Object-oriented implementation of the EnVE estimation/forecasting algorithm and its application to high-performance turbulence codes

By T. Bewley†, J. Cessna†, C. Colburn†, F. Ham, G. Iaccarino AND Q. Wang

The estimation/forecasting of multi-scale uncertain flow systems is one of the most visible computational grand challenge problems of our generation. Applications include

- short-term inclement weather forecasting (for hurricanes, etc.),
- contaminant plume forecasting in both urban environments (for coordinating emergency response) and battlefield environments (for coordinating troop movements),
- long-term ocean current forecasting (for El Niño, climate change, etc.), and
- MHD/plasma forecasting (for sunspot cycles, over terms of years, in order to plan space missions, and solar flares and solar wind, over terms of hours, in order to anticipate interruptions of satellite communications).

In order to address such challenges, in addition to improved *simulation* tools for such systems, there must be a concomitant emphasis on improved *data assimilation* algorithms. Data assimilation is essential for synchronizing such simulations with the current flow conditions in real time. It is also instrumental to *quantify the uncertainty* of the predictions, and to *target new observations* in order to minimize such uncertainty.

The new hybrid Ensemble/Variational Estimation (EnVE) algorithm proposed in Bewley, Cessna & Colburn (2008) stands to revolutionize the effectiveness of computational efforts to address such real-time estimation and forecasting problems in high-fidelity discretizations of complex multi-scale/multi-physics problems. The EnVE algorithm efficiently propagates non-Gaussian statistics of the forecast uncertainty while inheriting the favorable smoothing properties of a variational formulation and incorporating a consistent mechanism for revisiting past measurements in light of new data. The present paper provides a brief summary of this approach, then outlines our preliminary efforts in the application of such methods to complex turbulent flow systems.

The focus of the present effort is to implement the EnVE algorithm in a computationally efficient, portable, object-oriented framework which can easily be applied to a wide variety of legacy, high-performance simulation codes. Our initial application of this computational framework is centered on two high-performance MPI-based turbulence simulation codes: Diablo and CDP. Diablo is the pseudospectral stratified DNS/LES code developed at UCSD for simulating flows through simple geometries, and CDP is the unstructured collocated finite-volume LES code developed at Stanford for simulating flows in complex geometries.

1. Introduction

Chaotic systems are characterized by long-term unpredictability. Existing methods designed to estimate and forecast such systems, such as Extended Kalman filtering (a

† Flow Control Lab, Dept. of Mechanical and Aerospace Engineering, U.C. San Diego

"sequential" or "incremental" matrix-based approach) and 4Dvar (a "variational" or "batch" vector-based approach), are essentially based on the assumption that Gaussian uncertainty in the initial state, state disturbances, and measurement noise leads to uncertainty of the state estimate at later times that is well-described by a Gaussian model. This assumption is not valid in chaotic systems with appreciable uncertainties, thus motivating the development of tractable new techniques which revisit past meaurements in light of new data, as done by variational approaches, while summarizing the primary directions of uncertainty of the forecast, as done by Kalman-like approaches. The hybrid technique proposed in Bewley, Cessna & Colburn (2008), dubbed Ensemble Variational Estimation (EnVE), achieves both of these goals.

The two classes of tractable data assimilation strategies today for multi-scale uncertain systems (prior to the development of the hybrid EnVE approach) are the Ensemble Kalman Filtering (EnKF; see Evensen 1994) and the space/time variational (4DVar; see Le Dimet & Talagrand 1986) methods. The EnKF is particularly useful for nonlinear multi-scale systems with substantial uncertainties. In practice, it has been shown to provide significantly improved state estimates in systems for which the more traditional Extended Kalman Filter breaks down. The statistics of the estimation error in the EnKF are not propagated via a covariance matrix, but rather are implicitly represented via the distribution of several perturbed trajectories ("ensemble members") which themselves are propogated with the full nonlinear system model. The collection of these ensemble members (itself called the "ensemble") propagates the statistics of the estimation error accurately in many problems, even when a relatively small number of ensemble members is used. On the other hand, 4Dvar methods propagate state and sensitivity ("adjoint") simulations back and forth across an optimization window of interest. An optimization is performed based on these marches in order to minimize a cost function balancing (a) a term accounting for the misfit of the estimate with the measurements over the optimization window, with (b) a "background" term accounting for the "old" estimate based on the measurements obtained prior to the optimization window.

The application of a simulation-based estimation/forecasting scheme to atmospheric systems in particular presents several significant modeling challenges, including:
(1) those due to insufficient grid resolution in stratefied flows at very high Re;
(2) those due to parameter uncertainty in the physical represenation of the system, particularly in relation to temperature/density stratification and grossly simplified chemistry, thermal radiation, and evaporation/condensation models;
(3) those due to the inflow and outflow boundary conditions at the edges (in latitude, longitude and altitude) of the computational domain (such boundary conditions are often taken from very coarse global circulation models of the atmosphere).
Collectively, these complexities make the problem of estimation and forecasting in atmospheric flows exceedingly difficult. Ultimately, any fair comparison of data assimilation methods for such problems must be conducted in the face of all of these uncertainties. In the interest of developing a new data assimilation framework optimally suited to handle such complex uncertain systems, our team is currently considering the estimation and forecasting of multi-scale turbulent flow systems in a significantly more controlled setting, while developing a software framework which will extend easily to high-fidelity atmospheric codes developed at, e.g., NCAR, the Met Office and the DOE labs.

2. The EnVE algorithm

The EnVE algorithm is now presented as a consistent hybrid of the EnKF and 4DVar algorithms. Assume, without loss of generality, that an EnKF estimate $\widehat{\mathbf{X}}_{-j|-j}$ exists at some past time t_{-j}. This ensemble represents an estimate of the system state at time t_{-j} given measurements up to and including $\mathbf{y}_{-j}$. At this point, available measurements up to t_0 are considered. The EnVE algorithm is initialized via a traditional sequential march of the EnKF up to the time of the most recent measurement, t_0. This provides the updated ensemble $\widehat{\mathbf{X}}_{0|0}$ and all of its implicit statistics. The mean of the estimate is denoted $\bar{\mathbf{x}}_{0|0}$, and is found by taking the average of all the ensemble members. In a linear system with a very large number of ensemble members, this would give an accurate approximation of the best estimate at time t_0 given measurements up to and including time t_0. However, errors due to the nonlinearity of the chaotic system and approximations due to the finite size of the ensemble ultimately lead to a suboptimal estimate via the EnKF approach.

For forecasting applications, the most important estimate is the one at the most recent measurement time t_0, because it is this which is used as an initial condition for any forecasting calculation. With a linear system, any type of smoothing at this stage in the EnKF algorithm would have no effect on the estimate at t_0. The smoother would simply reduce the error in the past estimates, for some time $t < t_0$, using the information in the observations between t and t_0. However, for a nonlinear system, smoothing affects the entire estimate trajectory, even the most recent estimate at t_0. This is due to the dependence of the evolution of the estimate uncertainty on the trajectory of the estimate itself. For a linear system, the covariance propagation is independent of trajectory. However, for a nonlinear system, changes in a past estimate (via smoothing) will impact the future trajectory of the estimate and its associated covariance. This motivates the consistent revisiting of past measurements to help improve the resulting forecast.

To this end, the ensemble $\widehat{\mathbf{X}}_{0|0}$ is marched backward, using only the model equations. In so doing, the estimate retains the information captured by later measurements during the forward EnKF march (resulting in what is known in the language of estimation theory as a "smoothed" estimate). Thus, any point on this resulting trajectory is conditioned on all available measurements. At the conclusion of this backward march, the ensemble mean and its implied statistics are known at some past time, say t_{-K}. This retrograde march is monitored in such a way as to define the width of the observation window for the subsequent variational step of the EnVE algorithm. If the initial estimate at t_0 is poor, then a significant amount of useful information may be deduced from a small time window containing only a few observations. Including more observations in this case is superfluous, and in fact unnecessarily increases the complexity of the optimization surface. Conversely, if the initial estimate at t_0 is very accurate, then a significantly longer variational window can, and should, be included in the analysis in order to account for more measurements.

The backward march defines the window width by looking at the correlation between the initial estimate's trajectory and the past measurement history. Poor estimates diverge quickly from the measurements and should be analyzed with short optimization windows; conversely, accurate estimates march much further back in time before they begin to diverge from the measurements, and should be analyzed with longer optimization windows. To quantify this divergence, a "bias" measure, $B_k = \left\| \sum_{j=0}^{-k} (\mathbf{y}_j - H\,\bar{\mathbf{x}}_{j|0}) \right\|_1$, is calculated during the backward march. Through experimentation, a critical bias $\bar{B}$ is defined such that the ensemble mean is deemed significantly divergent from the observations once k is sufficiently large that B_k exceeds $\bar{B}$; the corresponding value of k is

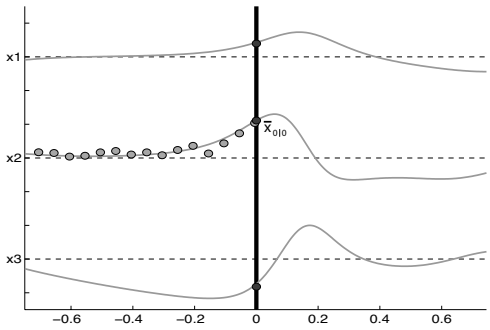

FIGURE 1. EnVE is initialized by marching a traditional EnKF forward through the available observations, making the appropriate updates. This provides an updated ensemble, $\hat{X}_{0|0}$, at the current time t_0, and the corresponding estimate given by the ensemble mean $\bar{\mathbf{x}}_{0|0}$. At this point, it may be beneficial to revisit past measurements in light of the most recent data.

denoted K. With the variational window $[\,t_{-K}, t_0\,]$ so defined, the initial best smoothed estimate of the state $\bar{\mathbf{x}}_{-K|0}$ is given as the mean of the ensemble $\widehat{\mathbf{X}}_{-K|0}$. At this point, variational methods are used to improve this estimate in a consistent manner.

To this end, the traditional 4DVar cost function is defined with a background estimate and covariance at t_{-K}. The background term of the cost function must now be defined carefully, as the correct background term is essential for EnVE to be consistent. In other words, properly defining the background terms in the variational cost function guarantees that erroneous updates are not made by using an observation more than once, and ensures that the result obtained reduces to that obtained by the Kalman Filter in the special case that the system considered happens to be linear. Therefore, the background term must be determined by returning to the original ensemble, $\widehat{\mathbf{X}}_{0|0}$, and marching it backward again to the left edge of the window t_{-K}, this time *removing* the effect of the measurements along the way.

A suitable formula for removing the effect of a measurement can be found by rearranging the standard forward KF update equations for the mean and covariance (Bewley, Cessna & Colburn 2008). This removal formula (combined with a suitable backward march of the system) may be used to produce the background ensemble $\widehat{\mathbf{X}}_{-K|-K}$ at the left edge of the variational window. From this background ensemble, the background mean $\bar{\mathbf{x}}_{-K|-K}$ and background covariance $\mathcal{P}^e_{-K|-K}$ can be extracted and used to define the variational cost function. Because the background terms of the cost function are consistently defined (in that, in the linear setting, they incorporate no information from the observations in the variational window), the corresponding n-dimensional optimization surface is, in the linear case, identical to what would have been used had no sequential march through those observations been completed. The global minimum of this surface is independent of any previous updates to the estimate within the variational window that have been computed.

With the cost function defined appropriately in this manner, a variational iteration can now be performed, similar to 4DVar. With traditional 4DVar, the first iteration is typically initialized using the background term, $\mathbf{u} = \bar{\mathbf{x}}_{-K|-K}$. However, with EnVE, a much better estimate than this is already known, namely $\bar{\mathbf{x}}_{-K|0}$. This is one of the strengths of EnVE: it initializes the variational iteration with an estimate that is known to be significantly better than the background. In either case, the optimization surface is

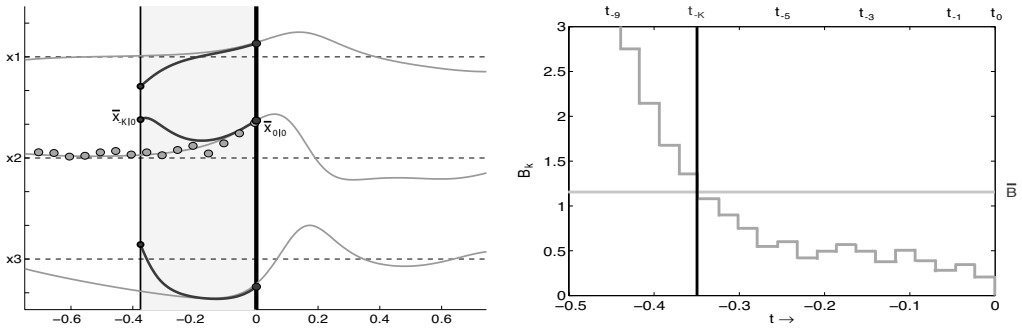

FIGURE 2. (left) To determine the accuracy of the current estimate (that is, its correlation with recent measurements), the ensemble at the current time is marched backward using the system equations until the trajectory of the ensemble mean is significantly divergent from the observations. This gives a "smoothed" estimate at the past time, $\bar{\mathbf{x}}_{-K|0}$. (right) The "bias" between the estimate trajectory and the observations is accumulated as the smoother is marched backward. Upon reaching a critical bias $\bar{B}$, the retrograde march is stopped. This time t_{-K} defines the width of the subsequent variational window.

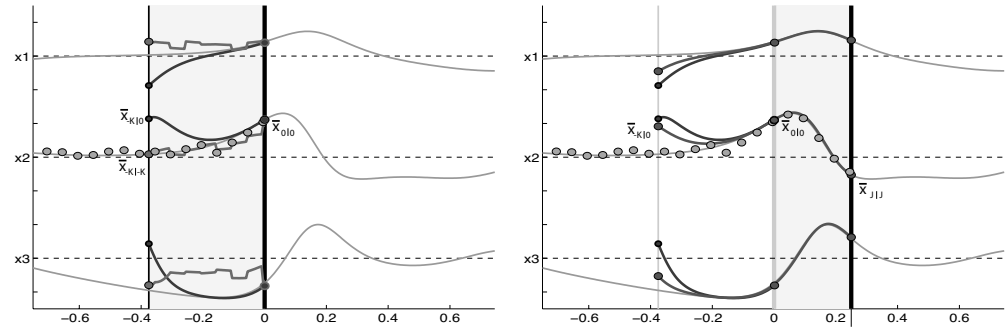

FIGURE 3. (left) To determine the variational cost function, the background terms at t_{-K} must be calculated. This is done by marching the original ensemble $\widehat{\mathbf{X}}_{0|0}$ backward through the window, sequentially *removing* the effect of each measurement update. This march results in a background ensemble $\widehat{\mathbf{X}}_{-K|-K}$ at the left edge of the optimization window; from this, the background mean and covariance can be inferred. (right) Upon completion of a variational iteration, the improved ensemble estimate $\widehat{\mathbf{X}}_{-K|0}$ at the left edge of the optimization window is propagated forward to the old current time t_0. No measurement updates are done during this march, as the available observations have already been accounted for by the variational analysis. Upon reaching t_0, the new ensemble estimate $\widehat{\mathbf{X}}_{0|0}$ is marched further to the new present time t_J using the EnKF to account for any new measurements recently received, and the process is repeated.

identical, but with EnVE, the initial estimate for $\mathbf{u}$ is much closer to the global minimum than the original background term. Consequently, if any significant improvement can be made upon the original best estimate, it will be discovered in the first variational iteration. Further, the original estimate is more likely to be in the region of attraction of this global minimum, so the probability of erroneous convergence to spurious local minima can be substantially reduced.

Rather than propagating the ensemble mean, the variational problem can be restructured and posed so as to measure the misfit between the mean of the ensembles and the measurements. Thus, an ensemble of adjoints is defined, where each adjoint is linearized about the trajectory of its corresponding ensemble member. Then, the gradient is found

as the mean of the ensemble of adjoints at the initial time. In practice, each member of the ensemble is "shifted" in phase space the same amount, effectively shifting the ensemble mean without affecting the higher-order ensemble statistics. In fact, with an adjusted estimate of this sort, a modified, if not improved covariance $\mathcal{P}^e_{-K|0}$ would be expected as well. However, as variational methods do not provide a means for tracking these changes, EnVE must simply use this shifted ensemble representation, which is a bit conservative. Note, though, that this is a significant improvement over 4Dvar, in which rigorous methods to march $\mathcal{P}$ are essentially unavailable. In contrast, with EnVE, the covariance associated with the original smoothed estimate is available, so it can be utilized. Though this is a conservative estimate of the covariance that does not account for the correction to the estimate due to the variational step, it correctly captures the main features of the covariance matrix, including the principle directions of estimate uncertainty.

To cycle the algorithm, the updated ensemble at t_k is marched forward to t_0. Note that the ensemble already accounts for the measurement in the window, so each ensemble member is propagated forward using the system equations only, with no additional measurement updates. This gives an improved best estimate at t_0, $\widehat{\mathbf{X}}_{0|0}$. During the time elapsed while completing this iteration, some new measurements $\{\mathbf{y}_1 \cdots \mathbf{y}_J\}$ will usually become available due to the computational time required to complete the variational step. The ensemble $\widehat{\mathbf{X}}_{0|0}$ can thus be marched forward again, using the EnKF to account for these new measurements, until the new current time t_J is reached. At this point, the time index is reset $t_0 \leftarrow t_J$, and the algorithm is repeated. Note that a significant computational burden can be avoided by storing the updated ensemble at the previous current time, $\widehat{\mathbf{X}}_{0|0}$. This point can serve as the initial condition for finding the background terms of the variational cost function, as opposed to starting from $\widehat{\mathbf{X}}_{J|J}$. Depending on the relative widths of the next variational window and the time elapsed during the current variational step, using this saved ensemble will result in either a shorter backward EnKF march (very beneficial due to the ill-posed nature of such a march) or possibly even a forward EnKF march (a well-posed march) to the left edge of the new variational window. This simple storage trick reduces the computational cost of the algorithm significantly and shortens (or removes altogether) one of the ill-posed backward marches.

2.1. *EnVE consistency*

Ultimately, sequential methods (EnKF) and variational methods (4DVar) are used to solve the same problem. Both methods work to minimize a cost function to optimize the estimate at t_0 conditioned on all available measurements. Thus, when these cost functions are defined appropriately, it is possible to switch back and forth between sequential and variational methods consistently, as EnVE does. For a linear system with a set of measurements defined on $[\,t_{-K}\,,\,t_0\,]$, the smoothed KF estimate at the back edge of the window, $\bar{\mathbf{x}}_{-K|0}$ (found by marching a KF forward through the observations and marching the resulting estimate backward to t_{-K}) is identical to the solution of a converged 4DVar algorithm with an appropriately defined background term. In other words, the optimal smoothed KF estimate $\bar{\mathbf{x}}_{-K|0}$ is the global minimum of the 4DVar cost function in the case of a linear system. For nonlinear systems, this relationship is still true, but the optimal smoothed KF estimate can not necessarily be found via a sequential estimator.

This relationship is what EnVE attempts to exploit to improve the estimate. Marching an Ensemble Kalman Smoother (EnKS) will not produce the optimal smoothed estimate $\bar{\mathbf{x}}_{-K|0}$ because of the nonlinearities in the system and the approximations required for the ensemble framework. However, by removing the effect of the measurements and appro-

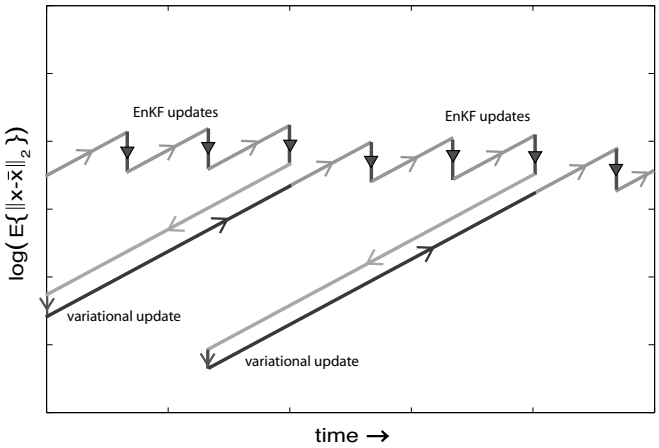

FIGURE 4. A cartoon illustrating the expected error for EnVE performed on a chaotic system. Exponential growth (linear growth in semi-log coordinates) in the expected error occurs during forward marches. Discrete reduction in the expected error occur at both the sequential updates and the variational update. Note that with a linear system, the variational update is necessarily zero, returning the estimate to its original value upon completion of the variational step.

priately defining the 4DVar cost function background terms, this suboptimal smoothed estimate can be used as an initial condition for the variational step. If the smoothed estimate $\bar{\mathbf{x}}_{-K|0}$ happens to be optimal, then the variational iteration is already converged and will produce a zero update to the estimate. Thus, EnVE uses the EnKS to initialize the 4DVar optimization, but does not reuse the information in the observations inconsistently. Thus, EnVE reduces to the expected optimal results of the Kalman Smoother (KS) for a linear system with both Gaussian measurement noise and disturbances.

An illustration of the expected estimation error as EnVE progresses for a typical chaotic system is shown in Fig. 4. Due to the chaotic nature of the system, any forward march of an estimate will lead to expected exponential growth of the estimation error (shown linearly in semi-log coordinates). Each EnKF measurement update creates a discrete drop in the expected estimation error. When a variational iteration is performed, the estimate is marched backward. This causes an exponential decrease in the expected error as trajectories of the chaotic system will converge (along the attractor) during the backward march. Then, a variational update is made, further reducing the expected error, and the resulting estimate is propagated forward again to the next available measurement. Recall that with a linear system, the update due to the variational step will have zero length, thus returning the estimate back to its original state to continue the sequential march. This helps illustrate the consistent nature of EnVE.

3. Advantages

By combining the statistical capabilities of the EnKF along with the batch processing/smoothing capabilities of a variational method, EnVE builds a better estimate of the system, possibly in real time, at a justifiable computational cost. Using the EnKF to initialize a 4DVar-like iteration allows for fewer iterations because full convergence is not required and the initial estimate is far more accurate than the background estimate alone. The intrinsic ability of the EnKF to represent the statistical properties of the esti-

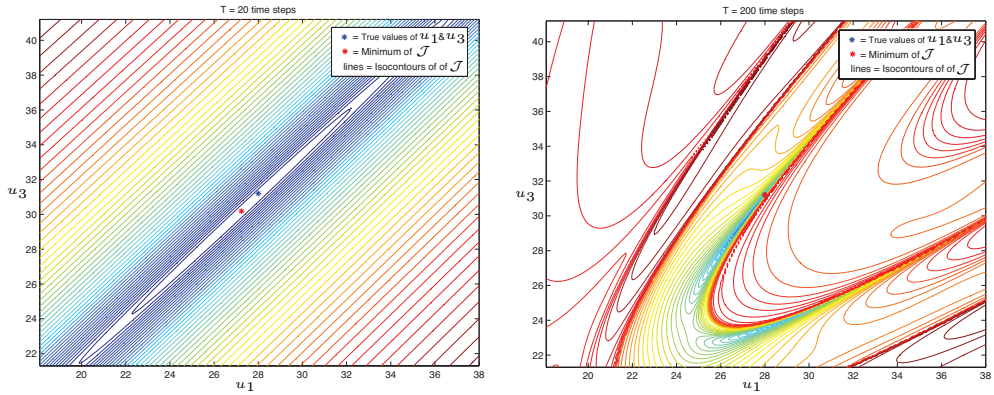

FIGURE 5. A cartoon illustrating the change in complexity of the optimization surfaces for a short variational window (left) and a long variational window (right). Also shown is the known truth model global minimum, which is more closely related to the global minimum of the highly irregular optimization surface of the longer window.

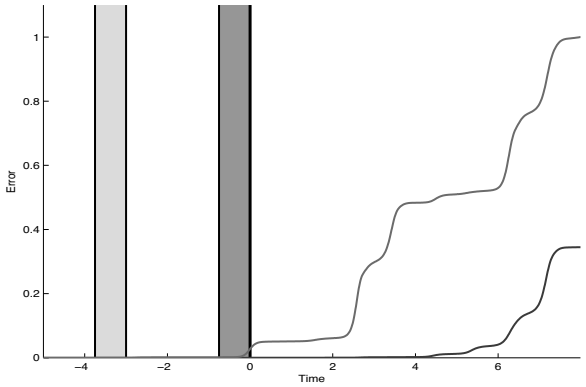

FIGURE 6. The accumulated forecast error from two forecasts. The left-most variational window is for a simple 4DVar without a receding-horizon framework. The right-most window is for EnVE with a receding-horizon framework. Note the difference in the accumulated errors of each forecast is due in large part to the time the forecast is ahead of the latest optimization window used. As this time is significantly reduced in the receding-horizon framework, forecasts made a certain amount of time into the future are greatly improved.

mate allows EnVE to repeatedly and consistently revisit past measurements and update the central trajectory of the ensemble (about which the system can be linearized when considering its covariance evolution) based on new measurements.

A key feature of the EnVE framework is that it combines a multiscale-in-time algorithm with a receding horizon optimization framework. The advantages of these properties are highlighted in the following section. Combined, they provide a dynamic optimization surface that tends to have desirable convergence properties for highly nonlinear systems.

3.1. *Multi-scale in time*

Because the variational window in EnVE is defined from the right (current time) by marching the current estimate backward until divergence, the width of this window can be selected during the iteration. In contrast, with traditional 4DVar, this window width must be specified in advance. The variable variational window widths of EnVE can be

used as a tool to precondition the optimization problem appropriately by coordinating this width with the accuracy of the initial estimate, as discussed previously.

Due to the noise in the measurements, a short window containing only a few observations is prone to inaccuracy. That is, the global minimum of the cost function defined over only a few observations is likely to deviate significantly from the "truth." However, because only a few measurements are included in this short window (with corresponding short marches of the chaotic system) this optimization surface tends to be more regular, with a larger region of attraction for the global minimum. The size of the region of attraction is important with gradient-based algorithms, as they are prone to converge to local minima.

As the estimate improves, longer windows with more observations included can be utilized. This will tend to make the optimization surface more irregular and shrink the region of attraction for the global minimum, and thus this extension of the variational window needs to be done gradually enough that the improved estimate remains in this reduced region of attraction. Because more measurements are included in this window, the effect of sensor noise is diminished from the shorter window, making this global minimum more accurate with respect to the "truth."

3.2. *Receding horizon*

A receding-horizon approach is defined by nudging the variational window forward in time to incorporate the most recent measurements obtained during each iteration of the variational optimization. Simplistic approaches to variational data assimilation leave the optimization window fixed until convergence. In contrast, EnVE redefines the optimization problem slightly at each iteration, updating it to include the newly obtained measurements. As this modification causes the optimization surface to shift constantly, the algorithm never completely converges. However, the receding-horizon optimization framework updates the current estimate at each iteration with maximal efficiency, as it is constantly using the most up-to-date information available. Further, the resulting dynamic evolution of the optimization surface in fact helps to "nudge" the estimate out of the local minima into which it might otherwise settle.

A typical contrast between the error of two forecasts (one generated with a fixed-horizon 4DVar algorithm and the other with a receding-horizon EnVE algorithm) is shown in Fig. 6. Unlike the receding-horizon EnVE algorithm, due to the computation required for convergence of the fixed-horizon 4DVar algorithm, the corresponding variational window has slipped into the past. Because of the chaotic nature of the systems of interest, any forecast will begin to diverge exponentially. Consequently, much of the relevant range of the fixed-horizon 4DVar forecast is wasted predicting events that have already taken place.

3.3. *Parallel state/adjoint marches*

As previously mentioned, another advantage of posing the variational optimization problem in a retrograde setting deals with the numerical implementation of EnVE. The adjoint equation is marched backward in time (from t_0 to t_{-K}), forced using the trajectory $\tilde{x}(t)$. Typically, this trajectory is found by marching the initial condition $\tilde{x}_{-K} = \mathbf{u}$ forward through the window (from t_{-K} to t_0). Especially for the multi-scale systems of interest, this poses a large storage constraint on the problem, because the adjoint is forced by the whole trajectory, but in reverse order. In other words, the trajectory of $\tilde{x}(t)$ needs to be computed and saved over the entire interval before the adjoint march can begin. Attempts to circumvent this problem for large atmospheric-scale systems include the checkpointing

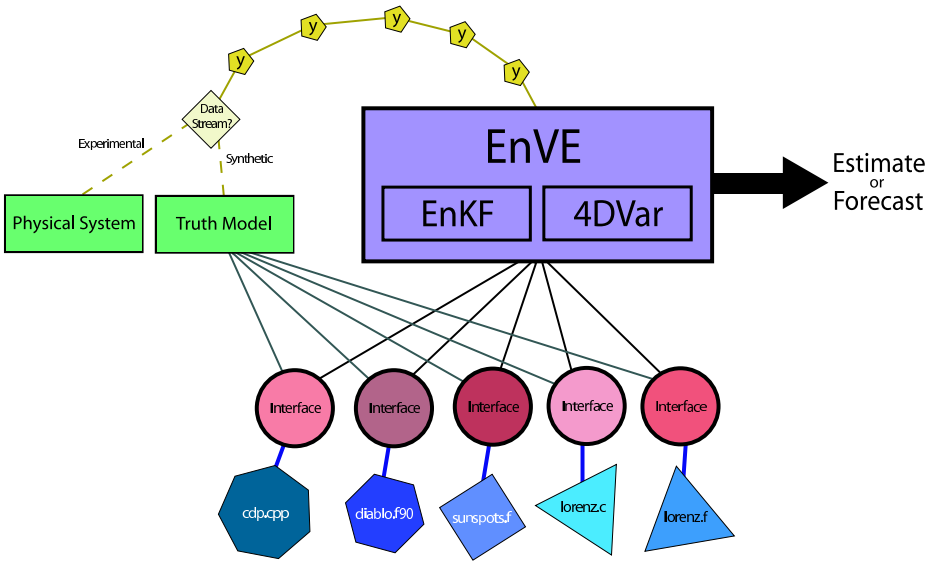

FIGURE 7. A depiction of the object-oriented strucutre of the EnVE implementation. The computational data assimilation algorithm (and code) is largely independent of the particular system model used. Consequently, much of the code specific to the individual model is isolated at a level below that of the main EnVE code. A compact, well-defined and largely portable "standard interface" is provided, enabling EnVE to drive a variety of system models and facilitating mixed-language programming.

algorithm, in which the trajectory is stored only on coarse time grid points, and then, as necessary, is either recomputed or linearly interpolated onto the fine (in time) grid used for time-stepping the adjoint calculation. Checkpointing requires a substantial amount of storage and significantly increases the computation required to compute the adjoint.

Note that, with EnVE, this required estimate trajectory is determined backward in time rather than forward in time. Thus, the corresponding adjoint may be computed simultaneously, eliminating this storage problem altogether.

3.4. *Object oriented framwork*

Because the theoretical development of EnVE is largely model-independent, an object-oriented hierarchical implementation of the EnVE algorithm has been developed. This C++ code is a wrapper that can be used to apply the EnVE algorithm to almost any desired model, with the development of an appropriate and fairly straightforward interface. By separating the data assimilation from the model simulation, the EnVE implementation code is able to adapt easily to complex legacy codes, such as Stanford's CDP code in the present work. Such a framework puts the minimum restrictions on the specific structure of the existing individual models.

In addition, the object-oriented framework extends naturally to handle the algorithmic overhead associated with the incoming observations. In real-life applications, observations will be streaming into the assimilation system from multiple sensors, each with their own types of measurements, time stamps, reliability and time delays (between when the measurement is taken and when this measurement is recieved by the data assimilation algorithm). The object-oriented EnVE implementation provides a means for handling this variability in a clean and efficient manner. For testing purposes, when no physical

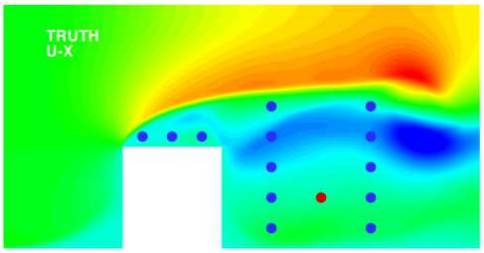

FIGURE 8. The computational setup of the 2-D problem. Flow moves from left to right in an open channel over a bluff body. For the data assimilation, 13 probes (on top of the 'building' and the two parallel veritcal rakes downstream) were placed in the flow measuring the velocity components and the scalar concentration. The isolated probe (located downsteam between the two rakes) is used only for retrospective analysis (see Fig. 11) and was not assimilated into the estimate.

data is available, the model interface can also be used to artificially generate this stream of observations.

One of the key difficulties when interfacing with an independently developed code base is dealing with mixed-language programming. Writing the outer EnVE shell in C++ enables more low level control with respect to memory management and mixed-language function calls, further increasing the overall versatility of the code.

4. Preliminary EnKF simulation and results

As a demonstration of the capabilities of our object-oriented implementation of EnVE, we consider the estimation of a 2-D flowfield with a passive scalar release in the wake of a bluff body. The ensemble members were simulated on a grid with both a factor of 64 fewer grid points and significantly larger time steps than the truth model. This under-resolution parallels one of the key challenges to be expected in operational atmospheric problems. Though only 25 grossly under-resolved ensemble members were used in the EnKF formulation, the ensemble accurately captured both the large-scale features of the truth model and its principle directions of uncertainty. The result thus suggests that, when assimilating experimental data, it is indeed possible to capture accurately the dominant dynamics of a complex system by replacing high-fidelity numerical simulations with an appropriately forced ensemble of under-resolved calculations.

Figures 9 and 10 show clearly that EnVE is capable of estimating the large-scale dynamics of interest in a 2-D vortex shedding problem past a simple bluff body. Note in Fig. 9 that the variation of concentrations over large length scales are generally estimated accurately, with more significant errors occuring over small length scales. Significantly, the variance of the ensemble accurately depicts where the mean of the ensemble actually differs from the truth model, thus providing a useful indicator of the accuracy of the estimate provided by the EnKF analysis. Note also that the structure of the uncertainty appears to be strongly correlated with the structures in the flow.

As another means for evaluating the estimator, an additional sensor was used to perform a retrospective pointwise comparison of the truth and the estimate. Fig. 11 compares the estimated current and forecast velocities at a point (indicated in Fig. 8) with the true velocities observed in the truth model. It is clear that the estimator provides a forecast that remains well-correlated with the flow itself over a time scale which is significant with respect to the dynamics of the large-scale features of the flow.

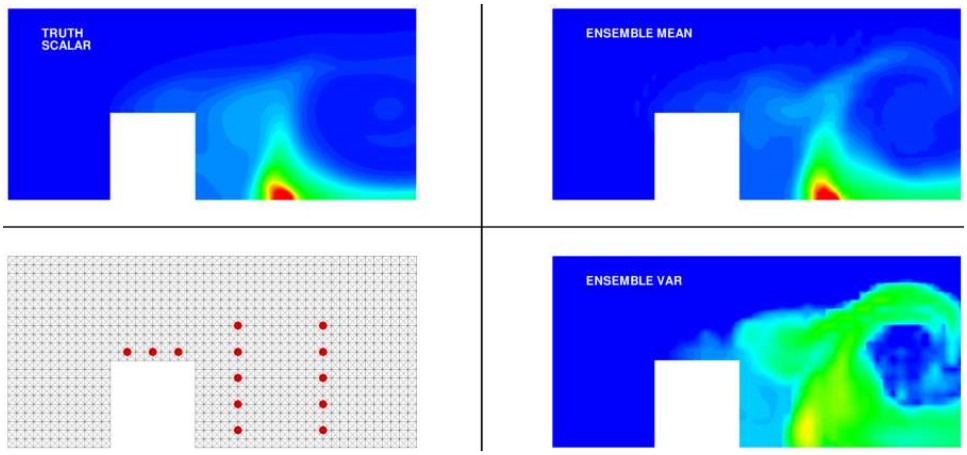

FIGURE 9. A typical snapshot of the EnKF assimilation showing the passive scalar concentration. In both the truth (top left) and estimate (top right), the passive scalar is being released at a known rate and location; in this case, just downstream from the bluff body. Though the grid on which the estimate is calculated (bottom left) is extremely course, we are able to capture many of the moderate to large-scale features in the passive scalar concentration. Additionally, note that the variance in the ensembles (bottom right) is small in the potential flow region upstream and near the center of the first shed vortex downstream.

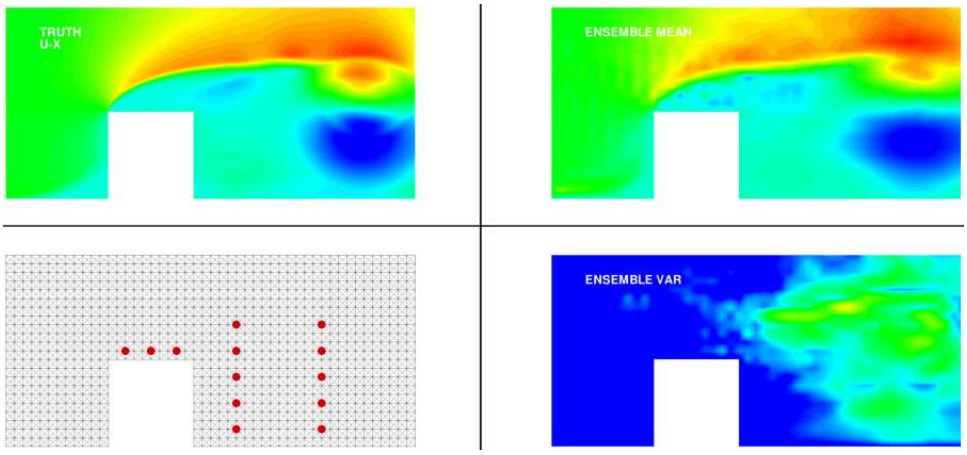

FIGURE 10. As in Fig. 9, showing now the horizontal component of the velocity. In the truth model (top left), the high-fidelity simulation clearly shows the separation of the flow at the leading edge of the bluff body and the location of the most recently shed vortex. Note again that we are able to estimate the approximate location of this vortex in the ensemble mean (top right). As in Fig. 9, the ensemble variance (bottom right) is low near the upstream boundary condition and higher in the turbulent wake, where the flow structure is less certain.

5. Summary and conclusions

This paper summarizes a new hybrid Ensemble Variational Estimation (EnVE) algorithm for data assimilation in complex systems, our implementation of this algorithm in a portable and extensible object-oriented framework, and our preliminary efforts to apply this framework to high-performance turbulence codes.

The EnVE method leverages the nonlinear statistical propagation properties of the

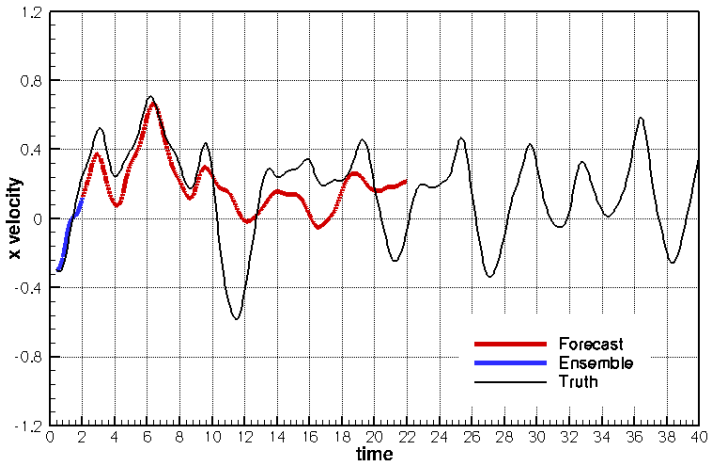

FIGURE 11. For the purposes of retrospective analysis, a probe was placed in the flow but not used in the assimilation process (see Fig. 8 for the precise location). The thin line shows the time history of the horizontal velocity of the truth model, where the oscillations due to turbulent vortex shedding become apparent. For the present plot, the thick line represents a time history of the EnKF estimate up to the present time ($t = 2$), after which the thick line represents the current operational forecast of the assimilation. Note that this forecast is statistically correlated with the known future truth for the next three shedding cycles.

sequential EnKF/EnKS to initialize and define properly an appropriate variational iteration, similar to 4DVar. This variational iteration is posed in such a way as to allow for a multiscale-in-time, receding-horizon optimization framework. The smoothed estimate from the EnKF is used as an accurate initial condition for the variational iteration, thus improving its overall performance. The multiscale-in-time framework is achieved via a retrograde march of the current estimate over the available observations, and appropriately preconditions the variational step. This also allows for a concurrent, parallel march of the appropriate adjoint ensemble. Thus, no additional storage is required for the gradient computation, as is otherwise typical with a 4DVar implementation.

The full EnVE algorithm has been develped into a compact, portable, object-oriented framework which can be applied to a wide variety of possible underlying simulation codes. By divorcing the EnVE algorithm from the underlying model, this framework is easily adapted to complex independently developed simulation codes. We have applied this framework to Stanford's CDP code; preliminary computational results are given in the present paper. Though the adjoint of the CDP solver is not yet fully operational, we have successfully incorporated the CDP code base into the present data assimilation framework, running (for now) in EnKF mode only. Computational results obtained thus far are quite promising, yet still show room for significant improvement, which we expect the full EnVE framework to provide.

Acknowledgments

The authors gratefully acknowledge the generous financial support of the National Security Education Center (NSEC) at Los Alamos National Laboratory (LANL) and the Center for Turbulence Research (CTR) at Stanford University.

REFERENCES

Bewley, T., Cessna, J. & Colburn, C. 2008 EnVE: A consistent hybrid ensemble/variational estimation strategy for multi-scale uncertain systems. *Tellus* A. *under review.*

Evensen, G. 1994 Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Review* **99**, 10143–10162.

Kim, J. & Bewley, T. 2007 A linear systems approach to flow control. *Annual Review of Fluid Mechanics* **39**, 383–417.

Le Dimet, F.-X. & Talagrand, O. 1986 Variational algorithms for analysis and assimilation of meteorological observations; theoretical aspects. *Tellus* **38A**, 97.