Augmenting Covariance Estimation for Ensemble-Based Data Assimilation in Multiple-Query Scenarios

Andrew F. Ilersich[a], Kyle A. Schau[b], Joseph C. Oefelein[b], Adam M. Steinberg[b], and Masayuki Yano[a]

[a]Institute for Aerospace Studies, University of Toronto, Toronto, Ontario, M3H 5T6, Canada; [b]Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA

ARTICLE HISTORY
Compiled July 19, 2022

ABSTRACT
We present and assess a method to reduce the computational cost of performing ensemble-based data assimilation (DA) for reacting flows in multiple-query scenarios, i.e. scenarios where multiple simulations are performed on systems with similar underlying dynamics. The accuracy of the DA, which depends on the accuracy 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 previous model evaluations. We assess our augmentation method using one-dimensional model problems and a two-dimensional synthetic reacting flow problem. We show in all these cases that ensemble size, and thus computational cost, may be reduced by a factor of three to four while maintaining accuracy.

KEYWORDS
Data assimilation; ensemble Kalman filter; data-driven modeling; reacting flow; multiple-query scenario.

1. Introduction

With few exceptions, e.g. [1, 2], the interface between combustion experiments and simulations typically has occurred either through design of 'clean' experimental apparatuses that are suitable for simulation validation, alignment of boundary conditions, and *a posteriori* comparison of statistical quantities. Although such interactions continue to be invaluable, the availability of dynamical (time-resolved) simulation and experimental data have the potential to enable new forms of interaction through data assimilation (DA). For example, data from high-fidelity time-resolved laser diagnostics may be used in a DA framework to adjust the evolution of large eddy simulations (LES) in order to better replicate an experiment and identify modeling challenges [1]. On the other hand, one may envision reduced order models of a combustor — perhaps even real-time digital twins [3] — that are meant to rapidly predict short-horizon behaviors being informed from relatively low fidelity sensors. This paper presents a method to reduce the computational cost of ensemble-based DA in the context of multiple-query scenarios —

CONTACT Andrew F. Ilersich. Email: andrew.ilersich@mail.utoronto.ca

when multiple (or long duration) simulations of a particular system are performed — through improved estimation of model covariance with low ensemble size.

Data assimilation seeks to incorporate experimental data into a mathematical model of the system of interest (in this case, a reacting flow) to estimate the system state or model parameters. DA solves for the estimate by weighing the uncertainty associated with the model and experimental data, finding a compromise that minimizes error in the estimate and is therefore most consistent with reality. In DA, we model the state estimate at a particular time t as a random field. This state random field is propagated forward in time in a two-step process. The first step is the *forecast step*, which utilizes a *process model* to produce a model-based prediction of the state random field at a future time $t + \Delta t$, where Δt is the observation period. The second step is the *analysis step*, which applies a correction to this state estimate based on experimental observation data. To recognize when a state estimate is discrepant with observation data, we use an *observation model* which estimates the observation data that would be produced by a given state. Now that we have a state estimate at time $t + \Delta t$ based on observation data up to time $t + \Delta t$, we move forward another increment Δt and repeat this two-step process.

DA is commonly used in meteorology, in particular numerical weather prediction (NWP), where sophisticated atmospheric models benefit from correction by sparse weather station data. The variational methods of 3dVar and 4dVar were commonly used historically [4, 5]; however, ensemble-based DA has seen increasing use over the past decade [6, 7]. This newer approach to DA [6–9] leverages increases in computational power to model the state random field in a Monte-Carlo fashion, i.e., with an ensemble of states that are propagated forward in time together. The prototypical example of this approach is the ensemble Kalman filter (EnKF) [10], which uses the ensemble representation to generalize the classical Kalman filter, allowing it to be applied to large-scale, nonlinear problems. In the forecast step, the EnKF propagates an ensemble of states forward in time by applying the process model to each ensemble member. In the analysis step, the EnKF weighs the uncertainties associated with the forecast estimate and observation data, and then applies an appropriate correction to the forecast estimate, obtaining the analysis estimate. The uncertainty is quantified using the sample covariance of the forecast ensemble and the known *a priori* covariance of noise in the observation data. Nevertheless, for complicated dynamical systems, as in atmospheric models, the ensemble size required to accurately represent the state mean and covariance can easily exceed 100 [6]. This proportionally increases computational cost, as each additional ensemble member requires another evaluation of the process model.

There exist two broad categories of methods that aid in resolving the state covariance, especially for a small ensemble size: *inflation* [9, 11] and *localization* [9, 12]. For a small ensemble size, the EnKF has a tendency to underestimate uncertainty in the estimate, leading the ensemble to converge towards the mean. Inflation increases the variance directly by spreading samples about the mean by a factor, making it more likely that the true state is within the confidence region. Localization is an alternative approach that suppresses spurious long-distance correlations by imposing a correlation length. There exist common practices to estimate the inflation factor [9]; however, estimating the correlation length requires some prior knowledge of the system. This is feasible in NWP where extensive literature is available [13–16], however not all systems are as well-characterized.

Though the application of DA to NWP is long-established, its application to reacting flow is still a recent development. Edwards *et al.* [17, 18] applied variational methods to an LES for two scenarios: a 2D hydrogen-air reaction and a 3D scramjet combustor.

Gao *et al.* [19] applied an EnKF, modified to preserve conservation and nonnegativity in the state estimate, to several test problems including a simple combustion simulation. Gray *et al.* [20] applied variational techniques to assess a shock-focusing geometry for a pulse detonation combustor for gas turbines, obtaining a refined numerical model that provided more detailed information than experimental data alone. Labahn *et al.* [1] applied the EnKF to a turbulent jet exhausting into a stationary fluid. Another study by Labahn *et al.* [21] applied the EnKF to a non-premixed turbulent flame, serving as a proof of concept for turbulent combustion as a whole. Yu *et al.* [22] applied the EnKF to a reduced-order model of a premixed flame to not only correct the state estimate, but to also correct model parameters, making the reduced-order model more accurate even without observation data. Yu *et al.* [23] used similar techniques to develop a reduced-order model for a ducted premixed flame.

All of the above works focus on DA for a single configuration; however, in practice, we often encounter multiple-query scenarios, in which we wish to solve a family of closely-related problems. This family may contain many different but closely related configurations [24], or it may contain a single configuration undergoing relatively slow changes in operating conditions, as in the digital twin configuration [3]. Whereas multiple-query problems are, by definition, more expensive than single-query problems, they also present an opportunity to reduce the marginal computational cost of each additional analysis by reusing the information gathered in previous analyses. To our knowledge, no published research has specifically considered strategies to improve the computational efficiency of DA in multiple-query scenarios.

The objective of this work is to reduce the cost of ensemble-based DA with an emphasis on multiple-query scenarios and on applications to reacting flows. We achieve this with a data-driven method for estimating the covariance, which we term *augmentation*, where information from statistically-resolved 'training' runs is retained and drawn upon in subsequent under-resolved runs. This statistical information characterizes the system and reduces the required ensemble size and hence the computational cost, while maintaining a desired level of accuracy in multiple-query scenarios. Moreover, unlike the aforementioned covariance modification techniques of inflation and localization, our data-driven approach does not rely on the user's prior knowledge of the system. This makes it suitable for reacting flow applications where prior knowledge may be limited.

We assess our augmentation method using two model problems based on the Lorenz 96 (L96) model – a simplified atmospheric model commonly used as a DA test problem [25] – and the Kuramoto-Sivashinsky (KS) equation, which models instabilities in laminar flame fronts and exhibits oscillatory and chaotic solutions [26–28]. The KS equation is derived from the species diffusion equation and the heat conduction equation in the limit of a large activation energy (i.e., the reaction rate strongly depends on temperature) and the Lewis number of near unity. We also assess our method with a case of simulated reacting flow, which we perform with an LES model with simplified chemistry. These are *observing-systems simulation experiments* (OSSEs), meaning that the 'ground truth' reference solution and the pseudo-observation data are both produced artificially [29]. Although the use of OSSEs admittedly masks some challenges associated with DA using real-world data, it allows us to quantitatively assess the accuracy of DA techniques against the known ground truth. We finally assess augmentation's ability to reduce the necessary ensemble size, and hence computational cost.

In Section 2, we discuss our OSSE methodology, in particular the ensemble DA technique and covariance estimation techniques considered. In Section 3, we discuss our proposed framework for covariance augmentation. In Section 4, we apply all covariance estimation techniques discussed to two model problems and assess the error and uncer-

tainty estimates. In Section 5, we apply augmentation to a reacting flow problem and assess the error and uncertainty estimates. Finally, in Section 6, we summarize our work and discuss its limitations.

2. Methodology

2.1. Problem Definition

We first introduce a 'ground truth' (or reference) solution $\{\mathbf{u}_1,\ \ldots,\ \mathbf{u}_{n_t}\}$ associated with the observation time instances $\{t_1,\ \ldots,\ t_{n_t}\}$, where n_t is the number of observation time instances. To this end, we start with an initial state vector $\mathbf{u}_0$. We then propagate the initial state forward from one observation time to the next using the *process model*, which takes the form

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

for a nonlinear operator $G\left(\cdot\right)$. In practice, a single application of the operator $G\left(\cdot\right)$ represents many successive time steps of a numerical time integrator. The process model may include a noise term to represent model inadequacy; however, in this work, we assume that the process model is exact.

Given the ground truth solution, we generate the associated observation data $\{\mathbf{y}_1^{\text{obs}},\ \ldots,\ \mathbf{y}_{n_t}^{\text{obs}}\}$. We do this by applying the *observation model* to the ground truth, which takes the form

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

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

2.2. Ensemble DA

Ensemble DA techniques represent the state estimate and associated uncertainty in a Monte-Carlo fashion using an ensemble of states $\{\mathbf{u}^j\}_{j=1}^{n_{\text{en}}}$. However, unlike a particle filter [30], the EnKF assumes that the state distribution is Gaussian. The Gaussian probability distribution is represented fully with the mean and covariance, which ensemble DA estimates with the sample mean and sample covariance of the ensemble. In general, the ensemble is initialized based on prior knowledge of the state distribution. For our OSSEs, the initial ensemble members are drawn randomly from the history of the ground truth

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

where $\mathcal{U}\{\cdot\}$ is the discrete uniform distribution. This ensures that the ensemble members are both physically plausible and, assuming the solution is unsteady, poorly converged. It is important for the initial condition to be physically plausible because some process models, especially for complex problems like the reacting flow in Section 5, may be very sensitive to whether the initial condition is nonphysical. If it is, then the forecast step may fail to propagate the state forward in time. Drawing the initial ensemble from

the ground truth ensures that the first forecast step is always successful, allowing a more straightforward comparison of error and uncertainty estimates between runs. The task of the filter therefore is to converge the ensemble around the ground truth and subsequently 'track' the ground truth, preventing error growth.

The ensemble DA technique used in this work is the ensemble transform Kalman filter (ETKF) from Bishop *et al.* [31]. Like the original ensemble Kalman filter (EnKF) from Evensen [10], and indeed all recursive DA techniques, the ETKF propagates a state forward in time via a two-step process, which we briefly summarize here.

To begin, suppose we are given an ensemble $\mathbf{u}^j_{k-1|k-1}$, $j = 1, \ldots, n_{\text{en}}$. The notation $\mathbf{u}^j_{k-1|k-1}$ indicates that the j-th ensemble member is associated with the state estimate at time step $k-1$ given observation data up to time step $k-1$. In the k-th *forecast step*, we apply the process model (1) to each ensemble member to produce the forecast estimate of the state $\mathbf{u}^j_{k|k-1} = G\left(\mathbf{u}^j_{k-1|k-1}\right)$, $j = 1, \ldots, n_{\text{en}}$. The notation $\mathbf{u}^j_{k|k-1}$ indicates that the j-th forecast estimate ensemble member is associated with the state estimate at time step k given observation data up to time step $k-1$. We estimate the forecast covariance $\mathbf{C}_{k|k-1}$ using the sample covariance of the ensemble,

$$\mathbf{C}_{k|k-1} = \frac{1}{n_{\text{en}} - 1} \sum_{j=1}^{n_{\text{en}}} (\mathbf{u}^j_{k|k-1} - \overline{\mathbf{u}}_{k|k-1})(\mathbf{u}^j_{k|k-1} - \overline{\mathbf{u}}_{k|k-1})^T, \tag{4}$$

where $\overline{\mathbf{u}}_{k|k-1}$ denotes the forecast ensemble mean. To rewrite (4) without the summation, we define the ensemble matrix $\mathbf{U}_{k|k-1} = \left[\begin{array}{ccc} \mathbf{u}^1_{k|k-1}, & \ldots, & \mathbf{u}^{n_{\text{en}}}_{k|k-1} \end{array} \right]$ and obtain

$$\mathbf{C}_{k|k-1} = \frac{1}{n_{\text{en}} - 1} \widetilde{\mathbf{U}}_{k|k-1} \widetilde{\mathbf{U}}^T_{k|k-1}, \tag{5}$$

where $\widetilde{\mathbf{U}}_{k|k-1} = \mathbf{U}_{k|k-1} - \overline{\mathbf{U}}_{k|k-1}$ and $\overline{\mathbf{U}}_{k|k-1} = \left[\begin{array}{ccc} \overline{\mathbf{u}}_{k|k-1}, & \ldots, & \overline{\mathbf{u}}_{k|k-1} \end{array} \right]$.

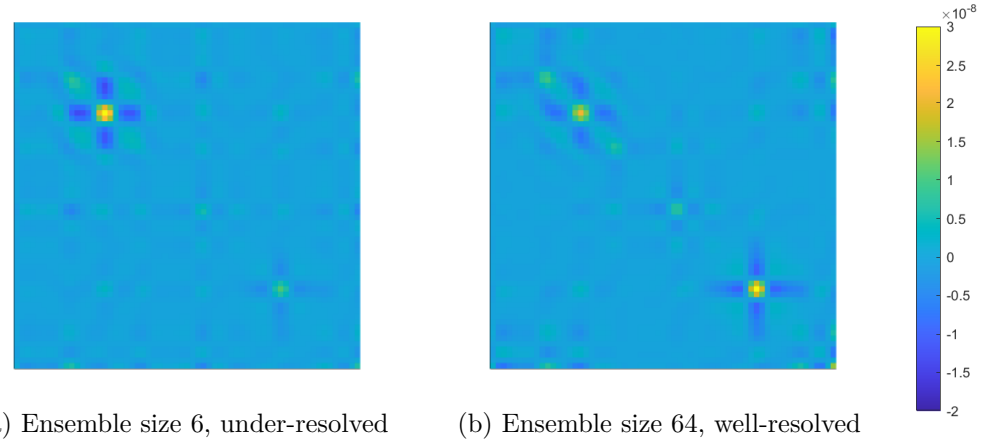
In the *analysis step*, we incorporate the pseudo-observation data $\mathbf{y}^{\text{obs}}_k$ to refine our forecast estimate. We apply a linear update to the ensemble mean $\overline{\mathbf{u}}_{k|k} = \overline{\mathbf{u}}_{k|k-1} + \mathbf{K}_k \left(\mathbf{y}^{\text{obs}}_k - H\left(\overline{\mathbf{u}}_{k|k-1}\right) \right)$ and transform the ensemble deviation matrix $\widetilde{\mathbf{U}}_{k|k} = \widetilde{\mathbf{U}}_{k|k-1} \mathbf{T}^{1/2}_k$ to produce the analysis estimate of the state. The Kalman gain $\mathbf{K}_k$ is given by

$$\mathbf{K}_k = \mathbf{C}_{k|k-1} \mathbf{H}^T \left(\mathbf{H} \mathbf{C}_{k|k-1} \mathbf{H}^T + \mathbf{R} \right)^{-1}, \tag{6}$$

where $\mathbf{H}$ is the linearized observation model. The Kalman gain $\mathbf{K}_k$ depends on the uncertainty associated with each source of information, i.e., the forecast covariance $\mathbf{C}_{k|k-1}$ and the observation covariance $\mathbf{R}$, and it can be shown that the gain minimizes the mean square error in the analysis estimate *if* both the process model and observation model are linear. There is some approximation in this formulation of the Kalman gain for nonlinear models, but it can still be computed.

2.3. Impact of the Ensemble Size on Covariance Estimation

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. The ensemble



(a) Ensemble size 6, under-resolved (b) Ensemble size 64, well-resolved

Figure 1.: Examples of sample covariance matrices from ETKF applied to the KS equation. This colour map is a visualization of the values in the covariance.

size required to accurately estimate the covariance depends on the number of dominant modes in the system, or, more precisely, the rate of decay of the eigenvalues of the underlying 'true' covariance; a 'true' covariance with many significant eigenmodes will require a larger ensemble size.

Figure 1 shows examples of a well-resolved and an under-resolved sample covariance. These are obtained from applying the ETKF to the KS equation, one of the model problems considered in Section 4. The 64×64 sample covariance matrices in this example are obtained from the forecast ensemble at the final DA time step. 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 spread of the ensemble is much smaller than the error in the state estimate. The underestimated covariance causes two problems: it can mislead the user to trust incorrect state estimates, and the state becomes resistant to correction in the analysis step of the filter. Section 4.3 provides a more detailed discussion of how ensemble size impacts ETKF performance for this example problem.

One should note that the linear correction applied by the analysis step lies in the subspace formed by the forecast ensemble. An extreme case of undersized ensemble is where the covariance is zero in most directions, increasing the likelihood that the desired update cannot be applied. When an ETKF is applied to a problem with complex dynamics and the ensemble size is several orders of magnitude smaller than the state dimension, this is a plausible occurrence.

2.4. Techniques to Improve Covariance Estimation

To avoid the consequences of an under-resolved sample covariance, there exist three broad categories of techniques to improve covariance estimation: adaptive ensemble sizing, covariance inflation, and covariance localization.

Adaptive ensemble sizing methods estimate the ensemble size required to sufficiently capture the system dynamics. If it determines the ensemble is not appropriately sized, it adds or removes members as necessary. Uzunoglu *et al.* [32] present an example of such a technique, which uses the decay of Shannon entropy in the eigenmodes of the ensemble covariance to determine whether the ensemble is undersized, oversized, or appropriately

sized.

Covariance inflation seeks to prevent false convergence of the ensemble members by spreading the ensemble members from the mean, increasing the covariance. Inflation requires an 'inflation factor' by which to spread the distribution, which may be fixed or estimated adaptively [9, 33]. An example adaptive inflation method by Anderson [11] maintains the ensemble spread before and after the analysis update.

Covariance localization seeks to suppress spurious correlation terms in the sample covariance by imposing a limiting correlation length appropriate to the system. Using this correlation length, apparent long-distance correlations may be suppressed. Localization may be performed by applying an elementwise mask to the forecast covariance, thus diminishing cross-correlation terms corresponding to spatially-distant nodes [12].

Inflation and localization utilize prior understanding about the system under consideration and are limited in their ability to artificially resolve the forecast covariance. Inflation requires one to estimate the degree to which the covariance is underestimated, and, because it is a rank-preserving operation on the forecast covariance, unrepresented eigenmodes are not accounted for. In the absence of prior understanding, there are common practices to estimate the inflation factor [9]; however, localization requires one to estimate the correlation length, which depends on the dynamics of the system. For the atmospheric models used in meteorology, this has been studied extensively in existing literature [13, 14]. For other problems, however, prior understanding of the system is more limited, making it more difficult to estimate a reasonable correlation length and apply localization.

3. Covariance Augmentation

The novel method proposed in this work is *covariance augmentation*, which takes a data-driven approach to improve forecast covariance estimation in a two-phase process.

(1) A *generating run* performs the DA with an ensemble of the size n_{en} that is large enough to resolve the sample covariance. Information from this generating run is retained in a library of distributions, each of which is represented by an ensemble.
(2) Subsequent *augmented runs* perform the DA using a combination of two ensembles: an (undersized) 'natural ensemble' of the size $n_{\text{en}}{}' < n_{\text{en}}$, which is used in both forecast and analysis steps; an 'artificial ensemble', which augments the 'natural ensemble' in the analysis step (only) to improve the quality of the sample covariance and is drawn from an appropriate distribution selected from the library produced in the generating run.

Covariance augmentation is designed for multiple-query scenarios. Under normal circumstances, the computational cost scales linearly with n_f, the number of scenarios considered. If we assume that the evaluation of the process model in the forecast step is much more costly than the state update in the analysis step, 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. This statistical model aims to capture all forecast covariances that the dynamical system may exhibit, informing the selection of artificial members. The more accurate this statistical model, the more that augmented runs may rely on the artificial members 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.

Augmentation is a general framework for training an ensemble filter with model-specific data, allowing it to reduce its natural ensemble size, and thus computational cost, while retaining accurate performance. Ours is only one potential realization of this framework. If we are given a library of all possible distributions of model state and a procedure to choose the most appropriate distribution from the library, then the ideal augmented ensemble would almost perfectly mimic a well-resolved ensemble, leading to very little difference between well-informed augmentation and a large natural ensemble. In practice, augmentation's performance is determined by

(1) The *classification scheme* that matches a natural forecast ensemble to the appropriate distribution from the library, and
(2) The *training strategy* that produces the library of distributions using generating runs.

3.1. Generating Runs

The generating run is an unaugmented evaluation of the EnKF that produces a library of n_B background covariance matrices $B = \{\mathbf{B}_1,\ \mathbf{B}_2,\ \ldots,\ \mathbf{B}_{n_B}\}$. This background library is used in later evaluations to augment the natural ensemble, improving forecast covariance estimation. The ensemble size in a generating run must therefore be large enough to ensure that the sample covariance is statistically converged.

To introduce the idea of background library in the simplest setting, consider the case of a background library with a single covariance matrix $\mathbf{B}$. From each time step, we retain the ensemble deviation matrix $\widetilde{\mathbf{U}}_{k|k-1}$, forming the background ensemble

$$\widetilde{\mathbf{W}} = \left[\ \widetilde{\mathbf{U}}_{1|0}\ \ \ldots\ \ \widetilde{\mathbf{U}}_{k|k-1}\ \ \ldots\ \ \widetilde{\mathbf{U}}_{n_t|n_t-1}\ \right], \tag{7}$$

which is then used to estimate the background covariance

$$\mathbf{B} = \frac{1}{n_{\mathrm{en}B} - 1}\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T, \tag{8}$$

for $n_{\mathrm{en}B} = n_t n_{\mathrm{en}}$ background ensemble members. This produces a 'long-exposure' background distribution which retains and emphasizes (statistically) steady behaviour.

In many physical systems however, the statistical behaviour can be highly unsteady. So rather than only one background, our augmentation method generates and draws from the library of n_B backgrounds. The structure of a generating run with $n_B = 3$ backgrounds is illustrated in Figure 2. For each time step, the natural ensemble is classified into one of the backgrounds in the library. It is then appended to the corresponding background ensemble

$$\widetilde{\mathbf{W}}_\ell \leftarrow \left[\ \widetilde{\mathbf{W}}_\ell\ \widetilde{\mathbf{U}}_{k|k-1}\ \right], \tag{9}$$

which leads to an updated background covariance $\mathbf{B}_\ell = \frac{1}{n_{\mathrm{en}B}-1}\widetilde{\mathbf{W}}_\ell\widetilde{\mathbf{W}}_\ell^T$. As the number of forecast ensembles in each background increases, the backgrounds yield a comprehensive library of n_B statistical regimes for the system. The classification scheme we use is detailed in Section 3.3.

Once all ensemble states from the generating run are clustered into n_B backgrounds, we refine the background clustering by applying a naive k-means clustering algorithm to

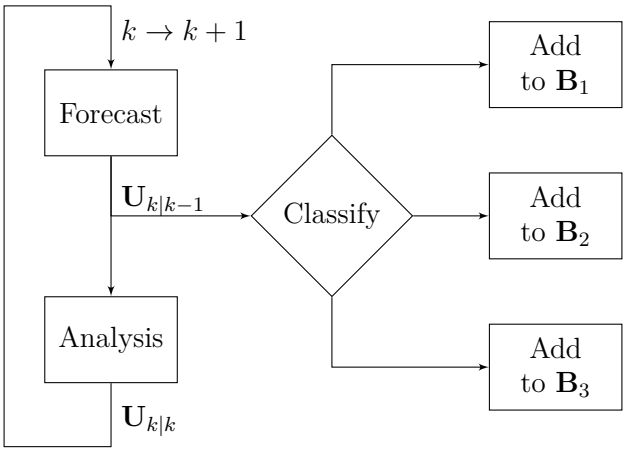


Figure 2.: A background-generating run

the background library. The ensembles are clustered into Voronoi cells, which minimizes variance between forecast ensembles within each background, providing a statistical regime classification given the forecast ensembles from the generating run.

3.2. Augmented Runs

The augmented run is an evaluation of the EnKF where the forecast ensemble is augmented with artificial members $\mathbf{U}_{\mathrm{art}_k}$ drawn from a steady distribution

$$\mathbf{U}_{\mathrm{art}_k} \sim \mathcal{N}\left(\overline{\mathbf{u}}_{k|k-1}, \mathbf{B}_\ell\right), \tag{10}$$

where the background covariance $\mathbf{B}_\ell$ is selected from a library $B = \{\mathbf{B}_\ell\}_{\ell=1}^{n_B}$ using a classification scheme as discussed in Section 3.3. We then form an augmented forecast ensemble matrix from the natural and artificial members,

$$\mathbf{U}_{k|k-1} \leftarrow \left[\begin{array}{cc} \mathbf{U}_{k|k-1} & \mathbf{U}_{\mathrm{art}_k} \end{array}\right]. \tag{11}$$

This augmented ensemble is used in place of the natural ensemble to calculate the Kalman gain in the analysis step. Assuming the artificial ensemble $\mathbf{U}_{\mathrm{art}_k}$ is appropriately chosen, the augmentation allows the natural ensemble $\mathbf{U}_{k|k-1}$ to be smaller than that required to statistically resolve the forecast covariance, reducing computational cost. The structure of an augmented run with $n_B = 3$ backgrounds is illustrated in Figure 3.

Although the ability to update the background distribution library using forecast ensembles from an augmented run on-the-fly would be attractive in practice, we do not consider this here. 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 training data that may be obtained from an augmented run is not independent of the information already present in the background library. This would introduce a potential source of bias in our library of background covariances. For our current implementation, generating runs and augmented runs are therefore mutually exclusive.

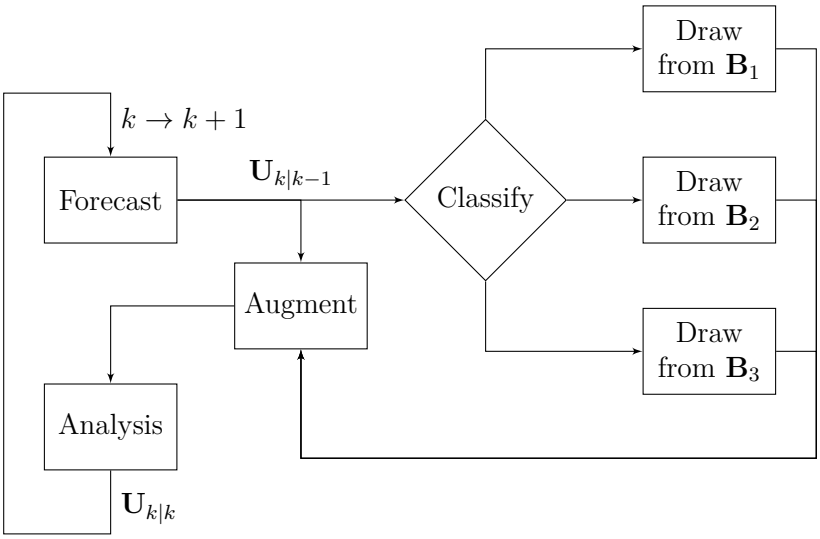


Figure 3.: An augmented run

3.3. Background Classification

The *classification scheme* takes a deviation forecast ensemble $\widetilde{\mathbf{U}}_{k|k-1}$, which has covariance $\mathbf{C}_{k|k-1}$, and finds its best match from a library of backgrounds B. A classification scheme is required in two scenarios. In the generating run, each forecast ensemble must be sorted among a library of backgrounds. The ensemble is then appended to the matching background ensemble. In augmented runs, the (statistically under-resolved) forecast ensemble must be matched to the appropriate statistically-resolved background. Artificial members are then sampled from this background.

This poses a statistical regime classification problem, which depends on two aspects of the generating or augmented run. The first is the particular trajectory of the ground truth, which determines the actual underlying flow regimes. The second is the uncertainty in the forecast ensemble, which depends on the initial conditions and the observation model. Because of this, two forecast ensembles that correspond to the same underlying ground truth may still have very different distributions.

We use the Wasserstein metric as the measure of distance between distributions. The Wasserstein metric is chosen for two reasons: (i) its stability with low-rank distributions, unlike e.g. the popular Kullback-Leibler divergence; (ii) it can be computed efficiently when the covariances admit low-rank decompositions. The metric has also been used successfully in the past in reacting flow regime classification [34]. When applied to two zero-mean Gaussian distributions with covariances $\mathbf{B}_\ell$ and $\mathbf{C}_{k|k-1}$, the (square of the) Wasserstein distance is

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

In practice, we do not explicitly form the matrices and instead leverage the low-rank decompositions of $\mathbf{C}_{k|k-1}$ and $\mathbf{B}_\ell$ to efficiently evaluate the distance. We classify $\mathbf{C}_{k|k-1}$ by matching it to the background $\mathbf{B}_\ell$ with the smallest Wasserstein distance. This is illustrated in Figure 4 for an example library of $n_B = 3$ backgrounds.

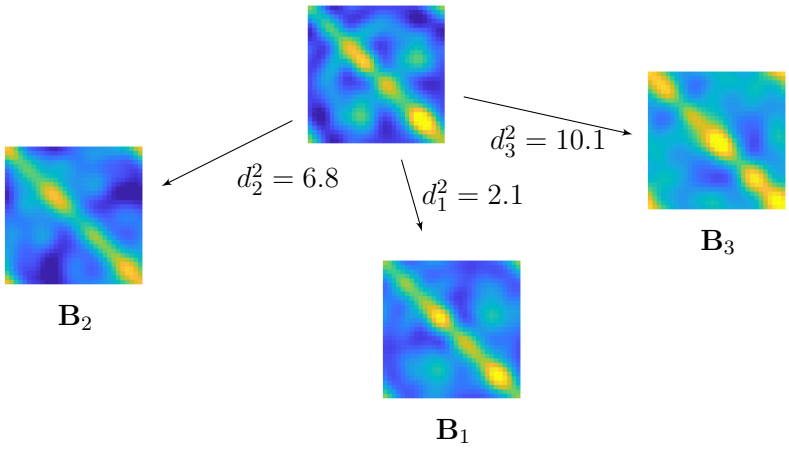


Figure 4.: Finding the Wasserstein distance d_ℓ^2 to each background covariance in B

3.4. Training Strategy

The *training strategy* is how we design the generating run or runs such that the resulting background library captures the relevant statistical behaviour for subsequent augmented runs. We wish to choose the generating runs such that they produce all forecast covariances necessary for the planned augmented runs.

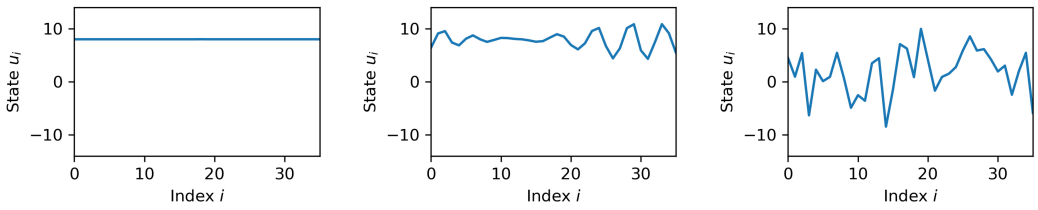
As discussed in Section 3.3, the background ensemble and the forecast ensemble must match in two characteristics: the physical state (i.e., mean) and the uncertainty (i.e., covariance). A generating run whose state estimate is converged about the true solution for most of its simulation time mostly captures only the low-uncertainty statistical regimes. For an augmented run whose time domain is wholly contained within the generating run's time domain, the large-uncertainty augmentation needed early in the filter run is unlikely to be well-matched by any background in the library. A preferable training strategy may be to perform many shorter generating runs, capturing a range of uncertainties for each time instance.

Covariance augmentation is a machine learning technique, and therefore the training strategy is critical to the method's performance. Computational cost is high for training, so we require an intelligent sampling method that minimizes the up-front cost and makes augmentation more useful for smaller multiple-query problems. We however have not performed a detailed study of training strategy in this work, but recommend it as a focus for future research. The strategies that we use are discussed in Sections 4 and 5.

4. Model Test Problems

In this section we apply our DA techniques to two model OSSEs. We follow the procedure detailed in Section 2.1:

(1) Evaluate the process model (1) for the full time interval under consideration to generate the ground truth reference solution $\mathbf{u}_k,\ k \in \{1, \ldots, n_t\}$, against which we compare the ETKF estimate.
(2) Apply the observation model (2) to the ground truth to obtain the pseudo-observations $\mathbf{y}_k^{\text{obs}},\ k \in \{1, \ldots, n_t\}$.
(3) Generate an initial ensemble as per (3) in Section 2.2.



(a) $t = 0$ (b) $t = 1$ (c) $t = 2$

Figure 5.: Reference solution for L96 problem

(4) Apply the ETKF with any covariance modification techniques of interest.
(5) Assess the error and uncertainty in the ETKF state estimate.

Because the process and observation models, which are used to generate the reference solution and observation data, are also used in the forecast and analysis steps of the filter, there is no inherent model deficiency. The error and uncertainty in the state estimate is solely due to the uncertainty in the initial condition.

4.1. Problem Definitions

The two model problems, L96 and KS, are used to assess the overall statistical behaviour of different methods. The problems are chosen because they are computationally inexpensive while still exhibiting behaviour of interest, including chaotic behaviour (for both) and either weak (for L96) or strong (for KS) correlation over long distances. As discussed in Section 1, the KS equation is derived from a simplification of reacting-flow equations [26–28], serving as a stepping stone between simple model problems and full reacting flow simulation.

4.1.1. Lorenz 96 Model

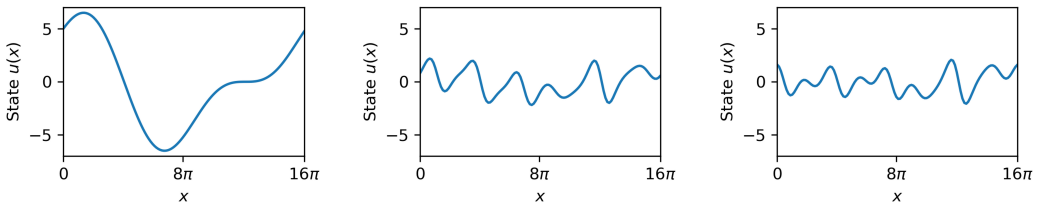
The L96 model is a system of first-order nonlinear ODEs

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = (u_{i+1} - u_{i-2})\, u_{i-1} - u_i + F, \qquad i = 1, \ldots, n_u, \tag{13}$$

which is a (significantly) simplified atmospheric model [25] that is commonly used to assess DA methods [33]. The three terms on the right hand side of (13) model advection, dissipation, and external forcing respectively. Following Lorenz [25], we use the model parameters $n_u = 36$ and $F = 8$, which is known to produce chaotic behaviour. The time interval is $[0, 2]$. We enforce a periodic boundary condition and the initial condition of $u_{i,k=0} = F$ for all i except $n_u/2$, where $u_{i=n_u/2,k=0} = F + 0.01$. We solve the ODEs using an adaptive-order backward differentiation formula (BDF) method with a relative error tolerance of 0.1%. The reference solution to this problem is shown in Figure 5.

We use a linear observation model $\mathbf{H}$ which consists of $n_y = 12$ observation nodes clustered into four evenly spaced groups of three; i.e., the observed elements of the state vector are at indices $i \in \{1, 2, 3, 10, 11, 12, 19, 20, 21, 28, 29, 30\}$. The analysis update therefore relies on local cross-correlation terms to correct the state estimate in the regions between observation nodes. We use a short observation period of $\Delta t = 0.003$.

The state in L96 is only weakly correlated over long distances, but strongly correlated



(a) $t = 0$ (b) $t = 30$ (c) $t = 60$

Figure 6.: Reference solution for KS problem

locally. This implies that the forecast covariance will likely be diagonally-dominant, and the eigenvalues decay slowly. A covariance estimate based on a small ensemble size is therefore missing significant eigenmodes.

4.1.2. KS Equation

The KS equation is a fourth-order nonlinear PDE

$$\frac{\partial u}{\partial t} = -\frac{\partial^4 u}{\partial x^4} - \frac{\partial^2 u}{\partial x^2} - u\frac{\partial u}{\partial x}, \tag{14}$$

which, as discussed in Section 1, models instabilities in laminar flame fronts and exhibits chaotic behaviour [26–28]. The fourth-derivative term provides damping in small scales, the second-derivative term destabilizes large scales, and the nonlinear transport term transfers energy between the two scales. We consider the time interval $[0, 60]$ and the spatial domain $[0, 16\pi]$. We enforce a periodic boundary condition and the initial condition is

$$u(x, t = 0) = 5\cos\left(\frac{x}{8}\right)\left(1 + \sin\left(\frac{x}{8}\right)\right). \tag{15}$$

We discretize the KS equation in space using second-order centered difference approximations on the grid with $n_u = 128$ nodes, so that $\Delta x = \pi/8$. The semi-discrete solution is given by $\mathbf{u}_i(t)$, $i = 1, \ldots, 128$. We then solve the semi-discrete equation using an adaptive-order BDF method, with a relative error tolerance of 0.1%.

As with L96, a linear observation model is used with $n_y = 12$ observation nodes clustered into four evenly spaced groups of three; i.e., the observed elements of the state vector are at indices $i \in \{1, 2, 3, 33, 34, 35, 65, 66, 67, 97, 98, 99\}$. The observation period is $\Delta t = 0.3$.

The ground truth reference solution is presented in Figure 6. After an initial transient period, the solution settles into a (statistically) 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. This leads to significant long-distance correlation over the state.

4.2. Assessment Procedure

We use two metrics to assess the filter performance: error and uncertainty. In problems with a state that is a discinuous function over a domain Ω, the error over time ϵ_k, $k \in$

$[1, n_t]$, is calculated by taking the L^2 norm over the domain Ω of the error between the ensemble mean and the ground truth solution,

$$\epsilon_k = \sqrt{\frac{\int_\Omega \left(u_k - \overline{u}_{k|k}\right)^2 dx}{\int_\Omega u_k^2 dx}}, \tag{16}$$

and the uncertainty $\sigma_k,\ k \in [1, n_t]$, is defined analogously in terms of the variance in the distribution,

$$\sigma_k = \sqrt{\frac{\int_\Omega \operatorname{var}\left(u_{k|k}\right) dx}{\int_\Omega u_k^2 dx}}. \tag{17}$$

In practice, given finite state vector $\mathbf{u}$ at grid points, we approximate the integrals with a quadrature rule. We take the error and uncertainty values at the final time step, ϵ_{n_t} and σ_{n_t} respectively, to assess the DA performance. When the ensemble size is sufficiently large, we can consider an ensemble filter to be 'well-resolved.' We identify this as the ensemble size where the error and uncertainty have reached their respective asymptotic limits.

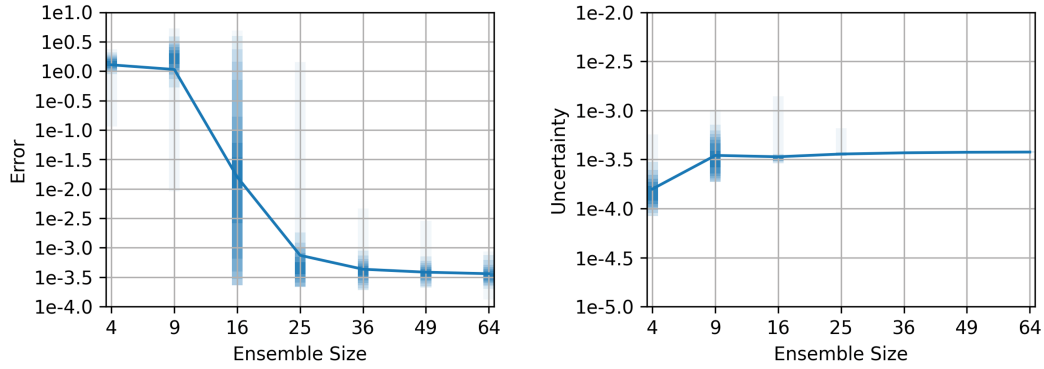
4.3. Baseline Performance

To assess the baseline performance of an unaugmented ETKF, we apply the ETKF to each of the model problems described in Section 4.1. We use a wide range of ensemble sizes appropriate to each model problem to characterize the ETKF performance and to find the required ensemble size to reach the asymptotic limits for error and uncertainty.

We take a moment now to discuss the figures used to present the results. An example is Figure 7, which shows the probability distributions $p(\epsilon|n_{\text{en}})$ and $p(\sigma|n_{\text{en}})$ over ensemble size n_{en}. These are empirical distributions constructed from repeated evaluations of the ETKF. The trace in each plot represents the mean of the distribution at a given ensemble size. The shaded colourbars represent the probability density function, as found by kernel density estimation. The x-axis of these plots is scaled quadratically; this is because estimating the mean of a normal distribution using the sample mean has a standard error that scales as ${n_{\text{en}}}^{-1/2}$.

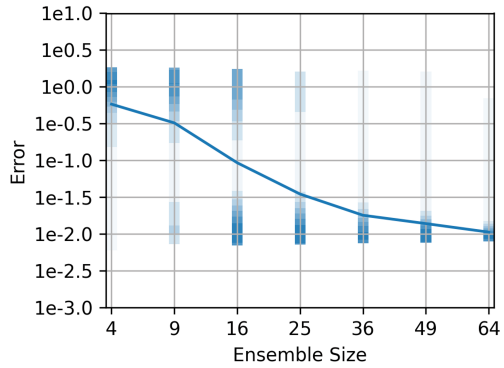
The baseline results for L96 are shown in Figures 7a and 7b. A large ensemble size relative to the state size is required to attain the asymptotic behaviour, which suggests that the eigenvalues of the covariance decay slowly. The uncertainty is underestimated only at the very small ensemble size of 4, and the mean uncertainty very quickly reaches the asymptotic value of $10^{-3.4}$ around ensemble size 9. At this ensemble size the error only begins to decrease, discinuing until it reaches the same limiting value of $10^{-3.4}$ around ensemble size 36. This shows that resolving the distribution requires much more than spreading the ensemble members, suggesting that inflation is not appropriate. For L96, we consider the ETKF 'well-resolved' at ensemble size 64.

The baseline results for the KS equation are shown in Figures 7c and 7d. Uncertainty in this case maintains a relatively consistent mean of around $10^{-3.25}$, though the spread converges with larger ensemble size. As with L96, the error mean decreases with increasing ensemble size; however, unlike L96, the distribution is almost exclusively bimodal. The same bimodal distribution is not present in the uncertainty, suggesting

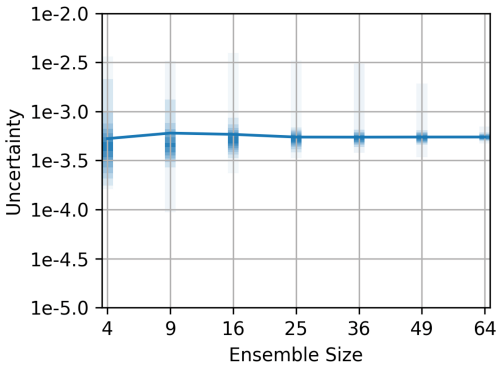


(a) L96 Error

(b) L96 Uncertainty



(c) KS Error



(d) KS Uncertainty

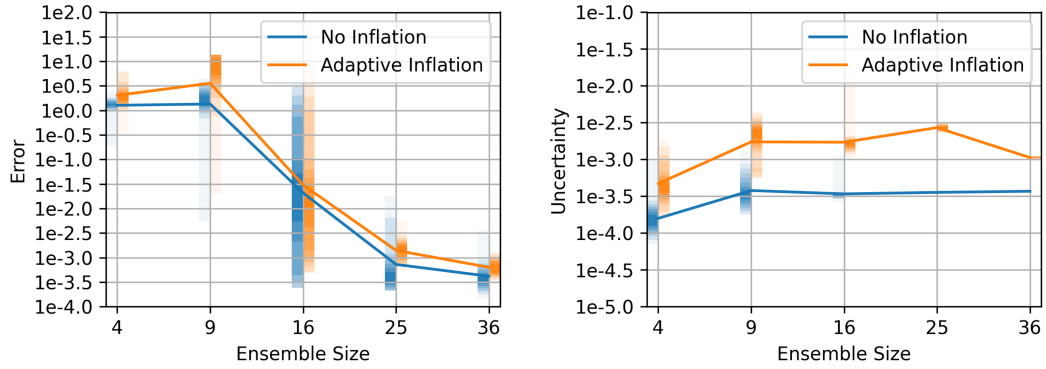
Figure 7.: Distributions of error and uncertainty for the ETKF applied to model problems

that the ensemble (falsely) converges about its mean regardless of whether it is tracking the true solution. A larger ensemble size allows it to more reliably track the true solution; this shift is mostly made between ensemble size 25 and 36. The performance discinues to improve with larger ensemble size, and the spread in the converged mode narrows as tracking becomes more reliable. The error approaches an asymptotic value of 10^{-2} around ensemble size 64, which is the ensemble size where we consider the ETKF 'well-resolved' for KS.

4.4. Inflation

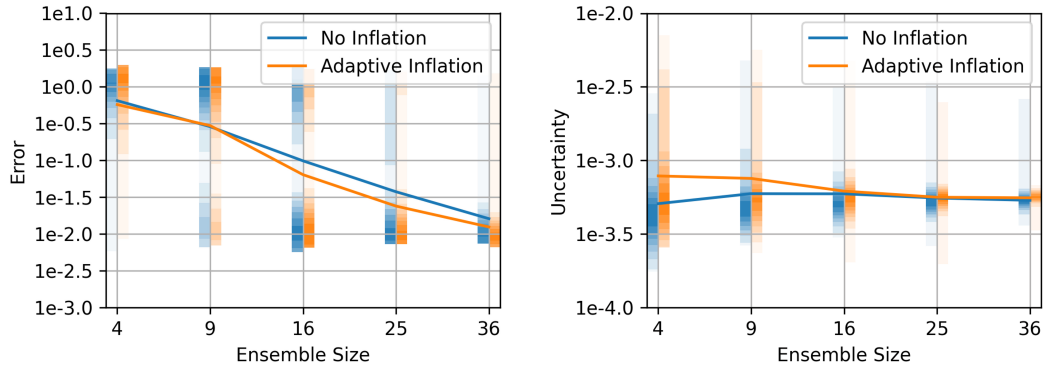
The results for adaptive covariance inflation of Anderson [11] are presented in Figure 8. We saw in the baseline results in Figure 7 that the uncertainty approaches its large-ensemble limit very quickly in both model problems, therefore inflation can only overestimate the forecast covariance, which we see in Figures 8b and 8d. Consequently, the effect of inflation on error in L96 and KS is limited. It is clear from the uncertainty plots that underestimated covariance (i.e., underestimated mode strength) is not a significant concern for most ensemble sizes in these model problems, therefore inflation is not an effective method to improve performance.

We however note that inflation can be an effective technique for L96 in some scenarios.



(a) L96 Error

(b) L96 Uncertainty



(c) KS Error

(d) KS Uncertainty

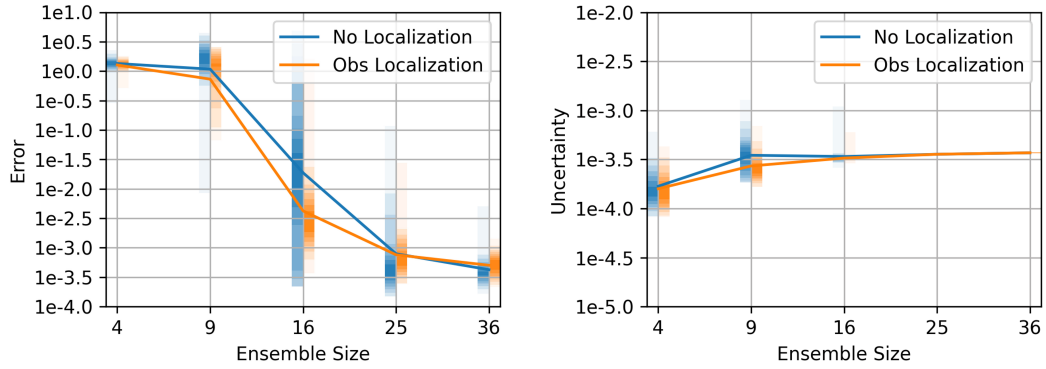
Figure 8.: Distributions of error and uncertainty for the ETKF with inflation applied to model problems

An example is Ahmed *et al.* [35], who use a denser observation model of $n_y = 18$ evenly-spaced nodes and a fixed, modest inflation factor of 1.04. With this observation model, each state element is either an observed node or adjacent to an observed node. This means that for each state element, the analysis update uses local observations, reducing the dependence on potentially under-resolved cross-correlation terms in the covariance. An inflated forecast covariance is therefore less likely to cause the state estimate to diverge due to spurious correlations, and this is reflected in the improved performance.

4.5. Localization

The results for observation localization of Greybush *et al.* [12] are presented in Figure 9. For each model problem, we consider a set of candidate localization length scales and report the results for the best case. The candidate length scales for L96 and the KS equation are $\{4, 5, \ldots, 16\}$ (in discrete node count) $\{2\pi, 4\pi, 8\pi, 16\pi, 32\pi\}$, respectively.

For L96, we use a length scale of 7 nodes. Localization provides a clear improvement over the unaugmented filter for small ensemble sizes. This is consistent with the known localized EnKF results for L96 [36]. Figure 9a shows that the greatest improvement to error happens around ensemble size 16, where localization decreases the error spread from four orders of magnitude to one. This makes the filter perform more reliably in



(a) L96 Error

(b) L96 Uncertainty

(c) KS Error

(d) KS Uncertainty

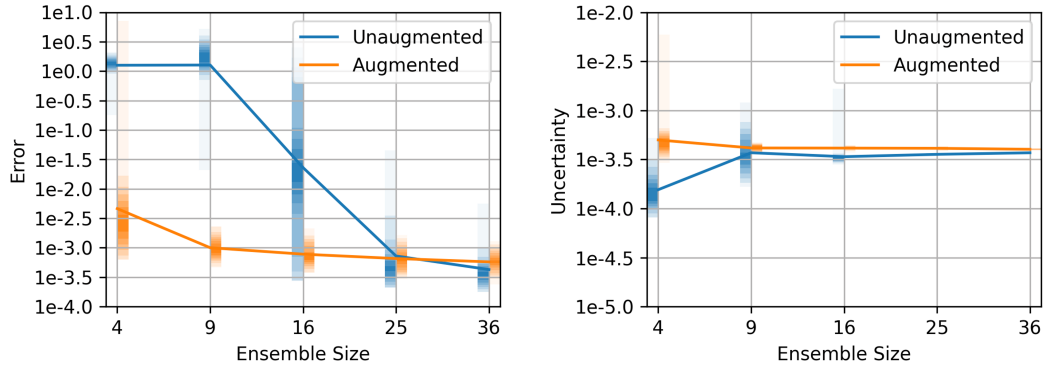
Figure 9.: Distributions of error and uncertainty for the ETKF with localization applied to model problems

cases where ensemble is only moderately undersized.

For the KS equation, we observe no improvements for all length scales considered. We recall from Figure 1 and the discussion in Section 4.1.2 that there are significant cross-correlation terms, which must be accurately resolved for effective analysis update because the observation data from KS is extremely sparse. However, localization suppresses these long-distance cross correlations, compromises performance for any effective length scale, and underestimates uncertainty. The results in Figures 9c and 9d use a large correlation length of $L = 16\pi$ to minimize localization's effect on performance. All smaller correlation lengths lead to larger error.

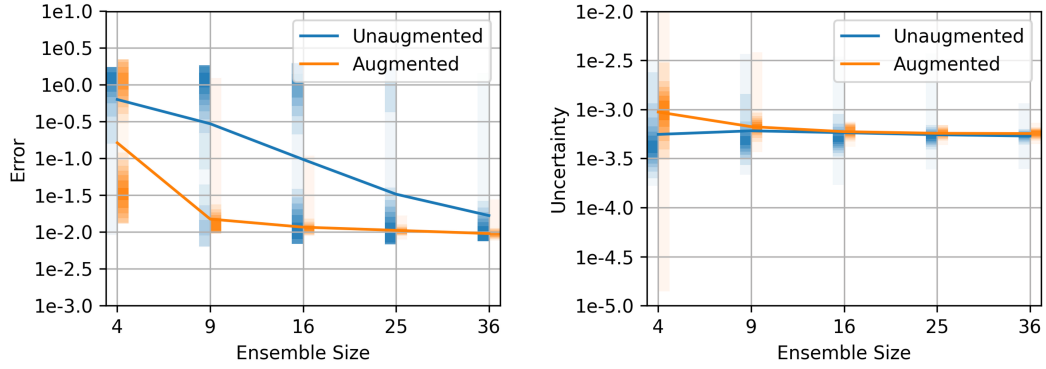
4.6. Augmentation

We consider three different use cases for covariance augmentation. 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 generalizability of the background library outside of the training data. The third is a scenario where the model parameters have changed between the generating and augmented runs, testing the generalizability of the background library to slightly different governing equations. For all three model problems, we produce $n_B = 50$ backgrounds



(a) L96 Error

(b) L96 Uncertainty



(c) KS Error

(d) KS Uncertainty

Figure 10.: Distributions of error and uncertainty for the ETKF with augmentation applied to model problems in the reproduction scenario

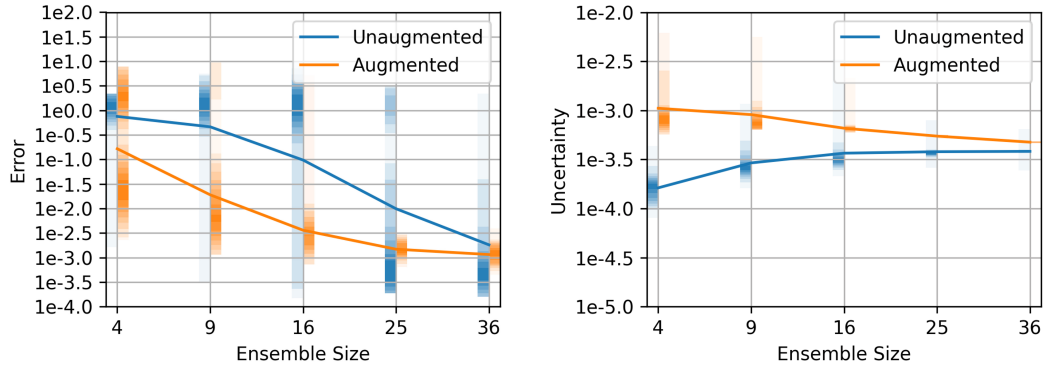
from two generating runs, each at the respective well-resolved ensemble size.

Note that in all augmentation results figures, the x-axis refers only to the natural ensemble size. All augmented runs use an augmented ensemble size of 64, as this is where both model problems were shown to be well-resolved in Section 4.3. The number of artificial members drawn from the background distributions therefore varies with the natural ensemble size, ensuring that the total remains 64.

4.6.1. Reproduction Scenario

The reproduction scenario represents the best-case scenario for covariance augmentation, as the training data and test data match as closely as possible. However, it does not reflect a practical use case, as it is necessarily more costly than simply using a well-resolved ETKF. Nonetheless, it is useful to quantify the best-case performance. The results are shown in Figure 10.

Because we know that there exist backgrounds in the library B that correspond almost perfectly with this trajectory, the test is ultimately of the classification scheme and the level of clustering, i.e., the background library size n_B relative to the number of forecast ensembles produced by generating runs. As expected, the error and uncertainty rapidly approach their respective asymptotic limits. The limiting factor determining ensemble size is no longer whether the covariance is well-resolved, but whether the clas-



(a) L96 Error

(b) L96 Uncertainty

(c) KS Error

(d) KS Uncertainty

Figure 11.: Distributions of error and uncertainty for the ETKF with augmentation applied to model problems in the different trajectory scenario

sification scheme can select the appropriate background covariance from the library B. For both model problems, the error mean and spread achieved with 9 natural ensemble members using the covariance-augmented ETKF is similar to that achieved with 36 members using the unaugmented ETKF, a fourfold reduction in natural ensemble size. If the cost to evaluate the Wasserstein distances (12) for n_B backgrounds to classify the covariance is negligible compared to the cost to evaluate process models, then this implies a fourfold reduction in computational cost. However, since we chose our model problems for their ease of computation, the cost to evaluate the process models is not so overwhelming.

4.6.2. Different Trajectory

A more practical test of augmentation is where the training data and the test data correspond to different model trajectories. The process model remains the same, but the generating run and the augmented run have different initial conditions. By observing the improvement obtained from backgrounds generated using a different trajectory, this test examines the level of steady statistical information present in the backgrounds. The results are shown in Figure 11.

Augmentation can lead to improved performance if there is steady, persistent statistical information in the background library that is relevant to most statistical regimes.

Whereas an underlying true covariance may have dynamic elements not present in the backgrounds, the overall covariance estimate may still be improved by steady-state characteristics that are present. A stronger condition for improved performance is whether the background library is *comprehensive*: i.e., any possible natural forecast ensemble has an appropriate corresponding background. In most cases, the construction of a comprehensive library requires a well-chosen training strategy.

This particular test uses background libraries of size $n_B = 50$ generated from the respective model time domains specified in Section 4.1. The augmented runs then start with an offset of half a simulation period from where the generating runs start, and then proceed for the same length of time. For example, with L96 the time domain is $[0, 2]$ for the generating run and $[1, 3]$ for the augmented run.

Using different simulation time windows results in the generating and augmented runs exhibiting different behaviors. Most notably, Figures 5 and 6 show that the state goes through the initial transient, where the state develops from initial conditions to its statistically steady behaviour, for a significant portion of the simulation period. In this different trajectory scenario, the generating run and consequently the background library include this initial transient, while the augmented run does not. Many backgrounds in the library are therefore superfluous to this augmented run. Note that, due to the difference in the initial time, the unaugmented performance in Figure 11 differs from the original baseline results in Figure 7.

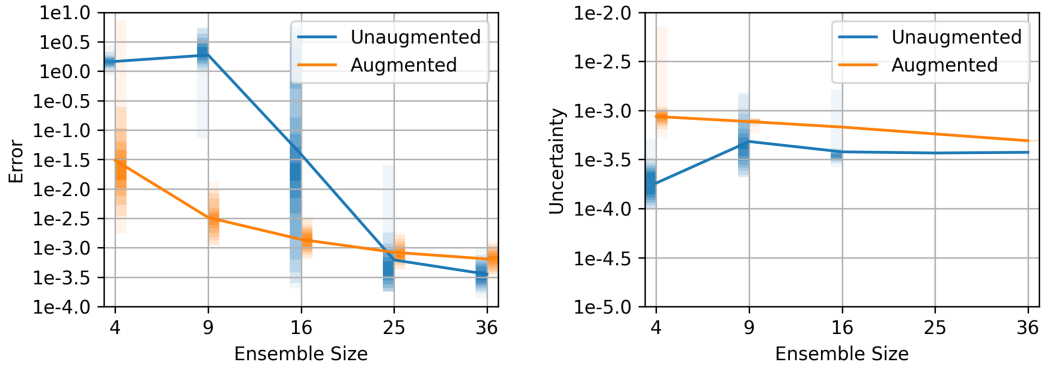
For L96, there is significant difference between the backgrounds collected from $t \in [0, 2]$ and necessary for $t \in [1, 3]$. Error in Figure 11a is highly bimodal at the smallest ensemble size, but the bimodality diminishes quickly as the ensemble size increases. The error spread remains wide for the unaugmented filter across all ensemble sizes, whereas the augmented filter is much more consistent, even for the ensemble sizes for which the mean performance is comparable. The clearest example is at natural ensemble size 16 with the augmented filter and 36 with the unaugmented filter; the augmented filter's error stays within an order of magnitude and the unaugmented filter's error spreads across almost three orders of magnitude.

For KS, the unaugmented performance over $t \in [30, 90]$ is significantly better than over $t \in [0, 60]$. Augmentation however still demonstrates improvement. Though both mean errors in Figure 11c are similar, the error spread in the unaugmented filter is much wider than in the augmented one. The augmented filter converges tightly on the asymptotic error limit by natural ensemble size 9, whereas the unaugmented filter does not until ensemble size 25–36, representing a 3–4× reduction in natural ensemble size. In Figure 11d, we see that the unaugmented filter underestimates uncertainty at most ensemble sizes, whereas the augmented filter again converges quickly to the asymptotic limit.

4.6.3. Parameter Variation

We finally assess the performance of an augmented ETKF when the model parameters differ between the generating runs and augmented runs. For L96, the parameter most readily modified is the forcing term F, which drives the chaotic motion. We set the forcing term to $F = 8$ for generating runs and $F = 10$ for augmented runs, which allows us to test the generalizability of augmentation when the process model is more chaotic than was anticipated when generating the backgrounds. The results are shown in Figure 12.

Figure 12 shows that the background library is robust to changes in the forcing term. Except for very small ensemble sizes, the augmented filter consistently converges to the



(a) Error (b) Uncertainty

Figure 12.: Distributions of error and uncertainty for the ETKF with augmentation applied to L96 in the parameter variation scenario

true solution, suggesting that, at least for this case, the background library is sufficiently generalizable such that it remains effective for different values of F. The augmented filter with L96 achieves stable, converged performance at natural ensemble size 9 that the unaugmented filter does not achieve until an ensemble size of approximately 20.

5. Reacting Flow Simulation

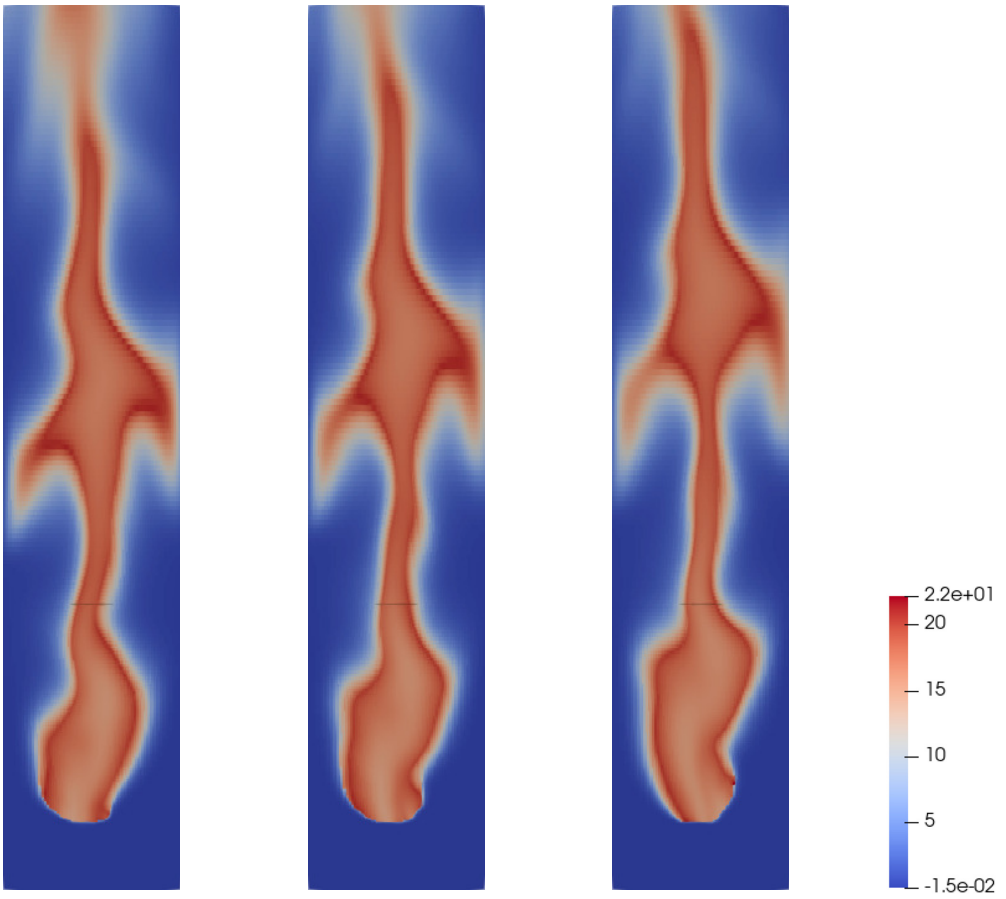
In this section we apply our DA techniques to a reacting flow problem using the same procedure detailed in Section 4. Again, the use of a noiseless process model and pseudo-observation data implies no inherent deficiency in the process or observation models.

5.1. Problem Definition

We consider data assimilation of two-dimensional non-premixed propane-oxygen diffusion flame at low Reynolds number using simultaneous PIV (velocity) and temperature measurements. We first discuss the process model, which consists of an LES model, and then discuss the observation model, which consists of simulated PIV and temperature measurements.

5.1.1. Process Model

The reacting flow is modelled with Raptor, a finite-volume solver developed by Oefelein [37]. The reacting flow model considers 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). The model considers the simplified process $2C_3H_8+3O_2 \rightarrow 6CO+8H_2O$ and $2CO + O_2 \rightarrow 2CO_2$ to simulate the combustion chemistry. The model hence works with nine flow variables: four fluid dynamic variables (pressure, two velocity components, and temperature) and five explicitly stored chemical mass species fractions (the last, N_2, is stored implicitly as 1 minus all other species). The Reynolds number is 12000, using the width of the domain as the length scale and the inflow speed as the velocity scale. The flow is unsteady but not turbulent, as it is not at a sufficiently high Reynolds number to have a well-established inertial range. The inflow Mach number is



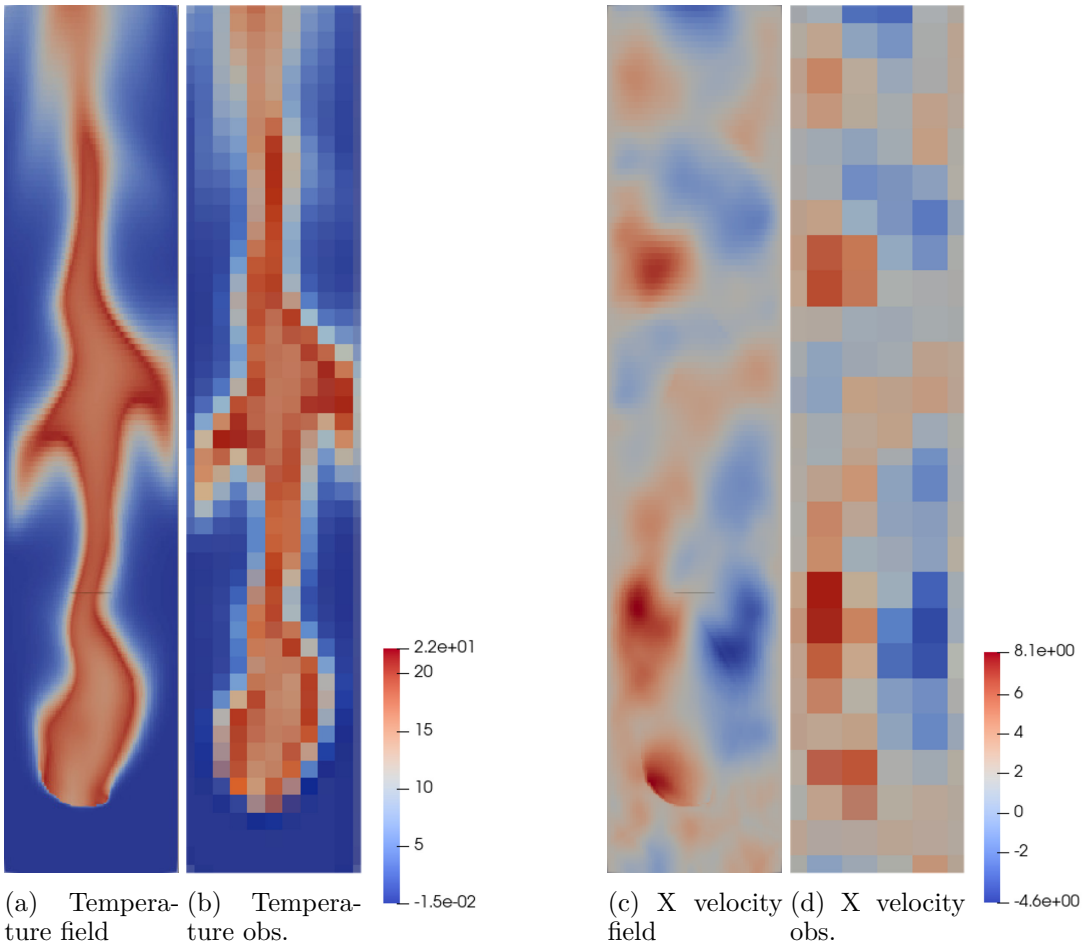
(a) $t = 0.0$ (b) $t = 0.5$ (c) $t = 1.0$

Figure 13.: Diffusion flame simulation. The plot is of non-dimensionalized temperature. The direction of flow is from bottom to top.

negligibly small at 0.003, implying incompressible flow. We impose a constant pressure boundary condition at the top, slip conditions on the sides, and a uniform inflow velocity boundary condition with a narrow propane inlet in the centre, one tenth of the domain width, at the bottom. The time interval is $[0, 0.5]$. Unlike the model problems in Section 4, there is no initial transient; the initial condition is a fully-developed flow. Snapshots of the ground-truth solution are shown in Figure 13.

5.1.2. Observation Model

For the observation model, we choose to replicate the effects of a PIV-type 2D velocity measurement and 2D temperature measurement. Although 2D temperature measurements are challenging in practice, they can potentially be accessed by, for example, Rayleigh scattering [38] or PLIF thermometry [39]. The pressure and chemical species mass fractions are not measured. The velocity measurements are obtained by averaging the ground-truth velocity field in each of 26×6 grid of PIV interrogation windows and then adding zero-mean Gaussian noise. Adjacent PIV interrogation windows have 50% overlap. The temperature measurements are obtained by sampling the ground-truth temperature field at each of 51×11 grid points and then adding zero-mean Gaussian noise. Figure 14 shows representative observed data, which are spatially under-resolved.



(a) Temperature field (b) Temperature obs. (c) X velocity field (d) X velocity obs.

Figure 14.: Spatial resolution of observation data in velocity and temperature

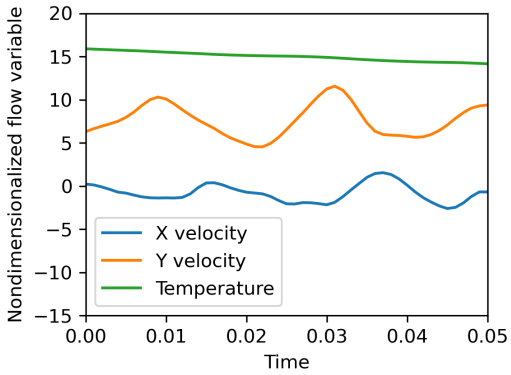
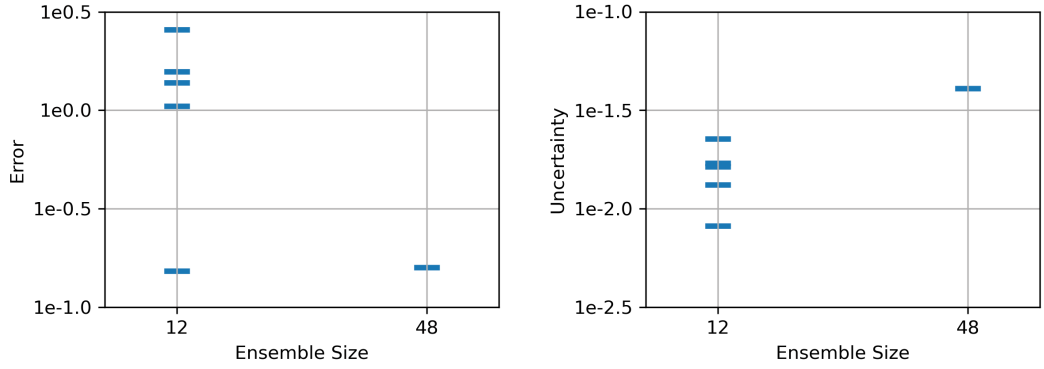


Figure 15.: Time history of nondimensionalized velocity and temperature over one observation period at a point on the edge of the flame

For non-dimensionalized flow variables of order 1, the PIV and temperature noise yield signal-to-noise ratios of roughly 16 and 8, respectively. That is, the Gaussian noise term has a standard deviation of 0.06 for PIV measurements and 0.12 for temperature mea-



(a) Error (b) Uncertainty

Figure 16.: Distributions of error and uncertainty for the ETKF applied to reacting flow

surements. We choose this limited, sparse, and noisy set of observations to assess DA's ability (i) to correct unobserved flow variables exclusively through cross-correlation terms in the forecast covariance, (ii) to estimate the small-scale flow structures not captured by the data, and (iii) to accommodate noisy measurements.

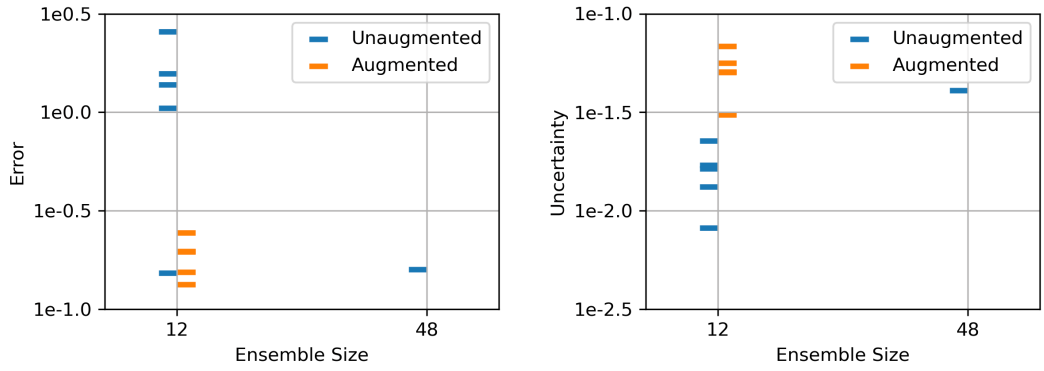
We choose an observation period of $\Delta t = 0.05$ in non-dimensionalized time. This provides observation data that temporally resolves the temperature field but not the velocity field. This is shown in Figure 15, where we plot the velocity and temperature over time for a point on the edge of the flame. The figure considers only one observation period, showing that the observation model leaves the velocity field temporally underresolved. Much like our choice to use a spatially sparse observation model, this allows us to assess DA's ability to estimate velocity field behaviour not captured by the data.

5.2. Baseline Performance

We define two baseline scenarios: a large-ensemble case with ensemble size 48, which provides the best-case performance, and a small-ensemble case with ensemble size 12, which underresolves the covariance and leads to poor performance. To characterize the typical performance of the ETKF in the small-ensemble case, we consider five different randomly chosen initial ensembles.[1] The large-ensemble case yields much more consistent performance, so only one initial ensemble is used.

Figure 16 shows the error and uncertainty over ensemble size as a scatter plot. In the large-ensemble case (i.e., $n_{\text{en}} = 48$), the ETKF achieves an error of roughly $10^{-0.8}$ and uncertainty of roughly $10^{-1.4}$; for nondimensionalized flow variables of order 1, this corresponds to approximately 16% error. This is lower than the level of velocity fluctuation shown in Figure 15, supporting that the ETKF achieves super-temporal resolution. This is also lower than the error in our low-resolution observations, which we observed to be roughly 30% to 40%, supporting that the ETKF also achieves super-spatial resolution. This error is achieved despite the pressure and species fractions being unobserved, relying on correlation with the observed velocity and temperature in order to apply the analysis updates. In the small-ensemble case (i.e., $n_{\text{en}} = 12$), there is a

[1]Due to the much higher computational cost of the reacting-flow simulation, we are unable to perform the same detailed statistical analysis performed for model problems in Section 4; nevertheless, we assess the method using five cases with different random ensembles to characterize the typical performance.



(a) Error

(b) Uncertainty

Figure 17.: Distributions of error and uncertainty for the ETKF with augmentation applied to reacting flow in the reproduction scenario

wide variation in the error; the best-case error of $10^{-0.8}$ is comparable to the large-ensemble case, whereas the worse-case error is roughly $10^{0.4}$. A typical error is large and is between 10^{0} and $10^{0.5}$. Uncertainty is consistently underestimated compared to the large-ensemble case, even in the atypical small-error result. As with the model problems, the underestimated uncertainty (relative to the error) causes the filter to incorrectly trust the state estimates and to resist updating the states based on observations; as a result, the filter fails to reduce the error.

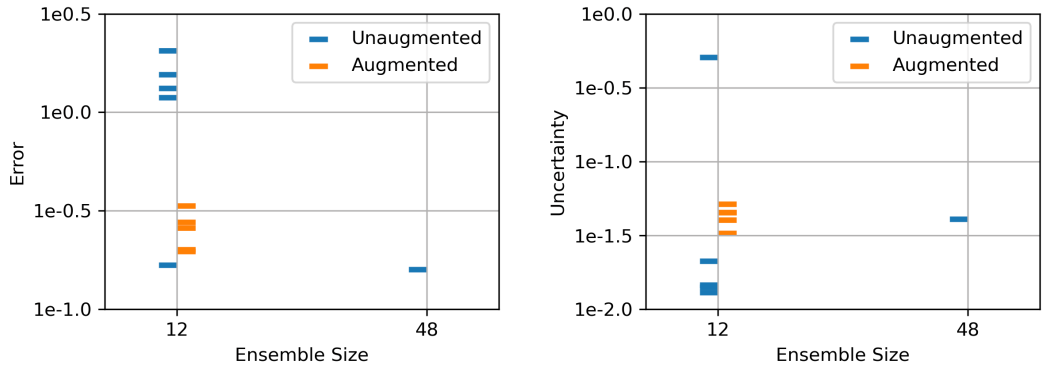
5.3. Augmentation

We consider two augmentation scenarios. We start with the reproduction scenario, where the model trajectory in the generating run is the same as that in the augmented run. We then consider a scenario where these trajectories differ, which tests the generalizability of the background library outside of the training data. For both the reproduction and the different trajectory scenarios, we produce $n_B = 50$ backgrounds from one generating run. The large-ensemble case in Section 5.2 serves as our generating run. We use the small-ensemble case as our augmented run, using the same five initial ensembles.

As in the augmentation results figures in Section 4, the x-axis refers only to the natural ensemble size. All augmented runs use an augmented ensemble size of 48 corresponding to the large-ensemble case. In an augmented small-ensemble run, there are therefore 36 artificial members and 12 natural members.

5.3.1. Reproduction Scenario

We first consider the reproduction scenario to illustrate the best-case scenario for covariance augmentation. Figure 17 shows the error and uncertainty for the augmented small-ensemble ETKF compared to the baseline results from Figure 16. The augmented results appear to strongly resemble the large-ensemble results. The mean error is decreased by an order of magnitude compared to the unaugmented small-ensemble results and the mean uncertainty raised by half an order of magnitude, bringing performance almost perfectly in line with the unaugmented large-ensemble case. With a natural ensemble size of 12 instead of 48, the augmented small-ensemble ETKF achieved this



(a) Error (b) Uncertainty

Figure 18.: Distributions of error and uncertainty for the ETKF with augmentation applied to reacting flow in the different trajectory scenario

error performance with a quarter of the large-ensemble ETKF's computational cost. Even with the more complicated test problem of a simulated diffusion flame, augmentation is robust enough to improve performance just as it did with the model problems in Section 4.

5.3.2. Different Trajectory

For the different trajectory scenario, we consider the same background library as the reproduction scenario, corresponding to the time interval $[0, 0.5]$. We perform an augmented run that is offset forward in time by two observation periods, i.e., the time interval $[0.1, 0.6]$. We recall from Figure 15 that this shift induces a significant change in the velocity field. Figure 18 shows the error and uncertainty for the augmented small-ensemble ETKF compared to new baseline results for this offset time interval. Again the augmented results strongly resemble the unaugmented large-ensemble results; the error is significantly reduced compared to the unaugmented small-ensemble ETKF and is only slightly higher than the reproduction scenario and large-ensemble results. The uncertainty has also increased toward the large-ensemble level. This matches our expectations for a slightly mismatched background library: although it does not reproduce the large-ensemble performance, the background library still contains relevant statistical information that improves performance. The result indicates that the proposed training strategy and classification schemes are effective in this more practical case with different trajectories.

6. Summary and Conclusions

In this work, we have developed and assessed a covariance augmentation technique for ensemble DA in multiple-query scenarios. In our proposed method, we first construct a library of background covariances from a number of generating runs. We subsequently invoke an augmented ETKF, which uses a small 'natural' ensemble of states augmented by the 'artificial' states drawn from the library. The key ingredients of the covariance augmentation technique are (i) a training strategy that identifies a comprehensive library of background covariances from which to construct the library and (ii) a classification

scheme that clusters closely related background covariances and identifies the appropriate cluster from which to draw the artificial states. We have assessed the approach using model problems and a reacting-flow problem. We have observed that the ensemble size required to classify an ensemble among statistical flow regimes is much smaller than the ensemble size required for an ensemble to resolve the covariance by itself, which allows us to use a much smaller natural ensemble size in the augmented ETKF. We have also observed that the covariance library is generalizable in the sense that it can be used when the system trajectory or the governing equations are different from those used in the generating runs; even if the library is not perfect, the classification scheme is robust enough to choose the most appropriate background among imperfect options, and that background contains enough relevant information to improve performance significantly. In practice, we have observed that an augmented ETKF can use a natural ensemble size that is three to four times smaller than that of an unaugmented ETKF, resulting in a commensurate reduction in the computational cost.

The first major limitation to the current work is the lack of consideration given to training strategies. Though we tested our augmentation scheme using reacting flow in the reproduction and different trajectory scenarios, there are more cases of practical interest that warrant examination. The reacting flow case in Section 5 provides an example of statistically stationary behaviour, which would be an appropriate test problem if one wants to quantify 'comprehensiveness' of statistical models and assess training strategies for producing them; however, that was beyond the scope of the current work.

The second major limitation is the chosen OSSE setup. Because our process and observation models were consistent with the ground-truth models, we likely achieved better performance than we would achieve in real-world applications of DA. The lack of an error term in the process model led to lower error and uncertainty than would have been obtained in a setup that included one. Additionally, it meant that the only initial uncertainty in the state estimate came from the initial ensemble, which was in turn drawn from ground truth. As discussed in Section 2.2, some process models may fail to propagate the state forward in time if the initial condition is nonphysical. Our DA performance was therefore more stable than it otherwise may have been; however, the focus was on a straightforward comparison of error and uncertainty estimates for large-ensemble and small-ensemble filters. Although the effect of ensemble size on stability was not discussed, in a practical application of DA it may be significant. Finally, the fact that all initial ensembles were drawn from the ground truth means that, for augmentation in the reproduction scenario, the initial ensemble would likely be well-matched to at least one background. It is clear however that augmentation's performance does not rely on this convenience, as the different trajectory and parameter variation scenarios both show significant reductions in error. Future assessments of augmentation as an approach ought to relax the assumptions that limit the realism of our OSSEs; however, it is not readily apparent that augmentation relies on any of these simplifications to perform. We speculate therefore that even with flawed models in real-world scenarios, if an (unaugmented) large-ensemble DA performs well, then covariance augmentation should still provide similar performance using a reduced ensemble size.

A minor third limitation is the relative simplicity of our reacting flow case; it is possible that for more complicated flows of practical interest (e.g. large, turbulent flows), constructing a comprehensive background library is prohibitively expensive, negating any possible benefits of covariance augmentation. However, at least for the model problems and the reacting flow case considered, augmentation was not compromised by the relative size and complexity of the test problems.

Acknowledgements

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

Funding

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.

The authors report there are no competing interests to declare.

References

[1] J.W. Labahn, H. Wu, S.R. Harris, B. Coriton, J.H. Frank, and M. Ihme, *Ensemble Kalman filter for assimilating experimental data into large-eddy simulations of turbulent flows*, Flow, Turbulence and Combustion (2019), Available at `https://doi.org/10.1007/s10494-019-00093-1`.

[2] S. Barwey, M. Hassanaly, Q. An, V. Raman, and A. Steinberg, *Experimental data-based reduced-order model for analysis and prediction of flame transition in gas turbine combustors*, Combustion Theory and Modelling 23 (2019), pp. 994–1020, Available at `https://doi.org/10.1080/13647830.2019.1602286`.

[3] M. Kapteyn, K. Willcox, and D. Knezevic, *Toward predictive digital twins via component-based reduced-order models and interpretable machine learning*, in *AIAA Scitech 2020 Forum*, 01. 2020.

[4] F. Rabier and Z. Liu, *Variational data assimilation: theory and overview*, in *Seminar on recent developments in data assimilation for atmosphere and ocean, 8-12 September 2003*, Shinfield Park, Reading. European Centre for Medium-Range Weather Forecasts (ECMWF), ECMWF, 2003, pp. 29–44, Available at `https://www.ecmwf.int/node/11805`.

[5] R.N. Bannister, *A review of operational methods of variational and ensemble-variational data assimilation*, Quarterly Journal of the Royal Meteorological Society 143 (2017), pp. 607–633, Available at `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2982`.

[6] P. Houtekamer and F. Zhang, *Review of the ensemble Kalman filter for atmospheric data assimilation*, Monthly Weather Review 144 (2016), pp. 4489–4532.

[7] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, *Data assimilation in the geosciences: An overview of methods, issues, and perspectives*, WIREs Climate Change 9 (2018), p. e535, Available at `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.535`.

[8] T. Hamill, *Ensemble-based data assimilation.*, in *Seminar on predictability of weather and climate, 9-13 September 2002*, Shinfield Park, Reading. ECMWF, ECMWF, 2003, pp. 83–112, Available at `https://www.ecmwf.int/node/9756`.

[9] G. Evensen, *Data assimilation: the ensemble Kalman filter*, Springer, 2009.

[10] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, Journal of Geophysical Research: Oceans 99 (1994), pp. 10143–10162, Available at `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572`.

[11] J.L. Anderson, *An adaptive covariance inflation error correction algorithm for ensemble filters*, Tellus A 59 (2007), pp. 210–224, Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2006.00216.x.

[12] S.J. Greybush, E. Kalnay, T. Miyoshi, K. Ide, and B.R. Hunt, *Balance and ensemble Kalman filter localization techniques*, Monthly Weather Review 139 (2011), pp. 511 – 522, Available at https://journals.ametsoc.org/view/journals/mwre/139/2/2010mwr3328.1.xml.

[13] L. Lei, J. Anderson, and G. Romine, *Empirical localization functions for ensemble Kalman filter data assimilation in regions with and without precipitation*, Monthly Weather Review 143 (2015), p. 150522112937007.

[14] B. Wang, J. Liu, L. Liu, S. Xu, and W. Huang, *An approach to localization for ensemble-based data assimilation*, PLOS ONE 13 (2018), pp. 1–23, Available at https://doi.org/10.1371/journal.pone.0191088.

[15] B. Ménétrier, T. Montmerle, L. Berre, and Y. Michel, *Estimation and diagnosis of heterogeneous flow-dependent background-error covariances at the convective scale using either large or small ensembles*, Quarterly Journal of the Royal Meteorological Society 140 (2014), pp. 2050–2061, Available at https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2267.

[16] B. Ménétrier, T. Montmerle, Y. Michel, and L. Berre, *Linear filtering of sample covariances for ensemble-based data assimilation. Part II: Application to a convective-scale NWP model*, Monthly Weather Review 143 (2015), pp. 1644 – 1664, Available at https://journals.ametsoc.org/view/journals/mwre/143/5/mwr-d-14-00156.1.xml.

[17] J.R. Edwards, C. Patton, H. Mirgolbabaei, T. Wignall, and T. Echekki, *4D data assimilation for large eddy simulation of high speed turbulent combustion*, in *51st AIAA/SAE/ASEE Joint Propulsion Conference*, 07. 2015, Available at https://arc.aiaa.org/doi/abs/10.2514/6.2015-3836.

[18] J.R. Edwards, L. Luo, C. Patton, T. Wignall, and T. Echekki, *Improved 4D data assimilation for large-eddy simulation of high-speed turbulent combustion*, in *46th AIAA Fluid Dynamics Conference*, 06. 2016, Available at https://arc.aiaa.org/doi/abs/10.2514/6.2016-3957.

[19] X. Gao, Y. Wang, N. Overton, M. Zupanski, and X. Tu, *Properties of a modified ensemble Kalman filter algorithm for combustion application*, in *46th AIAA Fluid Dynamics Conference*. 2016, Available at https://arc.aiaa.org/doi/abs/10.2514/6.2016-3484.

[20] J. Gray, M. Lemke, J. Reiss, C. Paschereit, J. Sesterhenn, and J. Moeck, *A compact shock-focusing geometry for detonation initiation: Experiments and adjoint-based variational data assimilation*, Combustion and Flame 183 (2017), pp. 144–156, Available at https://www.sciencedirect.com/science/article/pii/S0010218017301062.

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

[22] H. Yu, T. Jaravel, M. Ihme, M.P. Juniper, and L. Magri, *Data assimilation and optimal calibration in nonlinear models of flame dynamics*, Journal of Engineering for Gas Turbines and Power 141 (2019), Available at https://doi.org/10.1115/1.4044378, 121010.

[23] H. Yu, M.P. Juniper, and L. Magri, *A data-driven kinematic model of a ducted premixed flame*, Proceedings of the Combustion Institute (2020), Available at https://www.sciencedirect.com/science/article/pii/S1540748920302170.

[24] R. Yondo, E. Andrés, and E. Valero, *A review on design of experiments and surrogate models in aircraft real-time and many-query aerodynamic analyses*, Progress in Aerospace Sciences 96 (2018), pp. 23–61, Available at https://www.sciencedirect.com/science/article/pii/S0376042117300611.

[25] E.N. Lorenz, *Predictability: a problem partly solved*, in *Seminar on predictability, 4-8 September 1995*, Vol. 1, Shinfield Park, Reading. ECMWF, ECMWF, 1995, pp. 1–18, Available at https://www.ecmwf.int/node/10829.

[26] G.I. Sivashinsky, *Nonlinear analysis of hydrodynamic instability in laminar flames — I.*

derivation of basic equations, Acta Astronautica 4 (1977), pp. 1177 – 1206, Available at http://www.sciencedirect.com/science/article/pii/0094576577900960.

[27] Y. Kuramoto, *Diffusion-induced chaos in reaction systems*, Progress of Theoretical Physics Supplement 64 (1978), pp. 346–367, Available at https://doi.org/10.1143/PTPS.64.346.

[28] G.I. Sivashinsky, *On flame propagation under conditions of stoichiometry*, SIAM Journal on Applied Mathematics 39 (1980), pp. 67–82, Available at http://www.jstor.org/stable/2100687.

[29] C.P. Arnold and C.H. Dey, *Observing-systems simulation experiments: Past, present, and future*, Bulletin of the American Meteorological Society 67 (1986), pp. 687 – 695, Available at https://journals.ametsoc.org/view/journals/bams/67/6/1520-0477_1986_067_0687_ossepp_2_0_co_2.xml.

[30] K. Nakamura, N. Hirose, B.H. Choi, and T. Higuchi, *Particle filtering in data assimilation and its application to estimation of boundary condition of tsunami simulation model*, in *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg (2009), pp. 353–366, Available at https://doi.org/10.1007/978-3-540-71056-1_19.

[31] C.H. Bishop, B.J. Etherton, and S.J. Majumdar, *Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects*, Monthly Weather Review 129 (2001), pp. 420 – 436, Available at https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0420_aswtet_2.0.co_2.xml.

[32] B. Uzunoglu, S.J. Fletcher, M. Zupanski, and I.M. Navon, *Adaptive ensemble reduction and inflation*, Quarterly Journal of the Royal Meteorological Society 133 (2007), pp. 1281–1294, Available at https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.96.

[33] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation*, Springer International Publishing, 2015.

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

[35] S.E. Ahmed, S. Pawar, and O. San, *PyDA: A hands-on introduction to dynamical data assimilation with Python*, Fluids 5 (2020), Available at https://www.mdpi.com/2311-5521/5/4/225.

[36] J.L. Anderson, *Localization and sampling error correction in ensemble Kalman filter data assimilation*, Monthly Weather Review 140 (2012), pp. 2359 – 2371, Available at https://journals.ametsoc.org/view/journals/mwre/140/7/mwr-d-11-00013.1.xml.

[37] J.C. Oefelein, *Advances in modeling supercritical fluid behavior and combustion in high-pressure propulsion systems*, in *AIAA Science and Technology Forum and Exposition*. 2019, Available at https://arc.aiaa.org/doi/abs/10.2514/6.2019-0634.

[38] S.W. Grib, N. Jiang, P.S. Hsu, P.M. Danehy, and S. Roy, *Rayleigh-scattering-based two-dimensional temperature measurement at 100-kHz frequency in a reacting flow*, Optics Express 27 (2019), pp. 27902–27916, Available at http://www.opticsexpress.org/abstract.cfm?URI=oe-27-20-27902.

[39] M. Bruchhausen, F. Guillard, and F. Lemoine, *Instantaneous measurement of two-dimensional temperature distributions by means of two-color planar laser induced fluorescence (PLIF)*, Experiments in Fluids 38 (2005), pp. 123–131, Available at https://doi.org/10.1007/s00348-004-0911-2.