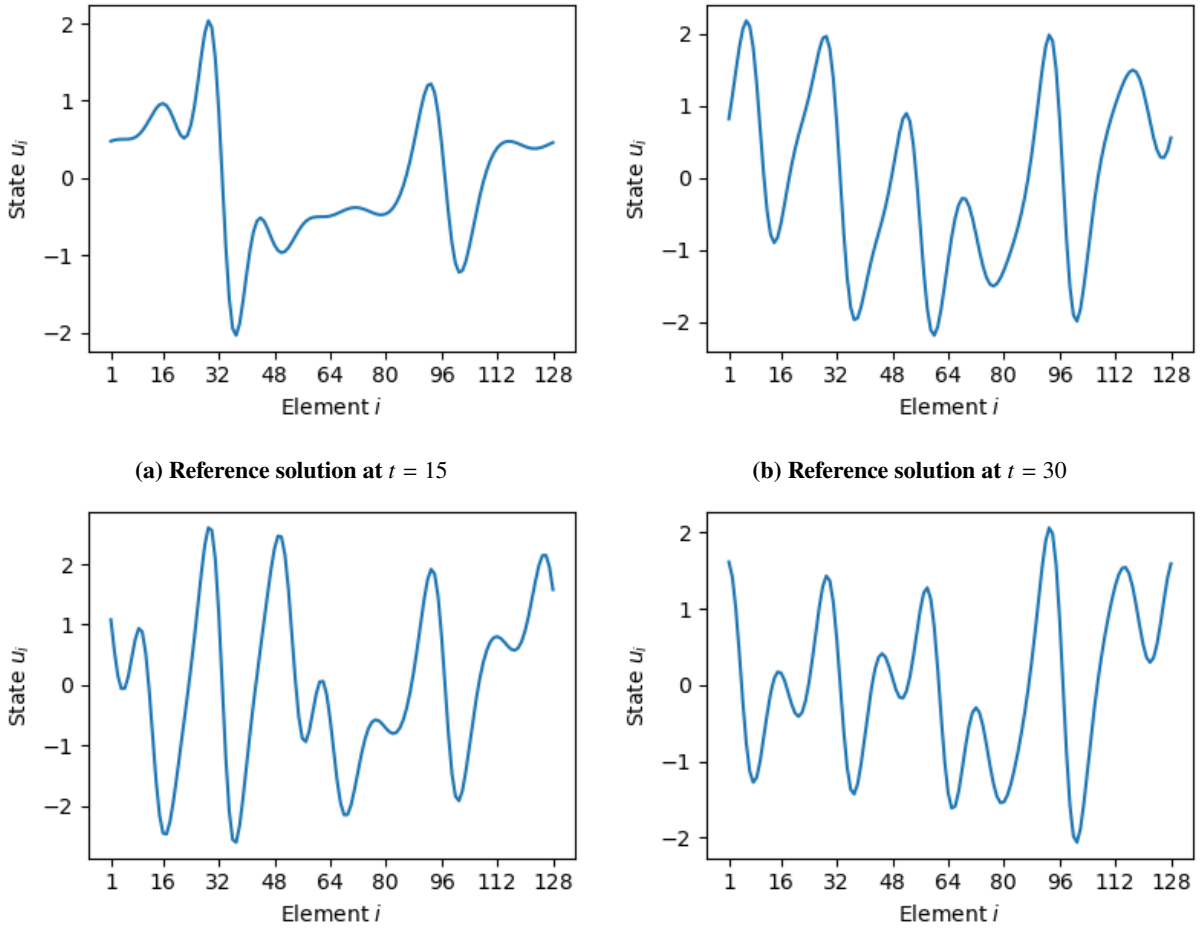
A. Problem Definition

The Kuramoto-Sivashinsky (KS) equation is a fourth-order nonlinear PDE

$$\frac{\partial u}{\partial t} = -\frac{\partial^4 u}{\partial x^4} - \frac{\partial^2 u}{\partial x^2} - u\frac{\partial u}{\partial x}, \tag{16}$$

which models instabilities in laminar flame fronts and exhibits chaotic behaviour [15, 16]. This model has one spatial dimension $x \in [0, 16\pi]$. There is a periodic boundary condition and the initial condition is

$$u(x, t = 0) = 5\cos\left(\frac{x}{8}\right)\left(1 + \sin\left(\frac{x}{8}\right)\right). \tag{17}$$



(a) Reference solution at $t = 15$

(b) Reference solution at $t = 30$

(c) Reference solution at $t = 45$

(d) Reference solution at $t = 60$

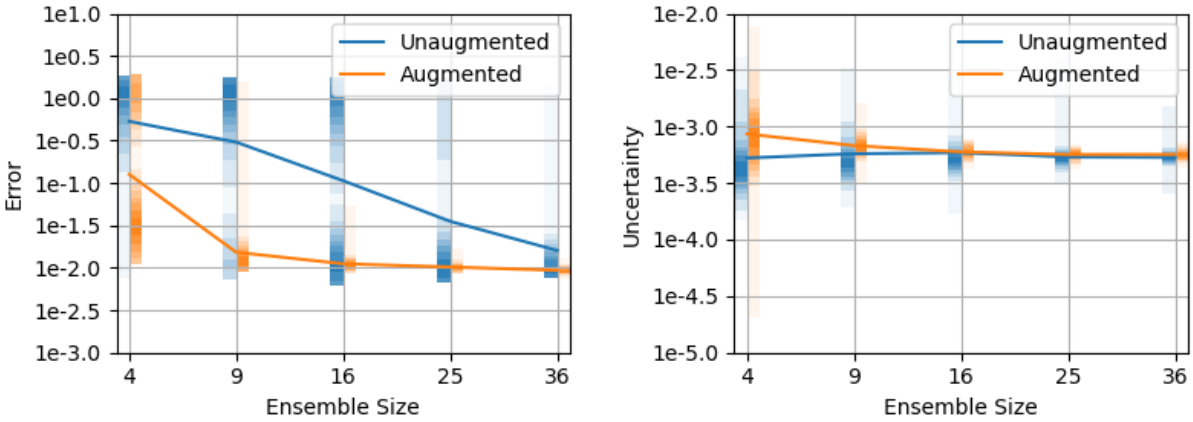
Fig. 4 Reference solution for KS equation

The KS equation is put into a semi-discrete form using second-order centered difference approximations for the spatial derivatives, which we then integrate in time with a backward differentiation formula. The spatial domain is discretized into $n_u = 128$ nodes, so $\Delta x = \pi/8$ and $i = 1, \ldots, 128$. A linear observation model is used with $n_v = 12$ observation nodes clustered into four evenly spaced groups of three.

The ground truth reference solution is presented in Figure 4. After an initial transient period, KS settles into a steady behaviour with “source" and “sink" terms roughly at $x = 12\pi$ (i.e., $i = 96$) and at $x = 4\pi$ (i.e., $i = 32$) respectively. The waves have a mostly uniform wavelength and velocity.

B. Baseline Performance

The baseline EnKF (i.e., without covariance augmentation) yields the results shown in blue in Figure 5. In this case, we consider the time interval $t \in [0, 60]$. Uncertainty in the baseline results maintains a relatively consistent mean around $10^{-3.25}$, though the spread converges with larger ensemble size. The mean error converges with the ensemble size, however the distribution is almost exclusively bimodal. The modes do not shift with increasing ensemble size, but the weight shifts from the unconverged mode to the converged mode. The same bimodal distribution is not present in the uncertainty however, suggesting that the ensemble converges about its mean regardless of whether it is tracking the true solution. The error approaches an asymptotic value of 10^{-2} around ensemble size 64, which is the ensemble size where we consider the EnKF “well-resolved."



(a) Error over ensemble size **(b) Uncertainty over ensemble size**

Fig. 5 KS Equation - Augmentation with Reproduction Scenario

C. Augmentation

There are two scenarios we consider when applying augmentation to the KS equation. First is the reproduction scenario, where the model trajectory in the generating run is the same as that in the augmented run. The second is a scenario where these trajectories differ, testing the applicability of the background set outside of the training data.

We produce $n_B = 50$ backgrounds from two generating runs with ensemble size of $n_{\text{en}} = 64$.

1. Reproduction Scenario

The reproduction scenario represents the best-case scenario for applying augmentation, as the training data and test data match as closely as possible. It does not reflect a practical use case, as it is necessarily more costly than simply running a well-resolved EnKF. Nonetheless, it is useful to provide an upper bound on performance. We consider the time interval $t \in [0, 60]$. The results are shown in Figure 5.

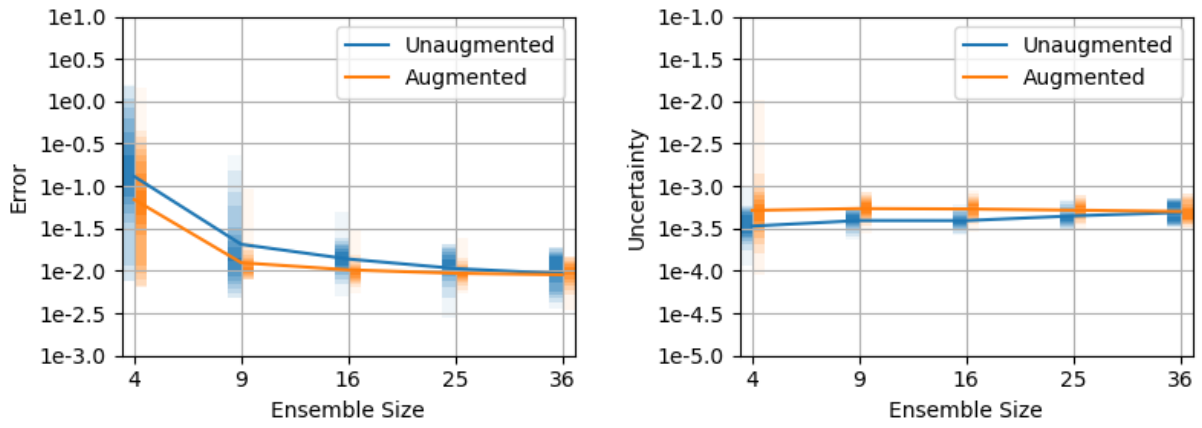
Because we know that there exist backgrounds in the set B that correspond almost perfectly with this trajectory, the test is really an evaluation of the chosen level of clustering and the classification scheme using Wasserstein distance. The performance is relatively consistent over all ensemble sizes and reliable even at extremely small ensemble sizes, suggesting that the clustered backgrounds are enough of a “compromise" that, though they are not perfectly suited to the natural ensemble, they are acceptable enough to improve performance where needed. Once the natural ensemble is large enough to be statistically-resolved, the augmentation becomes less consequential, as expected, but importantly does not adversely affect the performance.

The augmented EnKF demonstrates bimodal error at the smallest ensemble size of 4 which is comparable to the distribution at ensemble size 16, four times as large. It very quickly converges to a unimodal distribution by ensemble size 9, with an error distribution comparable to the unaugmented ensemble of size 36, again four times as large. Beyond this, the augmented EnKF approaches its asymptotic error limit far faster than the unaugmented one does, suggesting that if the background set is representative of model problem behaviour, the ensemble size required to match an appropriate background distribution is far smaller than that required to approximate the distribution itself.

2. Different Trajectory

A more practical test of augmentation is where the training data and the test data correspond to different model trajectories. By observing the improvement obtained from backgrounds generated using a different trajectory, and by starting the ground-truth simulation after the initial transient (i.e., from a statistically stationary state), this test examines the generalizability of the backgrounds, which is crucial for many-query scenarios. We use a background set generated from the time interval $t \in [0, 60]$. The results shown in Figure 6 with and without augmentation are for an offset time interval of $t \in [30, 90]$.

We expect to see improvement if at least one of two conditions holds. The first is if, when it comes to all possible trajectories, the generating runs can be considered to be reasonably comprehensive. In that case, any possible state has



(a) Error over ensemble size **(b) Uncertainty over ensemble size**

Fig. 6 KS Equation - Augmentation with Different Trajectory Scenario

an appropriate corresponding background. The second is if there is always steady, persistent statistical information in the backgrounds that may still be relevant to any possible state. While a given state may have a true covariance with dynamic elements not present in the backgrounds, the overall covariance estimate may still be improved by steady-state behaviour obtained from the backgrounds. This particular test

A consequence of this approach is that different flow behaviour is captured. Most notably, we can see in Figure 4 that a significant fraction of the original simulation period for the reproduction scenario in Section V.C.1 is spent in a transient period, where the state develops from initial conditions to its statistically stationary behaviour. In this different trajectory scenario, the generating run and consequently the background set include this initial transient, while the augmented run does not. Many backgrounds in the set are therefore superfluous to this augmented run, and we rely on the backgrounds produced from the developed, statistically-steady behaviour. Note also that because we consider fully-developed flow, the “no augmentation" trace in Figure 6 is different from the baseline results in Figure 5. It does not exhibit the same bimodal convergence.

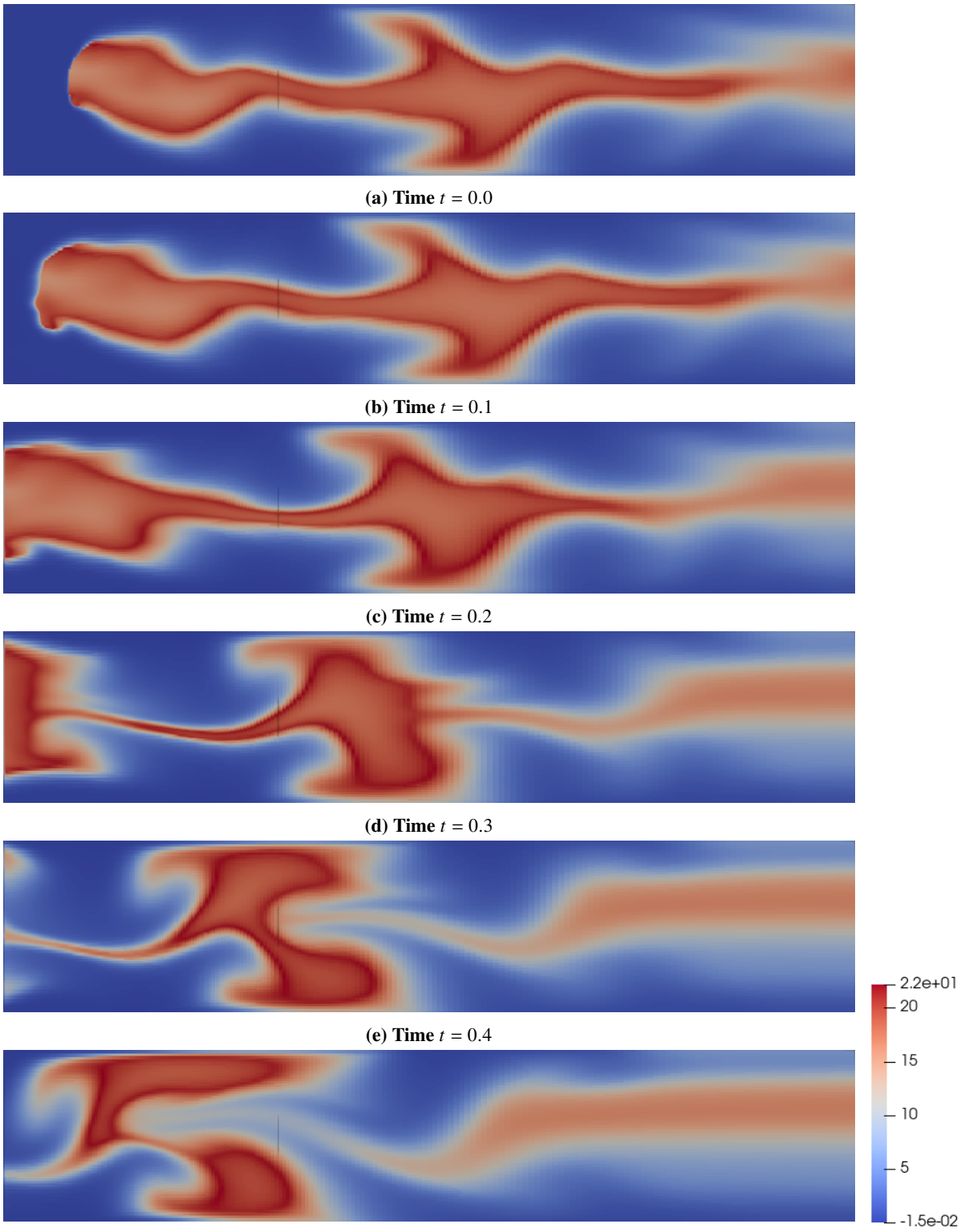
We note that that the unaugmented performance of EnKF over $t \in [3090]$ was better than over $t \in [060]$. Augmentation however still demonstrates improvement. Though the mean error is similar, the error spread in the unaugmented EnKF is much wider than in the augmented one. The augmented filter converges tightly on the asymptotic error limit by ensemble size 9, while the unaugmented filter does not until ensemble size 25–36, representing a 3–4× reduction in ensemble size. The unaugmented filter also underestimates uncertainty at most ensemble sizes, while the augmented filter again converges quickly to the asymptotic limit.

VI. Reacting Flow Results

We now apply the covariance augmentation technique to a reacting flow problem. Because this simulation is far more computationally intensive than the KS equation, the same level of statistical convergence with respect to ensemble size is not feasible. We plot individual runs with the error over time step, broken down by flow variable. Where there is significant variation in the results (i.e., at smaller ensemble sizes), the performance spread is characterized with a sample of five runs. The same initial ensembles used in these five runs are later reused in combination with augmentation. This ensures a fair comparison between the baseline and augmented results.

A. Problem Definition

We consider the extinction of a non-premixed propane-oxygen diffusion flame resulting from the shutdown of reactant flows. The simulation is performed with Raptor, a finite-volume solver developed by Oefelein [17]. Once the extinction event is triggered, it develops as shown in Figure 7. The simulation captures six chemical species: propane (C_3H_8), oxygen gas (O_2), carbon dioxide (CO_2), water (H_2O), carbon monoxide (CO), and nitrogen gas (N_2), following the simplified process



(a) Time $t = 0.0$

(b) Time $t = 0.1$

(c) Time $t = 0.2$

(d) Time $t = 0.3$

(e) Time $t = 0.4$

(f) Time $t = 0.5$

Fig. 7 Diffusion flame extinction event. The plot is of non-dimensionalized temperature. The direction of flow is from left to right.

$$2C_3H_8 + 3O_2 \rightarrow 6CO + 8H_2O,$$
$$2CO + O_2 \rightarrow 2CO_2,$$

to simulate the combustion chemistry.

The flow has a Reynolds number of 1200. The inflow boundary on the left side has a uniform flow velocity boundary condition, and the outflow boundary on the right side has a constant pressure boundary condition. The top and bottom boundaries are slip walls. We begin this simulation with a fully-developed flow as the initial condition. We then trigger the extinction event by disabling the inflow boundary. Once the extinction event begins, the model is run for $t \in [0, 0.5]$.

The observation model is a combination of simulated PIV and planar measurements to observe flow velocity and temperature respectively. The signal-to-noise ratios for PIV and temperature measurements are taken to be 250 and 60 respectively. The PIV interrogation boxes are a uniform 13×3 grid with 50% overlap over the domain shown in Figure 7 and the temperature observation nodes are a uniform 26×6 mesh over the domain. This PIV mesh would under normal circumstances be extremely coarse, however the application of DA allows us to resolve the flow at a much higher resolution than the raw observations. The pressure and chemical species fractions are not observed. Observations are incorporated with a period of $\Delta t = 0.05$. For the time interval $t \in [0, 0.5]$, there are 10 observation periods, i.e., 10 analysis steps. This excludes the initial time step.

To evaluate the performance of the EnKF, we use the same measures of error and uncertainty given in (15), respectively, extended to two dimensions. For this reacting flow, there are five fluid dynamic variables (pressure, three velocity components, and temperature) and five explicitly stored chemical species fractions (the last, N_2, is stored implicitly as $1 -$ all other species). The plots therefore have ten traces. Note that though we consider a 2-D case, Raptor is a 3-D solver and still produces 3-D velocity vectors. The z component of these however is consistently zero, hence not visible on the semilog plots of error.

B. Baseline Performance

We define two baseline scenarios that serve as reference points for evaluating the performance of augmentation. The first is a scenario where the ensemble is large enough to accurately approximate the forecast covariance. This provides the best-case performance and serves as the generating run. The second is a scenario where the ensemble size is deficient. This provides an example of poor performance that may be compared and contrasted with augmentation.

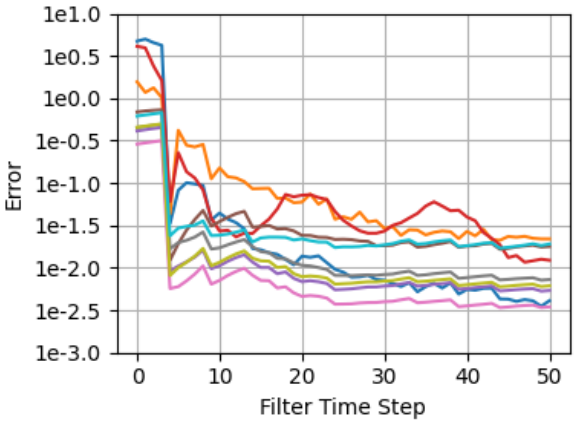
For the first case of a sufficiently large ensemble, we choose an ensemble size of 48, as it yields consistent performance with error between $10^{-1.5}$ and 10^{-2} for the velocity components and pressure, and between 10^{-2} and $10^{-2.5}$ for species fractions and temperature. For nondimensionalized flow variables of order 1, this corresponds to approximately 1–3% error in the velocity and pressure and 0.3–1% error in species fractions and temperature. An example of the performance is given in Figure 8a.

Here we also note the strength of the correlation across flow variables. This causes pressure and species fractions to converge despite not being directly observed. The largest persistent error is in the velocity and pressure components, despite velocity being directly observed via simulated PIV. This is due to the extinction event having a much greater effect on the fluid dynamics than on the chemistry of the flow, as well as the relatively lower spatial density of PIV observation compared to temperature observation.

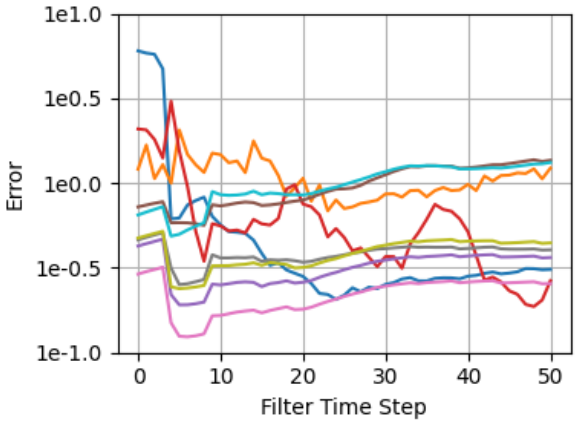
For the second case of an undersized ensemble, we choose an ensemble size of 12, as it yields highly variable performance with error between $10^{-0.5}$ and 10^{-2} and uncertainty below 10^{-2}. Figure 8b shows a typical case where the error remains high despite analysis updates, indicating that the ensemble has converged around a poor state estimate. There is drop in error and uncertainty after the first analysis update, similar to that in Figure 8a. The error however does not decrease significantly, and the ensemble thereafter is resistant to correction. The performance characteristics of five sample runs are shown in Figure 9.

C. Augmentation

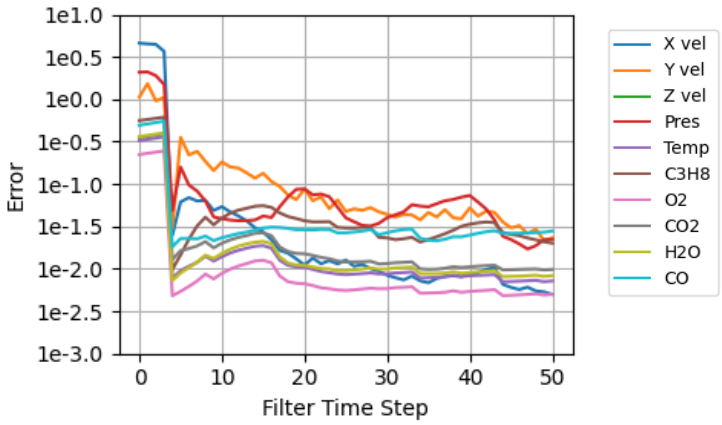
We consider the reproduction scenario here, where the model trajectory in the generating run is the same as that in the augmented run. We use a set of $n_B = 50$ backgrounds from one generating run. An example run is shown in Figure 8c and the spread is shown in Figure 9. The performance here is improved significantly from the unaugmented case. There remains a significant spread in the error among the five samples, though it is much reduced. The error mean is lowered by over half an order of magnitude. The uncertainty is more in line with the unaugmented filter of $n_{\text{en}} = 48$. It shows that when the entire trajectory of an augmented run is well-represented in the background set, the augmented



(a) Error over time with $n_{en} = 48$

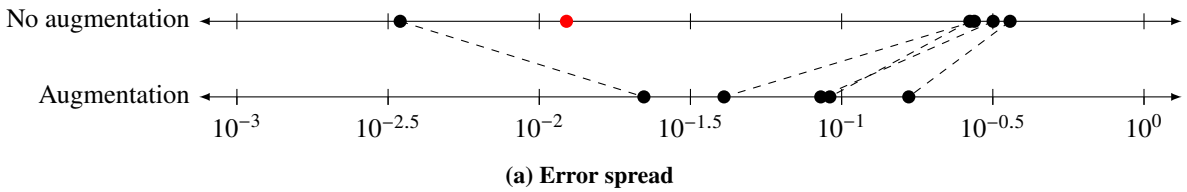


(b) Error over time with $n_{en} = 12$

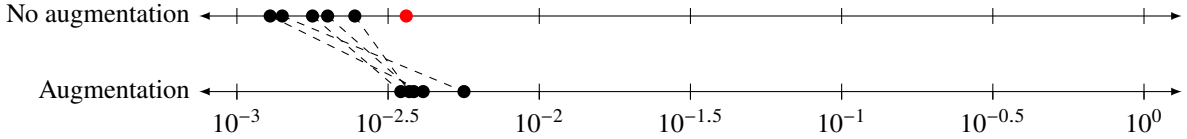


(c) Error over time with $n_{en} = 12$ & augmentation

Fig. 8 Reacting Flow - EnKF Performance - Example runs



(a) Error spread



(b) Uncertainty spread

Fig. 9 Reacting Flow with $n_{en} = 12$ - Change to error and uncertainty with augmentation - The red dot is the $n_{en} = 48$ case

ensemble performs more closely to a larger natural ensemble. This is despite it only being a quarter of the size of the original generating run ensemble. The more complicated test problem of reacting flow however makes it more difficult for augmentation to achieve a similar level of improvement to what it has shown for the KS equation, as it likely requires

more generating runs before its background set sufficiently captures most relevant statistical behaviour.

VII. Conclusion

In this paper we have demonstrated the technique of covariance augmentation, whereby we supplement the ensemble of an EnKF with artificial members drawn from background distributions. These backgrounds are produced from generating runs of the EnKF and reduce the computational cost of subsequent augmented runs. We have shown this method to be effective in a reproduction case with both the KS equation and a case of reacting flow. We have tested the method using a more practical case where the training and test data differ using the KS equation. Future work may further characterize this augmentation scheme or take a different approach to characterizing the statistical behaviour of a system, saving on computational cost and improving accuracy.

The present work is a *proof of concept* for data-driven approach to covariance augmentation for EnKF; the work has a number of limitations. First, we only demonstrated the augmentation technique for one configuration of reacting flow in a reproduction scenario; to accurately assess the robustness of the method and in particular the generalizability of the backgrounds, we must also consider non-reproduction scenarios as well as other flow configurations. Second, ultimately, the method should be tested for real flows and experimental observations, in which neither the process model nor the observation model, including the noise model, are perfect. Third, the proof-of-concept augmentation algorithm as presented leaves many potential areas of algorithmic improvement, ranging from better background training methods to classification schemes; the aforementioned more comprehensive assessment will inform future algorithmic development.

Acknowledgments

We thank Keishi Kumashiro at the University of Toronto Institute for Aerospace Studies for feedback, critique, and advice at all stages of this work.

This work was supported by the US Air Force Office of Scientific Research under Grant FA9550-17-1-0011 (Project Monitor Dr. Chiping Li) and the Ontario Graduate Scholarship. Computations were performed on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by the Canada Foundation for Innovation; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

References

[1] Pope, S. B., "Small scales, many species and the manifold challenges of turbulent combustion," *Proceedings of the Combustion Institute*, Vol. 34, No. 1, 2013, pp. 1–31. https://doi.org/10.1016/j.proci.2012.09.009.

[2] Talamelli, A., Persiani, F., Fransson, J., Alfredsson, P.-H., Johansson, A., Nagib, H., Rueedi, J.-D., Sreenivasan, K., and Monkewitz, P., "CICLoPE - A response to the need for high Reynolds number experiments," *Fluid Dynamics Research*, Vol. 41, 2009. https://doi.org/10.1088/0169-5983/41/2/021407.

[3] Houtekamer, P., and Zhang, F., "Review of the ensemble Kalman filter for atmospheric data assimilation," *Monthly Weather Review*, Vol. 144, No. 12, 2016, pp. 4489–4532. https://doi.org/10.1175/mwr-d-15-0440.1.

[4] Labahn, J., Wu, H., Coriton, B., Frank, J., and Ihme, M., "Data assimilation using high-speed measurements and LES to examine local extinction events in turbulent flames," *Proceedings of the Combustion Institute*, Vol. 37, No. 2, 2019, pp. 2259–2266. https://doi.org/10.1016/j.proci.2018.06.043.

[5] Yu, H., Jaravel, T., Ihme, M., Juniper, M. P., and Magri, L., "Data Assimilation and Optimal Calibration in Nonlinear Models of Flame Dynamics," *Journal of Engineering for Gas Turbines and Power*, Vol. 141, No. 12, 2019. https://doi.org/10.1115/1.4044378, URL https://doi.org/10.1115/1.4044378, 121010.

[6] Evensen, G., *Data assimilation: the ensemble Kalman filter*, Springer, 2009. https://doi.org/10.1007/978-3-642-03711-5.

[7] Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K., and Hunt, B. R., "Balance and Ensemble Kalman Filter Localization Techniques," *Monthly Weather Review*, Vol. 139, No. 2, 01 Feb. 2011, pp. 511 – 522. https://doi.org/10.1175/2010MWR3328.1, URL https://journals.ametsoc.org/view/journals/mwre/139/2/2010mwr3328.1.xml.

[8] Tucker, P., "Trends in turbomachinery turbulence treatments," *Progress in Aerospace Sciences*, Vol. 63, 2013, pp. 1–32. https://doi.org/https://doi.org/10.1016/j.paerosci.2013.06.001, URL https://www.sciencedirect.com/science/article/pii/S0376042113000547.

[9] Anderson, J. L., "An adaptive covariance inflation error correction algorithm for ensemble filters," *Tellus A*, Vol. 59, No. 2, 2007, pp. 210–224. https://doi.org/10.1111/j.1600-0870.2006.00216.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2006.00216.x.

[10] Law, K., Stuart, A., and Zygalakis, K., *Data Assimilation*, Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-20325-6.

[11] Wang, B., Liu, J., Liu, L., Xu, S., and Huang, W., "An approach to localization for ensemble-based data assimilation," *PLOS ONE*, Vol. 13, No. 1, 2018, pp. 1–23. https://doi.org/10.1371/journal.pone.0191088, URL https://doi.org/10.1371/journal.pone.0191088.

[12] Lei, L., Anderson, J., and Romine, G., "Empirical Localization Functions for Ensemble Kalman Filter Data Assimilation in Regions with and without Precipitation," *Monthly Weather Review*, Vol. 143, 2015, p. 150522112937007. https://doi.org/10.1175/MWR-D-14-00415.1.

[13] Poinsot, T., "Prediction and control of combustion instabilities in real engines," *Proceedings of the Combustion Institute*, Vol. 36, No. 1, 2017, pp. 1–28. https://doi.org/https://doi.org/10.1016/j.proci.2016.05.007, URL https://www.sciencedirect.com/science/article/pii/S1540748916300074.

[14] Johnson, R., Wu, H., and Ihme, M., "A general probabilistic approach for the quantitative assessment of LES combustion models," *Combustion and Flame*, Vol. 183, 2017, p. 88–101. https://doi.org/10.1016/j.combustflame.2017.05.004, URL http://dx.doi.org/10.1016/j.combustflame.2017.05.004.

[15] Sivashinsky, G., "Nonlinear analysis of hydrodynamic instability in laminar flames—I. Derivation of basic equations," *Acta Astronautica*, Vol. 4, No. 11, 1977, pp. 1177 – 1206. https://doi.org/https://doi.org/10.1016/0094-5765(77)90096-0, URL http://www.sciencedirect.com/science/article/pii/0094576577900960.

[16] Kuramoto, Y., "Diffusion-Induced Chaos in Reaction Systems," *Progress of Theoretical Physics Supplement*, Vol. 64, 1978, pp. 346–367. https://doi.org/10.1143/PTPS.64.346, URL https://doi.org/10.1143/PTPS.64.346.

[17] Oefelein, J. C., "Advances in Modeling Supercritical Fluid Behavior and Combustion in High-Pressure Propulsion Systems," *AIAA Science and Technology Forum and Exposition*, 2019, p. 0634. https://doi.org/10.2514/6.2019-0634, URL https://arc.aiaa.org/doi/abs/10.2514/6.2019-0634.