Data Assimilation with Gaussian Mixture Models using the Dynamically Orthogonal Field Equations. Part II. Applications

Sondergaard, T. and P.F.J. Lermusiaux (2012). Data Assimilation with Gaussian Mixture Models using the Dynamically Orthogonal Field Equations. Part II: Applications, Monthly Weather Review, doi:10.1175/MWR-D-11-00296.1, in press.

Data Assimilation with Gaussian Mixture Models using the Dynamically Orthogonal Field Equations. Part II: Applications

Thomas Sondergaard and Pierre F. J. Lermusiaux *

Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts

ABSTRACT

The properties and capabilities of the GMM-DO filter are assessed and exemplified by applications to two dynamical systems: (1) the Double Well Diffusion and (2) Sudden Expansion flows; both of which admit far-from-Gaussian statistics. The former test case, or twin experiment, validates the use of the EM algorithm and Bayesian Information Criterion with Gaussian Mixture Models in a filtering context; the latter further exemplifies its ability to efficiently handle state vectors of non-trivial dimensionality and dynamics with jets and eddies. For each test case, qualitative and quantitative comparisons are made with contemporary filters. The sensitivity to input parameters is illustrated and discussed. Properties of the filter are examined and its estimates are described, including: the equation-based and adaptive prediction of the probability densities; the evolution of the mean field, stochastic subspace modes and stochastic coefficients; the fitting of Gaussian Mixture Models; and, the efficient and analytical Bayesian updates at assimilation times and the corresponding data impacts. The advantages of respecting nonlinear dynamics and preserving non-Gaussian statistics are brought to light. For realistic test cases admitting complex distributions and with sparse or noisy measurements, the GMM-DO filter is shown to fundamentally improve the filtering skill, outperforming simpler schemes invoking the Gaussian parametric distribution.

1. Introduction

In part I of this two-part paper, we derived the GMM-DO filter: data assimilation with Gaussian Mixture Models (GMMs) using the Dynamically Orthogonal (DO) field equations. The result was an efficient, rigorous, data-driven assimilation scheme preserving non-Gaussian statistics and respecting non-linear dynamics. In the present study, we evaluate its performance against contemporary filters in a dynamical systems setting, including ocean and fluid flows. In section 2-a, we examine the application of the GMM-DO filter to the Double Well Diffusion Experiment (Miller et al. 1994; Eyink and Kim 2006), which is based on a classic stochastic system, time-dependent but of zero dimension in space. We compare our results with those of the Ensemble Kalman filter (Evensen 1994; Houtekamer et al. 1998) and the Maximum Entropy filter (Eyink and Kim 2006). For clarity, the latter filter is outlined in Appendix A of this paper. In section 2-b, we consider flows that are more realistic for coastal ocean or fluid dynamics. Specifically, we consider dynamic jets and eddies that occur in Sudden Expansion flows (Cherdron et al. 1978; Fearn et al. 1990; Durst et al. 1973) of two dimensions in space. We illustrate and study the results of the GMM-DO filter, including the evolution of probability density functions (pdfs) and of their DO decomposition, the Bayesian impacts of observations, and the overall capabilities of the filter. We also compare the GMM-DO filter's performance to that of an ESSE filter, scheme A (Lermusiaux and Robinson 1999).

For each test case, we critically evaluate the properties of the GMM-DO filter and outline the advantages that arise through its utilization. We illustrate and stress its equation-based and adaptive characteristics, which eliminate the need for heuristics or ad-hoc choices. We further conduct sensitivity studies in which we examine the filter's performance to variations in the following independent parameters: the model error; the number of observations and the observation error; and the number of subspace realizations. We give our concluding remarks in section 3. Table 1 provides the notation relevant to this manuscript.

2. Applications

a. Double Well Diffusion Experiment

The Double Well Diffusion Experiment has served as a test case for several assimilation schemes (e.g. Miller et al. 1994), recently among them the Maximum Entropy filter (see Appendix A). Due to the bimodal climatological distribution of the double well, the experiment lends itself to the evaluation of filters that aim to capture and extract non-Gaussian features.

Given the experiment's low dimensionality (specifically,

the state is a scalar, i.e. $n = 1$), the DO equations are here not needed and are thus not used. A first purpose of this test case is to evaluate the use of the EM algorithm and Bayesian Information Criterion (BIC) with GMMs in a dynamical systems setting. As described in Part I, such validations – applied to different test cases – were the used by Smith (2007), Dovera and Rossa (2010) and Frei and Kunsch (2011). A second purpose is to evaluate the sensitivities and capabilities of the GMM-DO filter as one varies key parameters, specifically the model error, the observation error and the number of ensemble realizations.

1) Description of Experiment

In the Double Well Diffusion Experiment, the goal is to track the location of a ball, $X(t)$, located in one of two wells. The ball is forced under 'pseudo-gravity' and externally excited by white noise. Specifically, the location of the ball evolves according to the following scalar stochastic differential equation (Miller et al. 1994):

$$dX = f(X)dt + \kappa d\Gamma(t), \quad \Gamma \sim \mathcal{N}(\gamma; 0, 1), \tag{1}$$

with

$$f(X) = 4X - 4X^3 \tag{2}$$

essentially acting as the gravitational force (see figure 1). The strength of the stochastic forcing is tuned by the diffusion coefficient κ. We also note that $X \in \mathbb{R}$.

We occasionally get access to direct, but noisy, measurements, y, of the ball location, modeled as:

$$p_{Y|X}(y|x) = \mathcal{N}(y; x, \sigma_o^2). \tag{3}$$

From these measurements, we wish to infer the current location of the ball. We are thus faced with a filtering task.

This experiment is an ergodic Markov Chain (e.g. Cover and Thomas 2006) and therefore possesses a stationary distribution (from hereon a *climatological* distribution), $q_X(x)$. It can be shown that this distribution satisfies (Eyink and Kim 2006):

$$q_X(x) \propto e^{-\frac{2x^4 - 4x^2}{\kappa^2}}, \tag{4}$$

which can be approximated by a GMM of complexity two, i.e.

$$q_X(x) \approx \sum_{m=1}^{2} w_m \times \mathcal{N}(x; \mu_m, \sigma_m^2), \tag{5}$$

with – by arguments of symmetry – the following parameters:

$$w_1 = w_2 = 0.5 \tag{6}$$

$$-\mu_1 = \mu_2 = \mu \tag{7}$$

$$\sigma_1^2 = \sigma_2^2 = \sigma^2. \tag{8}$$

For the particular case of $\kappa = 0.40$, Eyink and Kim (2006) estimated the mean and variance of the GMM to be around $\mu = 0.98$ and $\sigma^2 = 0.011$, respectively. This is plotted against the exact distribution in figure 2.

The choice of κ, the diffusion coefficient, determines the average time that the ball spends in a well before transitioning. For instance, according to Eyink and Kim (2006), for $\kappa = 0.40$, this residence time is $\tau_{res} \approx 10^5$ with transitions from one well to the other taking only $\tau_{trans} \approx 10^1$. For small values of κ, the system thus behaves in a manner similar to a noisy switch.

2) Test Procedure

We solve the governing stochastic differential equation, (1), by application of the Euler-Maruyama scheme (Higham 2001):

$$x_{i,k+1} = x_{i,k} + f(x_{i,k})\Delta t + \kappa\gamma_{i,k}\sqrt{\Delta t}, \quad i = 1, \ldots, N, \tag{9}$$

where γ is white in time and drawn from a normal distribution with zero mean and unit standard deviation, and $x_{i,k}$ is the i^{th} Monte Carlo realization at discrete time k.

Our goal is to evaluate the performance of the GMM-DO filter against the Ensemble Kalman filter (EnKF) and the Maximum Entropy filter (MEF) in its ability to track the ball. We did so by repeating the experiment for a large number of parameter values and, in Sondergaard (2011), we report results for a subset of these evaluations, specifically:

- Diffusion coefficient, $\kappa = \{0.4, 0.5\}$
- Observation error, $\sigma_o^2 = \{0.025, 0.050, 0.100\}$
- Number of realizations, $N = \{100, 1000, 10000\}$

In what follows, for the sake of simplicity, we primarily focus on the results for the case of $N = 1000$, $\kappa = 0.5$ and $\sigma_o^2 = 0.100$. We then summarize and briefly illustrate the effects of varying the three aforementioned parameters.

For a fair comparison, all filters are initialized (at discrete time $k = 0$) with the same Monte Carlo realizations, generated from the optimal Gaussian mixture approximation for the climatological distribution, equation (5). Furthermore, the stochastic forcing applied to the individual ensemble realizations of any one filter is identical to that of the others.

3) Results and Analysis

We show in figure 3 the results obtained for the case of $N = 1000$, $\kappa = 0.5$ and $\sigma_o^2 = 0.100$. Superimposed onto the true solution we show the temporal mean and standard deviation envelope ($\pm\sigma$) for each of three filters, as well as the observations with associated error bars. We have purposely centered the plot about a transition of the ball from one well to the other, as this event is of central interest to us. We have further framed the transition within

a suitable time window that will allow for an appropriate filter evaluation.

Following the assimilation of the first observation, at time $t = 2$, all three filters initially capture the true location of the ball (centered on $x = 1$; from hereon the *positive* well), as represented by their temporal means. They continue to do so until time $t = 20$, at which point the ball transitions, through stochastic diffusion, into the opposite well (centered on $x = -1$; from hereon the *negative* well). This transition is suggested by the observation at time $t = 22$. Both the MEF and the GMM-DO filter transition accordingly, the statistics of the latter settling completely to the negative well following the update at time $t = 26$. The EnKF, on the other hand, fails to recognize this transition despite observations at times $t = 22$ and $t = 26$ suggesting otherwise. In fact, not until time $t = 30$, following three information-rich observations, does it shift its course to the negative well. In what follows, we take a closer look at the mechanics of the three filters. Particularly, we investigate the prior and posterior distributions assigned by each filter as well as their ensemble representations at observation times $t = 18$, $t = 22$ and $t = 26$. We graphically depict this in figure 4.

In panel (a) of figure 4, we show the distributions assigned by each of the three filters (based on their ensemble representations, also shown) at time $t = 18$, at which point the ball has not yet transitioned into the negative well. All three filters correctly assign probability to the positive well, both prior and posterior to the recorded observation. We note, however, that both the GMM-DO filter and the MEF represent their estimates with greater certainty than the EnKF, as indicated by the spread of their respective distributions. This essentially derives from the former two's ability to differentiate between realizations located in separate wells. To illustrate the components of the GMM-DO algorithm, we also display (panel (a), left hand side) the optimal mixture complexity, M, obtained utilizing the BIC on the set of GMM-DO ensemble realizations. At time $t = 18$, this optimal complexity is one (i.e. $M = 1$). This is intuitively supported by previous measurements, having repeatedly suggested the true location of the ball to be in the positive well.

Panel (b) depicts the distributions assigned at time $t = 22$, at which point the ball has transitioned into the opposite well. This is supported by the available observation. Through the EM-BIC procedure, the GMM-DO filter optimally fits a GMM of complexity three to its prior set of ensemble realizations. While difficult to capture visually, one mixture component is centered on the negative well due to the presence of two local realizations (having diffused across from the positive well since time $t = 18$). As a consequence, following the Bayesian update, the GMM-DO filter satisfactorily assigns the majority of its probability to the negative well. This is depicted by its asymmetric bimodal posterior distribution; only few particles remain in the positive well. The MEF largely proceeds in a similar manner. Meanwhile, due to the imbalance of prior variance with measurement uncertainty for the EnKF, a Kalman gain of less than a half results (i.e. $K < 0.5$), which is insufficient to force individual ensemble members across into the negative well. As a consequence, the majority of its particles remain located in the positive well, albeit biased towards the center; the EnKF does not capture the transition.

In panel (c) of figure 4, at time $t = 26$, the majority of realizations of each filter have – since the update at time $t = 22$ – been forced under gravity into the nearest well, displaced from its minimum only by stochastic diffusion. In particular, most of the GMM-DO filter's realizations are now centered on $x = -1$, consequently causing this well to be probabilistically weighted during the GMM fitting procedure. Following the Bayesian update, in which information on the true location of the ball is extracted from the observation, *all* of its realizations are satisfactorily located in the negative well, coinciding with the true location of the ball. The posterior distribution assigned by the MEF agrees with the GMM-DO filter. Again, however, the EnKF's conservative estimate for the Kalman gain is insufficient to completely force particles across into the negative well. Rather, after their update, the ensemble members lie centered on $x = 0$, a state which – due to its instability (see figure 1) – is highly improbable. The EnKF continues in this manner, gradually forcing ensemble members across, and not until time $t = 30$ has it captured the transition (see figure 3).

In figure 5, we briefly investigate the filter sensitivity to each of the three parameters: κ, σ_o^2 and N. We vary each parameter independently (holding the other two fixed) and compare results with these of the standard test case, figure 3. We then generalize the conclusions based on all of our simulations, more of which are presented in Sondergaard (2011).

(i) Filter sensitivity to the number of realizations, N

In figure 5–(i), we reduce the number of ensemble realizations for each of the three filters to $N = 100$. As expected, we see a deterioration in performance for all three filters. In particular, while the statistics in figure 3 of the MEF and the GMM-DO filter settled at times $t = 22$ and $t = 26$, respectively, this settling is now postponed by another observation period, i.e. four time units. The EnKF, however, fails entirely to settle statistically within the time window: while it correctly estimates the true location for the ball following the assimilation at time $t = 34$, it continues to exhibit large variance. In general, based on our experience with varied dynamical systems including results shown in Sondergaard (2011), we found that the GMM-DO filter better handles the task of filtering in the case of fewer

realizations. The more limited the number of ensemble realizations, the more important it is to try to capture the proper shape of the pdfs.

(ii) Filter sensitivity to the diffusion coefficient, κ

In figure 5–(ii), we reduce the diffusion coefficient to $\kappa = 0.4$. While this has little effect on the performances of the GMM-DO filter and MEF, the EnKF again fails to transition during the time interval of focus. This is confirmed in other cases (Sondergaard 2011). As such, for models exhibiting low noise, the approximations made on the prior distribution employed in a Bayesian update become crucial. On the other hand, when the model uncertainty is large, the model noise then dominates the prior pdf and if that noise is Gaussian, a Gaussian update is warranted. An advantage of the GMM-DO filter is that it adapts to all these situations as they occur, in part by updating its shape (its complexity M).

(iii) Filter sensitivity to the observation error, σ_o^2

In figure 5–(iii), we reduce the observation error to $\sigma_o^2 = 0.025$, with marked improvements for all three filters. In fact, here the GMM-DO filter and MEF become indistinguishable, both transitioning at the first suggestion by an observation (at time $t = 22$). On the other hand, when observation errors increase (Sondergaard 2011), we found that the GMM-DO filter significantly outperforms the EnKF. This is because the prior distribution then dominates the update (i.e. the posterior pdf is influenced more by the prior pdf than by the observation pdf) and thus gains importance. As such, when working with systems in which observations are sparse or noisy – and therefore contain useful but relatively limited information – the gain of moving beyond the simple parametric Gaussian distribution becomes substantial.

4) Discussion

For the parameter values investigated in the Double Well Diffusion Experiment, the GMM-DO filter has been shown to outperform the EnKF in its ability to track the transitions of the ball. This enhanced performance is due the former's ability to capture and retain non-Gaussian features during the updates. Moreover, for just a moderate number of ensemble realizations, the performance of the GMM-DO filter is comparable to that of the MEF even though the MEF may be considered tailored to the given test case, i.e. a solution that uses structural information not known by the other two filters.

The MEF shares a number of similarities with the GMM-DO filter, particularly in its use of GMMs for approximating the prior distribution. However, while the MEF enforces its structure through an imposed climatological distribution (see Appendix A), the GMM-DO filter infers this structure in real time by use of the EM algorithm applied to its set of ensemble realizations. As a consequence, the GMM-DO filter is substantially more generic, needing no specification of a climatological pdf and learning only from information contained in the available data. Furthermore, the minimization procedure required by the MEF quickly becomes intractable for systems of increasing dimensionality. In any event, for cases in which the climatological pdf is known or is well approximated with a pdf of low dimensions, the two schemes – GMM-DO filter and MEF – can be merged in a beneficial manner.

We have examined the effects of parameters N, κ and σ_o^2 on the performances of the three filters. With only a few realizations, the GMM-DO filter satisfactorily captures the ball transitions. Specifically, it only requires enough ensemble realizations to sufficiently explore the state space; the optimal fitting of the GMM in turn completes the appropriate assignment of probability. As we increase the number of ensemble realizations, we expect the GMM-DO filter to converge to the optimal Bayes filter. This claim is supported by the results obtained for the case of $N = 10000$ (Sondergaard 2011). For trials with increased observation error, we found the GMM-DO filter substantially more capable than the EnKF. This was also the case for a reduced diffusion coefficient, κ. The extrapolation of these results to ocean and atmospheric data assimilation is interesting. This is because situations with limited number of realizations, limited measurements, or reduced model errors frequently arise, specifically: i) running realistic computational models remain costly and the number of DO modes will remain limited even with distributed computing; ii) the number of platforms and sensors remains small compared to the scales of interest and data errors of representativeness can be significant; and iii), the sustained progress in computational models continues to reduce model errors (e.g. Deleersnijder and Lermusiaux 2008; Deleersnijder et al. 2010). Taken together, these limitations highlight the need for refined data assimilation schemes.

The bimodal distributions of the present experiment is reminiscent of that which arises, for instance, in the dynamics of the Kuroshio (Sekine 1990; Miller 1997; Qiu and Miao 2000; Schmeits and Dijkstra 2001). We consequently hypothesize that many of the conclusions drawn here may be extrapolated to larger systems with more complicated dynamics. This is explored in the following test case on Sudden Expansion fluid and ocean flows.

b. Sudden Expansion Flows

In this section, we examine and discuss the performance and results of the GMM-DO filter in more realistic fluid and ocean dynamics with variable jets and eddies. Specifically, we consider two-dimensional Sudden Expansion flows. In fluid dynamics, such flows have been of considerable interest (Durst et al. 1973; Cherdron et al. 1978; Fearn et al.

1990) and continue to do so. Due to the breaking of symmetries with increasing Reynolds number and the consequent development of at least bimodal statistics, it provides a test case particularly well-suited to the evaluation of data assimilation schemes. In ocean dynamics, such flows are analogous to a uniform barotropic jet (2D flow in the horizontal plane) exiting a narrow strait or an estuary, in the case of a width that is small enough for the effects of the earth rotation (Coriolis acceleration) to be neglected. Such strait or estuary flows are common in the coastal ocean, generally leading to meanders and vortices as the jet exits the constriction.

After describing the test case, we will outline the numerical method used. As with the Double Well Diffusion Experiment, we evaluate the performance of the GMM-DO filter by application of 'identical twin experiments' (Bengtsson et al. 1981; Ide and Ghil 1997a,b): we generate a simulated true solution over a suitable time frame at a Reynolds number that allows for interesting dynamics. Based on sparse and intermittent synthetic measurements of velocities, we ultimately wish to reconstruct the true solution with knowledge only of initial uncertainties. We compare the GMM-DO filter against an ESSE-DO scheme A (the ESSE scheme-A (Lermusiaux and Robinson 1999) combined with the DO equations for priors), using as performance metric the root mean square difference between the true solution and their respective mean fields. We further provide detailed results of the Bayesian update procedure at assimilation times and conclude with an in-depth analysis of their performances.

1) Description of Experiment

It is a well known fact that flows, seemingly symmetric both in initial conditions and geometry, may develop asymmetries with increasing Reynolds numbers, Re; a phenomenon sometimes referred to as the 'Coanda' effect (Fearn et al. 1990). A classical example of such is the development of the *von Karman vortex street* in the wake of a blunt body placed in a uniform flow (Kundu and Cohen 2008). Vortex streets are also ubiquitous in the ocean and atmosphere, especially around islands or other geometric features with rapidly varying aspect ratios. Sudden Expansion flows exhibit similar behavior.

Sudden Expansion flows, here limited to two spatial dimensions, are perhaps most easily understood visually, see figure 6. A developed, symmetric flow of maximum inlet velocity U_{max} in a channel of height h expands into a larger channel of height H, denoting H/h as the expansion ratio. Depending on the Reynolds number,

$$Re = \frac{(h/2)U_{max}}{\nu}, \tag{10}$$

where ν is the kinematic viscosity, a number of phenomena may occur. Experimental results show that, for low Re, the flow is symmetric about the channel centerline, with circulation regions formed at the corners of the expansion (Durst et al. 1973). This is the case depicted in figure 6, in which the flow is described by streamlines. As the Re is increased, instabilities develop, giving rise to asymmetric flows, steady or unsteady. In this paper, we will consider the case of an intermediate Re for which the two-dimensional flow develops asymmetries, yet remains steady and laminar. Specifically, we utilize an expansion ratio of 3 and Re = 250, for which Cherdron et al. (1978) suggested the onset of asymmetries (for the case of three-dimensional flows). We expect results similar to those predicted numerically and verified experimentally by Fearn et al. (1990) for the case of $Re = 140$, as shown in figure 7. The symmetric inlet velocity initially breaks to one side of the centerline. Further downstream, a second region of circulation forces the flow to the opposite side before eventually restoring its initial symmetry. Clearly, the favored direction of breaking depends sensitively on perturbations in the initial conditions, thus giving rise to at least bimodal statistics.

2) Test Procedure

(i) Physical setup

In figure 8, we present the setup for our test case. Placing variables in a non-dimensional form, we let $h = \frac{1}{3}$; $l = 4$; $H = 1$; and $L = 16$. We further impose a *uniform* inlet velocity of $U_{in} = 1$. By conservation of mass and assuming a steady, fully developed Navier-Stokes flow, we obtain the velocity profile at the expansion, $x = 0$ (Kundu and Cohen 2008):

$$U(x = 0, y) = \frac{2}{h^3}\left(\frac{h^2}{4} - y^2\right). \tag{11}$$

and thus a maximum inlet velocity of $U_{max} = U(x = 0, y = 0) = \frac{3}{2}$.

(ii) Initialization of DO decomposition

- Mean Field, $\bar{\boldsymbol{x}}$: the x-component of the mean field velocity is everywhere 1 in the inlet and $\frac{1}{3}$ at any point in the channel, in accordance with continuity; the y-component of the mean field is initially zero everywhere.

- Orthonormal modes, $\tilde{\boldsymbol{x}}_i$: following Sapsis and Lermusiaux (2009), the orthonormal modes are generated by retaining the dominant eigenvectors of the correlation operator $\boldsymbol{C}(\cdot,\cdot)$, defined by:

$$\boldsymbol{C}\left(\mathbf{r}_1, \mathbf{r}_2\right) = \boldsymbol{\mathcal{M}}\left(\mathbf{r}_1, \mathbf{r}_2\right) C(r), \tag{12}$$

where r is the Euclidean distance between points $\mathbf{r}_1$ and $\mathbf{r}_2$, and $\boldsymbol{\mathcal{M}}(\cdot,\cdot)$ is a mollifier function globally taking the value 1 apart from at solid boundaries,

at which it vanishes smoothly. We let $C(r)$ take the form (Lynch and McGillicuddy 2001):

$$C(r) = (1 + 5r + \frac{5^2 r^2}{3})e^{-5r}. \quad (13)$$

We initialize the stochastic subspace, $\boldsymbol{\mathcal{X}}_0$, by retaining the twenty most dominant eigenvectors (i.e. $s = 20$) and hold this size constant throughout the GMM-DO simulations. We note that we have also run cases where s varies in time, as governed by the system dynamics and an adaptive criterion (Sapsis and Lermusiaux 2012).

- Ensemble Members, $\{\boldsymbol{\phi}\} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N\}$: we generate $N = 10000$ subspace realizations, $\boldsymbol{\phi}_i$, from a zero mean, multivariate Gaussian distribution with diagonal covariance matrix. We thus initialize the modes as being uncorrelated with marginal variances proportional to the eigenvalues of the matrix defined by correlation (13).

In general, both N and s evolve but N remains much larger than s to capture the unknown dynamic structure of the pdf in the evolving subspace. This is feasible because the cost of evolving the *scalar* coefficients $\{\boldsymbol{\phi}\}$ is much smaller than that of the evolving the modes $\boldsymbol{\mathcal{X}}_k$.

(iii) Generation of the true solution

We initialize the true solution by generating an arbitrary field according to the aforementioned initial pdf, restricted, however, to the five most dominant modes. Since the true solution is generated from the same statistics as the one imposed by the initial pdf, we ensure that our initial statistics capture the true solution. We note that we have also studied cases where this is not the case and the results remain similar; here, we focus on evaluating assimilation schemes, so we assume the statistics is representative of the unknown truth.

The true solution is propagated deterministically forward in time under the governing Navier-Stokes equations for a total time of $T = 100$, after which the simulation settles into its steady state.

(iv) Observations

We make a total of three sets of measurements of both u- and v-velocities of the true solution at times $T_{obs} = \{50, 70, 90\}$ at the locations indicated in figure 9. The measurements are independent of each other and are made with an observation noise distributed according to a zero-mean Gaussian with variance $\sigma^2_{obs} = 0.1$. This variance is comparable to that expected at the measurement locations during the first assimilation, $T = 50$. We note that other data errors were also employed (not shown here).

(v) Numerical method

Based on Ueckermann et al. (2012), we solve the DO decomposition of the stochastic Navier-Stokes equations numerically, using a flexible and efficient finite volume framework:

- Geometry: The Sudden Expansion geometry is discretized on a uniform, two-dimensional, structured grid of 40 by 30 elements in the x- and y-direction, respectively. A staggered c-grid is utilized to avoid spurious pressure modes.

- Discretization in space: The diffusion operator is approximated using a second order central differencing scheme; the advection operator makes use of a Total Variation Diminishing scheme with a monotonized central limiter (van Leer 1977).

- Discretization in time: For the modes, a first-order accurate, semi-implicit Projection method is employed, where the diffusion and pressure terms are treated implicitly, and the advection is treated explicitly (for details see Ueckermann et al. (2012)). In all cases we limit the time step in accordance with the Courant-Friedrichs-Lewy (CFL) condition. For the scalar coefficients, a Runge-Kutta scheme is employed.

- Boundary conditions: As depicted in figure 8, we assume no-slip boundary conditions at all solid boundaries, while imposing a uniform velocity of 1 across the inlet opening. At the open outlet boundary we restrict the flow by eliminating the first x-derivative of the v-velocities and the second x-derivative of both pressure and u-velocities:

$$\frac{\partial v}{\partial x} = 0; \quad \frac{\partial^2 u}{\partial x^2} = 0; \quad \text{and} \quad \frac{\partial^2 p}{\partial x^2} = 0. \quad (14)$$

3) Results and Analysis

In what follows, we focus on the results of the GMM-DO filter at times $T = \{10, 50, 70, 100\}$. These shows the DO evolution, two assimilation times and the final time, respectively. They allow to appreciate the mechanics of the filter, both prior and posterior to the assimilation of data. We refer to Sondergaard (2011) for complete analyzes every 10 time units.

At each of these times, we display in

- panel (a): the true field, $\boldsymbol{x}^t$;

- panel (b): the mean field, $\bar{\boldsymbol{x}}$;

- panels (c) and (d): the two most dominant modes, $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$;

- panels (e) and (f): the marginal pdfs of the stochastic coefficients Φ_1 and Φ_2, approximated using a kernel dressing method (Silverman 1992);
- panel (g): a scatter plot of the ensemble set, $\{\boldsymbol{\phi}\} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N\}$, projected in the two-dimensional plane defined by the pair of modes: ($\tilde{\boldsymbol{x}}_1$, $\tilde{\boldsymbol{x}}_2$);
- panel (h): a time history of the variances of all the stochastic coefficients, Φ_j; and
- panel (i): a time history of the RMS error of both the GMM-DO filter and the 'ESSE-DO filter', the latter being equivalent to the GMM-DO filter with a forced mixture complexity of one (i.e. M=1).

These series of figures illustrate the way in which the flow and its uncertainties develop, ultimately settling into a steady mean state. It also shows the manner in which the DO equations evolve the state representation and how the GMM update is completed.

At the two assimilation times shown (i.e. $T = \{50, 70\}$), we further display:

- the optimal fitting of the GMM to the set of ensemble realizations within the DO subspace based on the EM algorithm and the BIC; and,
- the spatially local Bayesian updates at each of the measurement points.

From here on, all figures depicting the fluid flow will be described by streamlines overlaid on a color-plot denoting the magnitude of velocity.

(i) $T = 10$

After 10 non-dimensional time units (figure 10), the initial perturbations in the true solution, panel (a), have not yet broken the symmetry of the flow, ultimately causing the appearance of eddies as shown in figure 7. The symmetric mean field, panel (b), consequently still provides a good approximation for the true solution, as quantified by the low RMS error, panel (i). We note that as no data has yet been assimilated, the GMM-DO filter and ESSE-DO scheme provide identical solutions (the two schemes differ mainly in their manner of carrying out the update). The DO statistics (only two of twenty modes shown here) have seemingly evolved little from the initial Gaussian seeding, as represented by the scatter plot, panel (g), and marginal pdfs, panels (e) and (f). The corresponding modes, panels (c) and (d), further give an indication of the initial correlations and probabilistic structures that exist in the flow by combinations with the coefficients (panel g).

(ii) $T = 50$ – *prior distribution*

At the time of the first assimilation of data, dynamics has drastically broken the symmetry of the true solution, as visualized in figure 11–(a). Meanwhile, the DO mean field has remained symmetric, panel (b), thus causing a substantial increase in the RMS error, as shown in panel (i). Fittingly, the filter uncertainty has increased accordingly, witnessed by the inflation of variances of each of the stochastic coefficients in panel (h). Moreover, the marginal distributions of the two most dominant modes, panels (e) and (f), suggest the presence of at least bimodal statistics, reflecting the ambiguity of direction with which the Sudden Expansion flow may break. As such, we expect that the prior DO distribution still statistically encompasses the true solution. Panel (g) also shows that the dynamical system manifold leads to 2d-marginal pdfs with seemingly "harder boundaries", in part due to the limited width of the physical domain and size of the eddies and meanders.

(iii) $T = 50$ – *fitting of GMM*

We show the fitting of the GMM using the EM algorithm to the prior set of ensemble realizations, $\{\boldsymbol{\phi}^f\} = \{\boldsymbol{\phi}^f_1, \ldots, \boldsymbol{\phi}^f_N\}$, in figure 12. Based on the BIC, we determine the optimal mixture complexity to be $M = 29$. We display the one-standard-deviation contours of the 29 mixture components (shown in red) marginalized across pairs of modes (2d joint pdfs), considering only the *four* most dominant modes. We further project the optimal GMM onto the domain of each of the stochastic coefficients, thus giving their respective 1d marginal distribution. We find that the GMM-DO filter successfully captures the complicated, multi-dimensional and nonlinear features present in the set of 10000 DO realizations. For example, the 2d-projections of the GMM clearly identify the localized regions of the subspace (and thus of the state space) where solutions are dynamically possible while the 1d-projections differ little from the marginal distributions predicted by a 1d kernel dressing method (shown in blue). Finally, while not shown in the figure, any other scheme (e.g. ESSE-DO) that fits a single Gaussian to the set of DO realizations undoubtedly results in a severe loss of dynamical information. This is supported by the ESSE-DO performance at later assimilation times.

(iv) $T = 50$ – *local Bayesian update in the data space*

Based on the GMM-DO filter's optimal GMM fit, we display in figure 13 the *local* Bayesian updates projected at each of the observation locations (indicated in panel (a) of figure 11). Shown are the: local true solution, obtained observations, and prior and posterior GMM-DO distributions, computed in exact accordance with Bayes' Law. The local prior distributions are logically more bimodal for u than for v and more uniform near the center of the flow. Overall, they are found to consistently capture the true solution, as particularly evidenced in panel (a). Had we instead used a Gaussian approximation for the prior distribution, the true solution would have been within the

tail of the Gaussian and thus inadequately represented. Of further notice is the shape of the GMM-DO filter's posterior distribution, generally placing greater weight on the mixture components surrounding the true solution. This is again clearest for the Bayesian update in panel (a), in which the left lobe of the bimodal distribution encompasses the true solution.

(v) $T = 50$ – posterior distribution

We show the resulting posterior state description in figure 14. The GMM-DO mean estimate has slightly smaller RMS error than the ESSE-DO scheme mean. However, in accord with the significant data uncertainty, both filters show limited improvements in their estimates for the true solution, as indicated by the small reduction in the RMS error in panel (i). Meanwhile, the stochastic subspace – here just visualized by modes $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$ – has remained unchanged, as explained in part I. Yet, the two filters differ in one crucial aspect: while the posterior statistics of the ESSE-DO scheme-A is Gaussian (not shown), the GMM-DO filter has retained an accurate description of the true statistics of the flow. As such, we expect a superior performance of the latter at the next assimilation step at $T = 70$.

(vi) $T = 70$ – prior distribution

By the time of the second assimilation, figure 15, the prior state estimate of the GMM-DO filter still suggests the presence of at least bimodal statistics, most notably reflected in the marginal distribution of the most dominant stochastic coefficient, panel (e). The RMS error has, since the first assimilation at time $T = 50$, further increased for both filters, the GMM-DO filter providing a slightly superior estimate over the ESSE-DO scheme. The variances of the stochastic coefficients, panel (h), have also slightly increased.

(vii) $T = 70$ – fitting of GMM

In figure 16, we show the fitting of the GMM to the set of realizations in the DO subspace at time $T = 70$. Here, we determine the optimal mixture complexity to be $M = 20$ based on the BIC, reflecting the multi-dimensional structures of the true probability distribution. As before, we note the satisfactory representation of the non-Gaussian features by the GMM, both in one- and multi-dimensional space.

(viii) $T = 70$ – local Bayesian update in the data space

In figure 17, we show the local Bayesian updates projected at each of the observation locations, using the optimal GMM just determined. The local prior GMM-DO pdfs at these data points are now more bimodal than at $T = 50$. This is in part because at $T = 70$, the data points are right at the location of a wide meander. Overall, we again note that the prior pdfs capture the true solution relatively well, especially for the lower data points, see panels (d), (e) and (f), which are in a recirculation eddy (see figure 15). Considering the GMM-DO posterior estimate at data points, the local probability densities have increased where the observations were most expected; this is most clearly visible from increased lobes close to the observations.

(ix) $T = 70$ – posterior distribution

Based on the prior fitting and analysis, we show the global Bayesian update and corresponding posterior distribution at time $T = 70$ in figure 18. We now note how the updated mean field adequately captures the true solution, in particular having determined the direction of the main meander. The RMS error has been reduced, see panel (i). An added strength of the GMM-DO filter is its ability, again, to retain the bimodal structure, as witnessed in panels (e), (f) and (g). As such, it stores the possibility that the flow may in fact have meandered in the opposite direction.

(x) $T = 100$

At the final time, $T = 100$, the true solution is settling into a steady state, exhibiting the characteristic asymmetric flow (e.g. figure 7). This is nearly perfectly captured by the GMM-DO filter. In particular, the RMS error of the GMM-DO mean has been reduced to that at which it started, at time $T = 0$, before the perturbations in the true solution were dynamically evolved. The bimodal structure of the GMM-DO filter, while still present, is much reduced, suggesting an added confidence in its estimate for the mean. This is further supported by the reduced variances of the stochastic coefficients, displayed in panel (h). As such, we conclude that the GMM-DO filter has accurately captured the true solution, exhibiting little uncertainty in its estimate.

4) Discussion

We have examined the application of the GMM-DO filter to fluid and ocean dynamics with variable jets and eddies. The illustration consisted of a two-dimensional Sudden Expansion flow of aspect ratio 3 and Re = 250, at which the true solution becomes asymmetric. Given the sensitivity of the meanders and eddies to the initial perturbations, the corresponding stochastic flow admits complex, far-from-Gaussian distributions and as such is well-suited to evaluate the performance of the GMM-DO filter.

Based on the root mean square errors between the filter mean and true solution, we found the GMM-DO filter to significantly outperform the ESSE-DO scheme A, the latter referring to the GMM-DO filter with a forced mixture complexity of $M = 1$. Specifically, assimilating temporally and spatially sparse measurements, the GMM-DO filter accurately predicted the structure of the true solution at time

$T = 100$ (figure 19). In particular, based on the RMS error, the GMM-DO filter showed a four-fold improvement over the ESSE-DO scheme. (While these results naturally depend on the chosen truth and observations, similar conclusions were drawn based on many other runs not shown here.)

We found the performance of the ESSE-DO scheme to be comparable to that of the GMM-DO filter up until the *second* assimilation step (i.e. $T = 70$), after which the latter showed marked improvements. This is because the GMM-DO filter accurately captures and retains the inherent far-from-Gaussian statistics, both prior *and* posterior to the melding of data, in exact accordance with Bayes' Law. With this, the statistical representation of the state following the first assimilation of data remains accurate, reflected in the successful updates at later assimilation times. At $T = 70$, the dynamics is well captured by the data loca tions and the GMM-DO filter compellingly corrects the mean and uncertainties.

A further strength of the GMM-DO filter is its ability to adapt to the complexity of the subspace realizations at assimilation times. In particular, for the illustrated case, the optimal mixture complexity at the first update was found to be $M = 29$; at the second, $M = 20$; and at the third, $M = 14$; each as determined by the BIC. The accuracy of the fitting procedure is illustrated by figures 12 and 16. This adaptation suggests that Bayes' Law is accurately carried out during the update, neither under- nor over-fitting the true prior pdf.

Finally, by adopting the DO equations, we render computationally tractable the optimal fitting of GMMs. Rather than working in n-dimensional space, the focus is the s-dominant subspace (with $s \ll n$) defined by $\boldsymbol{\mathcal{X}}_k$. The minor loss of information incurred by the reduced dimensionality is more than counterbalanced by the optimal GMM fitting: the complex pdf structures can be captured in the subspace. The subsequent non-Gaussian GMM update in this subspace is then also computationally efficient. Ultimately, the result is an accurate estimation of the posterior pdf, in some sense the central goal of data assimilation.

3. Conclusion

In part II of this two-part paper, we evaluated the performance of the GMM-DO filter in a dynamical systems setting, applying it to (1) the Double Well Diffusion Experiment and (2) Sudden Expansion flows. We illustrated the overall capabilities of the filter including: equations-based and adaptive characteristics; dynamical evolution of the pdfs and DO decompositions; estimation of the GMM parameters in the DO subspace using the EM algorithm and Bayesian Information Criterion (BIC); and, efficient Bayesian updates and the corresponding data impacts. We also compared results to those of contemporary filters including the Ensemble Kalman Filter, Maximum Entropy Filter and ESSE-DO filter. Results clearly showed the advantages of respecting nonlinear dynamics and preserving non-Gaussian statistics.

With the Double Well Diffusion Experiment, we validated the use of the EM algorithm and BIC with GMMs in a filtering context. In particular, we have shown the GMM-DO filter to outperform the Ensemble Kalman filter in its ability to track the transition of the ball from one well to the other. We attribute this skill to the former's ability to capture and retain non-Gaussian features during the data assimilation update. We have further suggested the benefits of adopting the GMM-DO filter over the otherwise novel Maximum Entropy filter; the GMM-DO filter is adaptive, generic and substantially more efficient, learning from information contained in the dynamics and available data. We also examined the sensitivity to variations in the input parameters, finding the GMM-DO filter especially superior for cases of few realizations, sparse and noisy measurements, and moderate model errors – all commonly encountered in ocean and atmospheric applications.

With the Sudden Expansion flows, we showed the properties of the GMM-DO filter in problems of non-trivial dimensionality, specifically flows with dynamic jets and eddies. By focusing on the evolving dominant subspace of the full stochastic state space, the GMM-DO filter enables an otherwise computationally intractable procedure. Specifically, it allows the prediction of prior uncertainties using nonlinear differential equations, the optimal fitting of GMMs to large sets of realizations in the subspace, and the subsequent efficient non-Gaussian update of the GMM pdfs by Bayesian data assimilation. We found the GMM-DO filter to consistently capture the non-Gaussian features of the flow uncertainties, and, critically, preserve them through the Bayesian update. As a consequence, the GMM-DO filter gave a fourfold improvement over the ESSE-DO scheme at the final time step for the given test case. We note that we have obtained similar results with systems of lower (e.g. the Lorenz-95 model) and higher dimensionality (other 2D flows with larger state vectors and more complex features).

A research direction that is now feasible is the study of the dynamics and evolution of the GMMs: they identify the localized nonlinear regions of the stochastic subspace that correspond to dynamically realizable solutions of the uncertain governing equations. Such studies would also be useful for data-model comparisons, adaptive sampling and learning of model errors. If the pdfs of real ocean or atmospheric fields are complex and far-from-Gaussian, we showed that refined data assimilation schemes such as the GMM-DO filter are needed. Should these pdfs be Gaussian, an advantage of the GMM-DO filter is that it automatically adapts to a linear Kalman update. Another obvious next step is the efficient implementation of the GMM-DO filter and its

variations to a full, 4D ocean model, evaluating its performance in a multiscale ocean setting (Haley and Lermusiaux 2010).

Acknowledgments.

We are very thankful to the MSEAS group members, especially to Mr. M.P. Ueckermann for invaluable help with running the Sudden Expansion flows. We are grateful to the Office of Naval Research for support under grants N00014-08-1-1097 (ONR6.1), N00014-09-1-0676 (Science of Autonomy – A-MISSION) and N00014-08-1-0586 (QPE). PFJL is also thankful to Sea Grant at MIT for the "2009 Doherty Professorship in Ocean Utilization" award.

APPENDIX A

The Maximum Entropy Filter

The Maximum Entropy filter (Eyink and Kim 2006) – developed to handle far-from-Gaussian distributions in a dynamical systems setting – is based on a Monte Carlo approach and is applicable to cases in which a climatological distribution for the system of interest exists, is known, and further can be well approximated by a (semi-)parametric distribution that allows for tractable Bayesian inference. For simplicity, in what follows we restrict our attention to univariate distributions. We note, however, that the analysis generically extends to the multivariate case.

We assume that our system is defined such that a climatological distribution, $q_X(x)$, exists and is known. While the method holds for arbitrary distributions, we restrict our attention to a GMM of complexity M:

$$q_X(x) \approx \sum_{m=1}^{M} w_m \times \mathcal{N}\left(x; \mu_m, \sigma_m^2\right). \tag{A1}$$

For a system modeled as a non-periodic Markov Chain with a single recurrent class (Bertsekas and Tsitsiklis 2008), it can be shown that any distribution, $p_X(x)$, forced under the transition kernel (i.e. model) converges to the stationary (i.e. climatological) distribution of the system, $q_X(x)$ (Cover and Thomas 2006). We write this as:

$$\lim_{k\to\infty} D_X(p^k \,||\, q) = 0, \tag{A2}$$

where k is a discrete time index, and $D_X(p \,||\, q)$ denotes the Kullback-Leibler divergence (Kullback 1968) between probability density functions $p_X(x)$ and $q_X(x)$:

$$D_X(p \,||\, q) = \int_{\mathcal{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)}\, dx. \tag{A3}$$

Adopting this framework, an ensemble of forecast realizations (or particle forecasts) is assumed available at the time of a new observation, y. The prior probability density function of the system is then fit to these realizations using an *information projection*,

$$\hat{p}_X^k(x) = \underset{p\in\mathcal{S}_k}{\operatorname{argmin}}\, D_X\left(p \,||\, q\right), \tag{A4}$$

where $\mathcal{S}_k$ denotes a chosen set of distributions consistent with Monte Carlo moment constraints on the set of forecast realizations, $\{x\} = \{x_1, \ldots, x_N\}$. Qualitatively, we understand (A4) as finding the distribution, $p_X(x)$, that satisfies the moment constraints given by $\mathcal{S}_k$ and that is 'closest' to the climatological distribution, $q_X(x)$, having chosen the Kullback-Leibler divergence for measure of distance. A hat is used on the prior pdf, $\hat{p}_X(x)$, to note that it has arisen through an information projection.

For the purposes of tractability, we will concern ourselves only with the first and second moments of the particles, i.e.

$$\begin{aligned}\mathcal{S}_k = \{p_X(x) : E\left[X \mid p_X(x)\right] &= \frac{1}{N}\sum_{i=1}^{N} x_i \equiv \bar{x}_k, \\ \operatorname{var}(X \mid p_X(x)) &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \equiv s_k^2\},\end{aligned} \tag{A5}$$

although, the analysis holds for arbitrary constraints. We note that $\bar{x}_k$ and s_k^2 refer to the sample mean and variance, respectively, at discrete time k. When limiting our attention to the first two moments of the set of realizations, $\{x\} = \{x_1, \ldots, x_N\}$, we will show that the prior distribution, too, takes the form of a GMM.

With $\mathcal{S}_k$ defined as in (A5), it can be shown that $\hat{p}_X^k(x)$ is a member of the following exponential family (Cover and Thomas 2006):

$$\hat{p}_X^k(x) = q_X(x)\frac{e^{\lambda_1 x + \lambda_2 x^2}}{Z(\lambda_1, \lambda_2)}, \tag{A6}$$

with λ_1 and λ_2 chosen such that (A5) is satisfied (i.e. $\lambda_1 = \lambda_1(\bar{x}_k, s_k^2)$ and $\lambda_2 = \lambda_2(\bar{x}_k, s_k^2)$), and where $Z(\lambda_1, \lambda_2)$ is the partition function ensuring that $\hat{p}_X^k(x)$ is a valid distribution. By substituting (A1) into (A5), and completing the square (dropping the explicit notation of time with the understanding the update occurs at discrete time k), it can be shown (Sondergaard 2011) that $\hat{p}_X(x)$ takes the form of a GMM:

$$\hat{p}_X(x) = \sum_{m=1}^{M} \hat{w}_m \times \mathcal{N}(x; \hat{\mu}_m, \hat{\sigma}_m^2) \tag{A7}$$

with parameters

$$\hat{w}_m = \frac{w_m \times e^{-\frac{1}{2\sigma_m^2}\left(\mu_m^2 - \frac{(\mu_m+\sigma_m^2\lambda_1)^2}{1-2\sigma_m^2\lambda_2}\right)}}{Z(\lambda_1,\lambda_2)\sqrt{1-2\sigma_m^2\lambda_2}} \tag{A8}$$

$$\hat{\mu}_m = \frac{\mu_m + \sigma_m^2\lambda_1}{1-2\sigma_m^2\lambda_2} \tag{A9}$$

$$\hat{\sigma}_m^2 = \frac{\sigma_m^2}{1-2\sigma_m^2\lambda_2}. \tag{A10}$$

where w_m, μ_m and σ_m^2 are assumed known (parameters of the fixed background or climatology pdf). Having determined the prior pdf (here, left as a function of $\lambda_1(\bar{x}, s^2)$ and $\lambda_2(\bar{x}, s^2)$), we proceed with the Bayesian update based on observation y. We showed in part I (Sondergaard and Lermusiaux 2012), however, that for a Gaussian observation model,

$$p_{Y|X}(y|x) = \mathcal{N}\left(y; x, \sigma_o^2\right), \tag{A11}$$

with a GMM as prior, the posterior distribution equally takes the form of a GMM. We specifically arrive at

$$p_{X|Y}(x|y) = \sum_{m=1}^{M} \tilde{w}_m \times \mathcal{N}\big(x; \tilde{\mu}_m, \tilde{\sigma}_m\big), \tag{A12}$$

with parameters

$$\tilde{w}_m = \frac{\hat{w}_m \times \mathcal{N}\big(y; \hat{\mu}_m, \sigma_o^2 + \hat{\sigma}_m^2\big)}{\sum_{i=1}^{M} \hat{w}_i \times \mathcal{N}\big(y; \hat{\mu}_i, \sigma_o^2 + \hat{\sigma}_i^2\big)} \tag{A13}$$

$$\tilde{\mu}_m = \hat{\mu}_m + \frac{\hat{\sigma}_m^2}{\sigma_o^2 + \hat{\sigma}_m^2}(y - \hat{\mu}_m) \tag{A14}$$

$$\tilde{\sigma}_m^2 = \frac{\hat{\sigma}_m^2\sigma_o^2}{\hat{\sigma}_m^2 + \sigma_o^2}. \tag{A15}$$

At this point, we generate a new set of realizations, $\{x\} = \{x_1, \ldots, x_N\}$, from the updated GMM and evolve these in time using the governing equation for the system.

REFERENCES

Bengtsson, L., M. Ghil, and E. Kallen, 1981: *Dynamic Meteorology: Data Assimilation Methods*. Springer, New York, 330 pp.

Bertsekas, D. P. and J. N. Tsitsiklis, 2008: *Introduction to Probability*. 2d ed., Athena Scientific.

Cherdron, W., F. Durst, and J. H. Whitelaw, 1978: Asymmetric flows and instabilities in symmetric ducts with sudden expansions. *J. Fluid Mech.*, **84 (1)**, 13–31.

Cover, T. M. and J. A. Thomas, 2006: *Elements of information theory*. Wiley-Interscience, New York, NY, USA.

Deleersnijder, E., V. Legat, and P. Lermusiaux, 2010: Multi-scale modelling of coastal, shelf and global ocean dynamics. *Ocean Dynamics*, **60**, 1357–1359, doi:10.1007/s10236-010-0363-6.

Deleersnijder, E. and P. F. J. Lermusiaux, 2008: Multi-scale modeling: nested grid and unstructured grid approaches. *Ocean Dynamics*, **58**, 335–336, doi:10.1007/s10236-008-0170-5.

Dovera, L. and E. D. Rossa, 2010: Multimodal ensemble Kalman filtering using Gaussian mixture models. *Computational Geosciences*, 1–17.

Durst, F., A. Melling, and J. H. Whitelaw, 1973: Low Reynolds number flow over a plane symmetric sudden expansion. *J. Fluid Mech.*, **64**, 111–128.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *Journal of Geophysical Research-Oceans*, **99 (C5)**, 10 143–10 162.

Eyink, G. L. and S. Kim, 2006: A maximum entropy method for particle filtering. *Journal of Statistical Physics*, **123 (5)**, 1071–1128.

Fearn, R. M., T. Mullin, and K. A. Cliffe, 1990: Nonlinear flow phenomena in a symmetric sudden expansion. *J. Fluid Mech.*, **211**, 595–608.

Frei, M. and H. R. Kunsch, 2011: Mixture ensemble Kalman filters. *Computational Statistics and Data Analysis*.

Haley, P. J. and P. F. J. Lermusiaux, 2010: Multiscale two-way embedding schemes for free-surface primitive-equations in the multidisciplinary simulation, estimation and assimilation system. *Ocean Dynamics*, **60**, 1497–1537, doi:10.1007/s10236-010-0349-4.

Higham, D. J., 2001: An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, **43 (3)**, 525–546.

Houtekamer, P. L., H. L. Mitchell, and L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, **126**, 796–811.

Ide, K., P. Courtier, M. Ghil, and A. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational. . *Meteor. Soc. Japan*, **75**, 181–189.

Ide, K. and M. Ghil, 1997a: Extended Kalman filtering for vortex systems. Part I: Methodology and points vortices. *Dyn. Atmos. Oceans*, **27**, 301–332.

Ide, K. and M. Ghil, 1997b: Extended Kalman filtering for vortex systems. Part II: Rankine vortices and observing-system design. *Dyn. Atmos. Oceans*, **27**, 333–350.

Kullback, S., 1968: *Information Theory and Statistics*. Dover Publications, Inc.

Kundu, P. and I. Cohen, 2008: *Fluid Mechanics*. 4th ed., Academic Press.

Lermusiaux, P. F. J. and A. Robinson, 1999: Data assimilation via error subspace statistical estimation, Part I: Theory and schemes. *Monthly Weather Review*, **127 (8)**, 1385–1407.

Lynch, D. and D. McGillicuddy, 2001: Objective analysis of coastal regimes. *Cont. Shelf Res.*, **21**, 1299–1315.

Miller, R. N., 1997: Data assimilation in nonlinear stochastic models. *'Aha Huliko'a Hawaiian Winter Workshop*, 23–34.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical-systems. *Journal of the Atmospheric Sciences*, **51 (8)**, 1037–1056.

Qiu, B. and W. Miao, 2000: Kuroshio path variations south of japan: Bimodality as a self-sustained internal oscillation. *Journal of Physical Oceanography*, **30**, 2124–2137.

Sapsis, T. and P. F. J. Lermusiaux, 2009: Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D*, **238**, 2347–2360, doi:10.1016/j.physd.2009.09.017.

Sapsis, T. and P. F. J. Lermusiaux, 2012: Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty. *Physica D*, **241**, 60–76, doi:10.1016/j.physd.2011.10.001.

Schmeits, M. J. and H. A. Dijkstra, 2001: Bimodal behavior of the Kuroshio and the Gulf Stream. *Journal of Physical Oceanography*, **31**, 3435–3456.

Sekine, Y., 1990: A numerical experiment on the path dynamics of the Kuroshio with reference to the formation of the large meander path south of Japan. *Deep-Sea Research Part a-Oceanographic Research Papers*, **37 (3)**, 359–380.

Silverman, B., 1992: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Smith, K. W., 2007: Cluster ensemble Kalman filter. *Tellus Series a-Dynamic Meteorology and Oceanography*, **59**, 749–757.

Sondergaard, T., 2011: Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. M.S. thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering.

Sondergaard, T. and P. F. J. Lermusiaux, 2012: Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. Part I: Theory and scheme. *Monthly Weather Review*, sub-judice.

Ueckermann, M. P., P. F. J. Lermusiaux, and T. P. Sapsis, 2012: Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. *Journal of Computational Physics*, doi:10.1016/j.jcp.2012.08.041, in press.

van Leer, B., 1977: Towards the ultimate conservative difference scheme III. upstream-centered finite-difference schemes for ideal compressible flow. *J. Comp. Phys.*, **23**, 263–275.

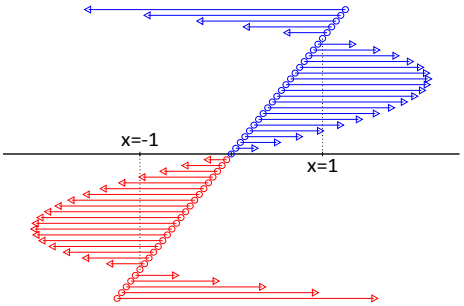

FIG. 1. Forcing Function, $f(x)$. At any location (o) in the horizontal, x, the ball is forced under pseudo-gravity in the direction indicated by the appropriate vector. The magnitude of the vector corresponds to the strength of the forcing. We note that there exists an unstable node at the origin, and two stable nodes at $x = \pm 1$, corresponding to the minima of the wells.

FIG. 2. Climatological distribution and Gaussian mixture approximation for $\kappa = 0.40$. In accordance with intuition, the distributions are bimodal, appropriately centered on the minima of each of the two wells.

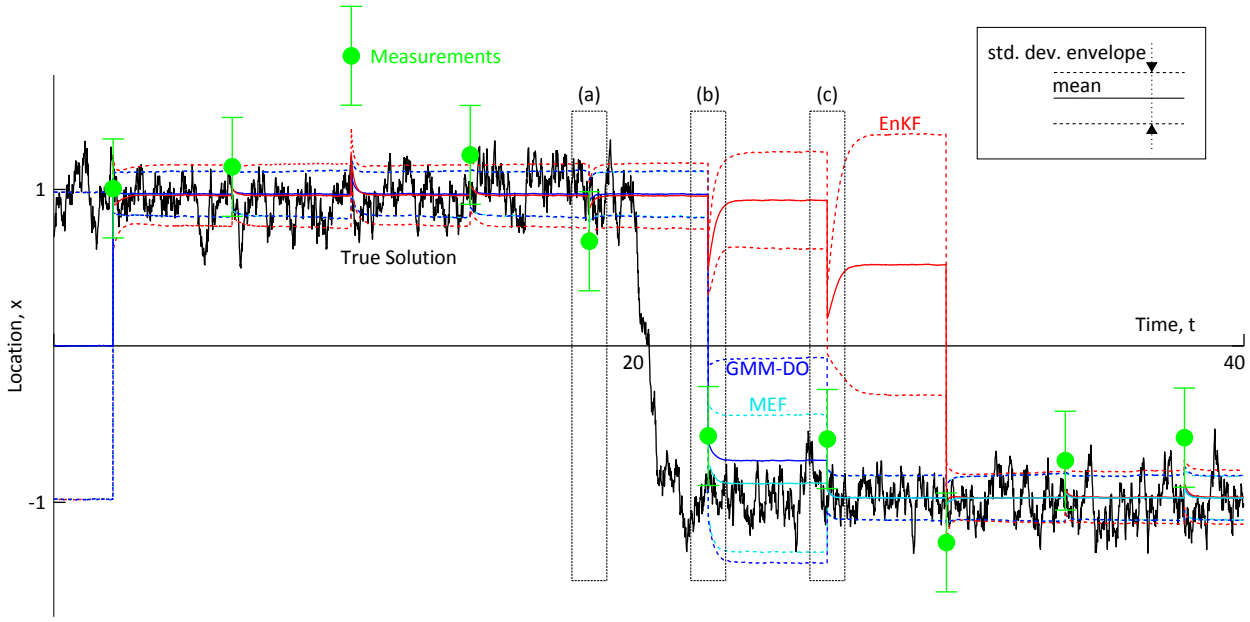

FIG. 3. Results of the three filters – GMM-DO filter (blue); MEF (yellow); and EnKF (red) – for the case of $N = 1000$, $\kappa = 0.5$ and $\sigma_o^2 = 0.100$. The black curve denotes the true solution for the location of the ball, with the green markers representing observations with associated standard deviation envelope. The highlighted instances – (a), (b) and (c) – are examined in detail in figure 4.

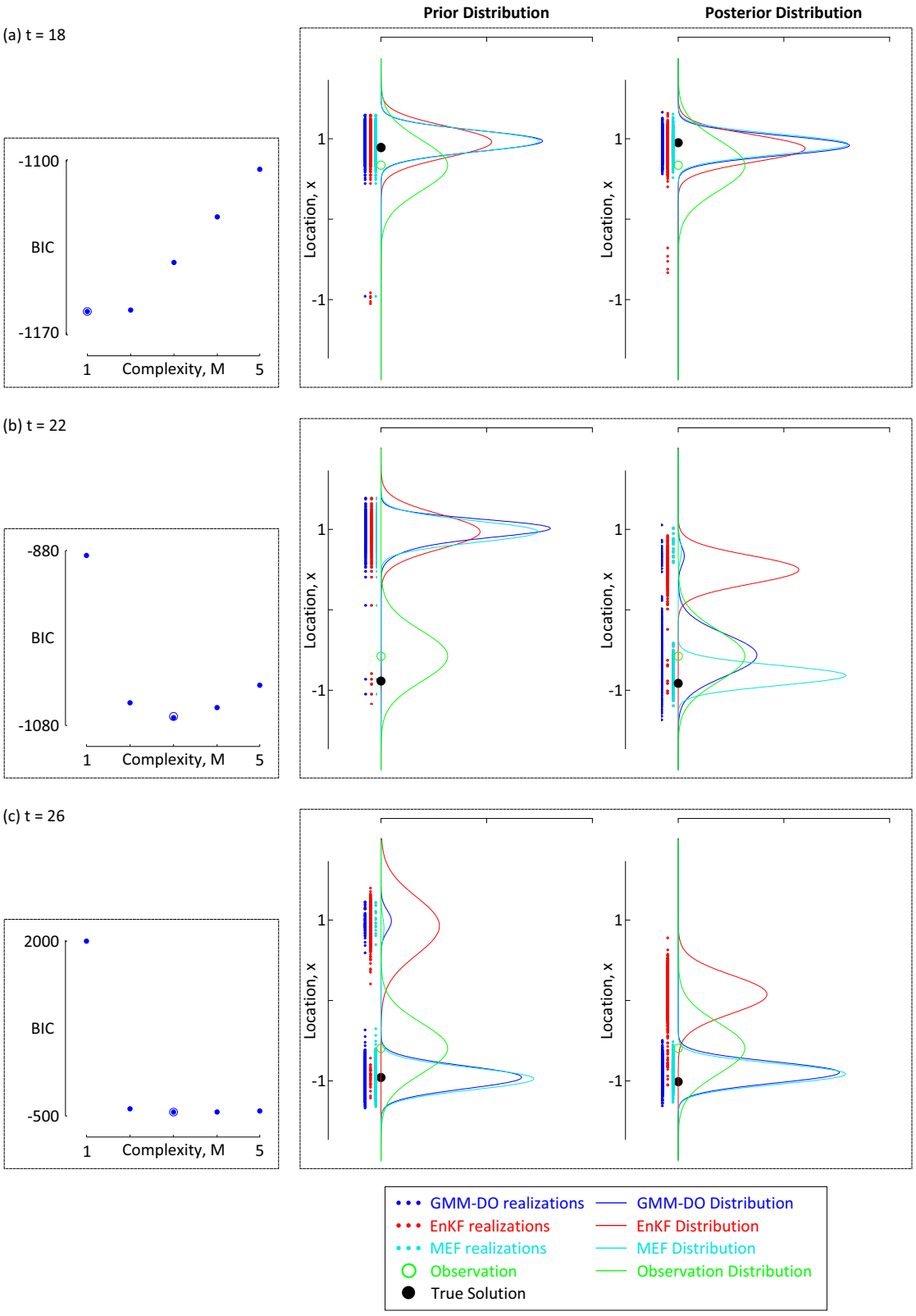

FIG. 4. An analysis of the prior and posterior distributions assigned by each of the three filters – the GMM-DO filter, MEF and EnKF – at times $t = 18$, $t = 22$ and $t = 26$ for the standard test case, displayed in figure 3, with parameters $N = 1000$, $\kappa = 0.4$ and $\sigma_o^2 = 0.025$. We also display the optimal mixture complexity for the prior distribution of the GMM-DO filter, as obtained by application of the EM algorithm and the BIC.

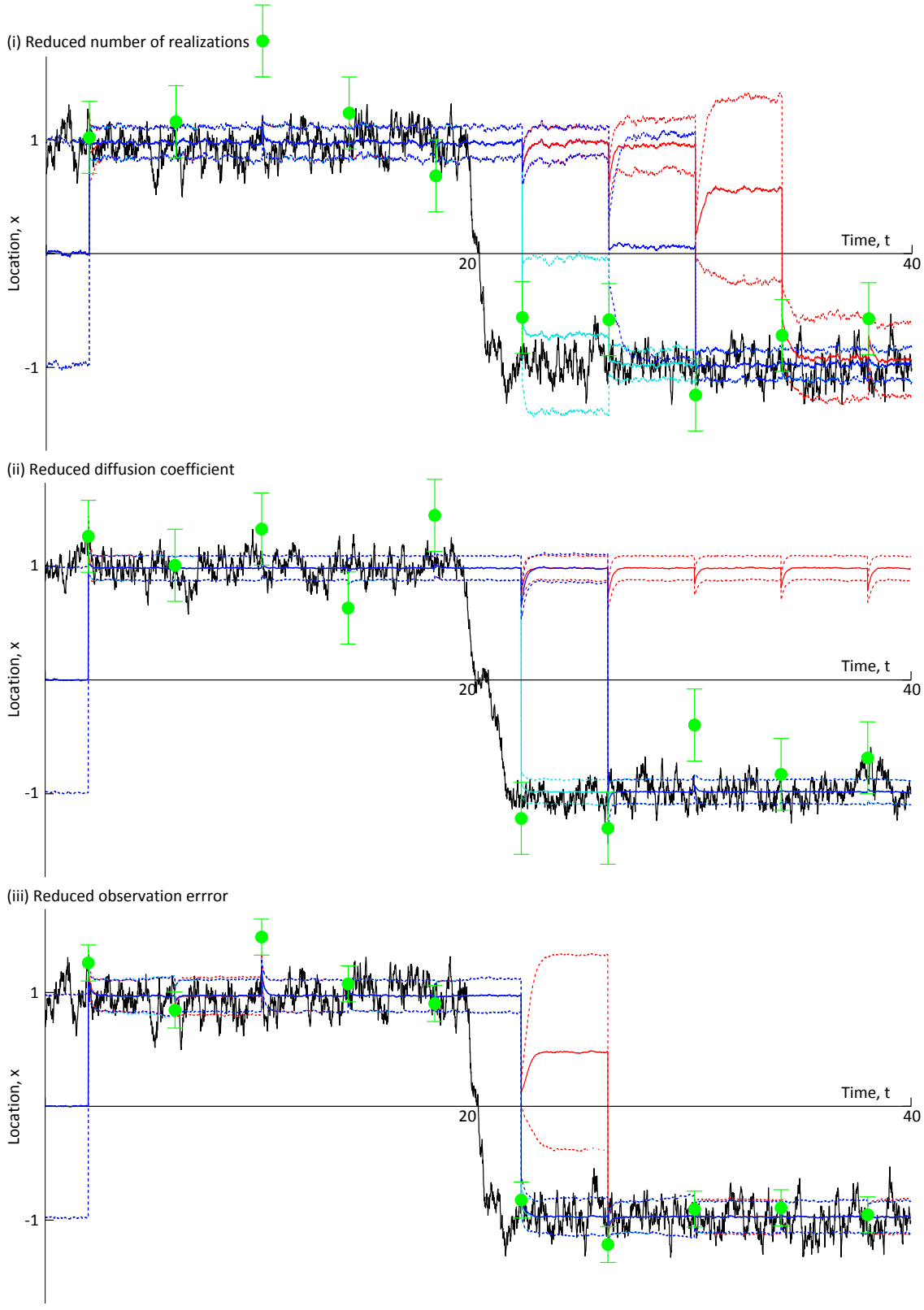

FIG. 5. A study of the filter sensitivities to variations in each of the three parameters κ, σ_o^2 and N. In panel (i), we reduce the number of ensemble realizations to $N = 100$, while holding the other two parameters constant. In panel (ii), we reduce the diffusion coefficient to $\kappa = 0.4$, while in panel (iii), we reduce the observation error to $\sigma_0^2 = 0.025$. These results are to be compared with the standard test case, shown in figure 3, with parameters $N = 1000$, $\kappa = 0.5$ and $\sigma_o^2 = 0.025$.

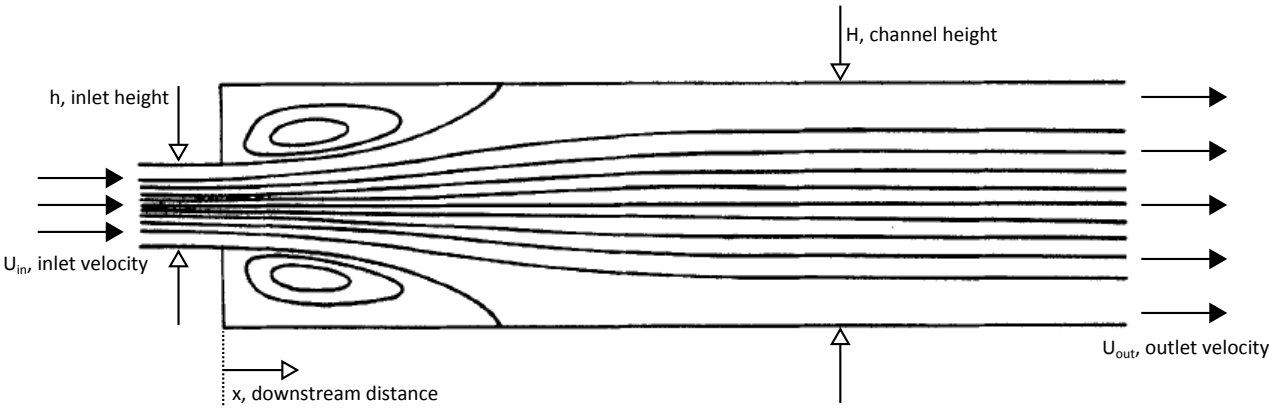

FIG. 6. Setup of the Sudden Expansion test case (Fearn et al. 1990).

FIG. 7. Calculated streamlines at $Re = 140$. (Fearn et al. 1990).

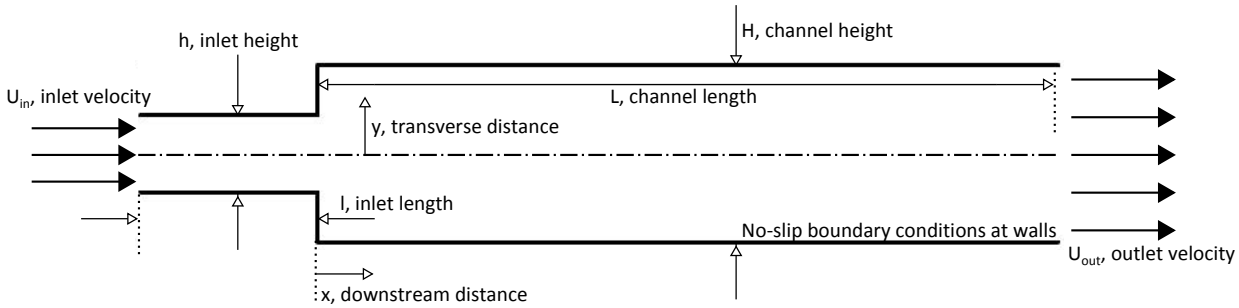

FIG. 8. Sudden Expansion Test Setup.

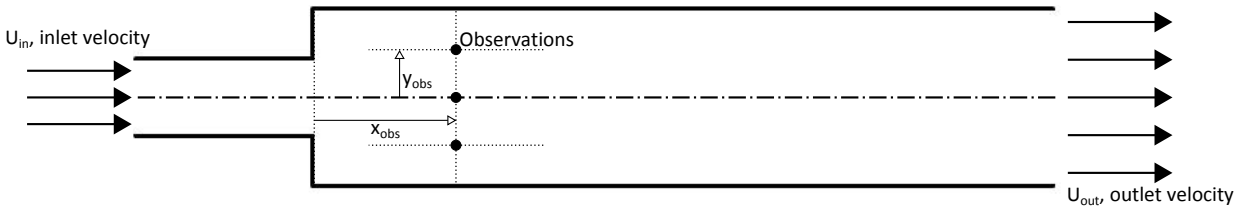

FIG. 9. Observation Locations: $(x_{obs}, y_{obs}) = \{(4, -\frac{1}{4}),\ (4, 0),\ (4, \frac{1}{4})\}$.

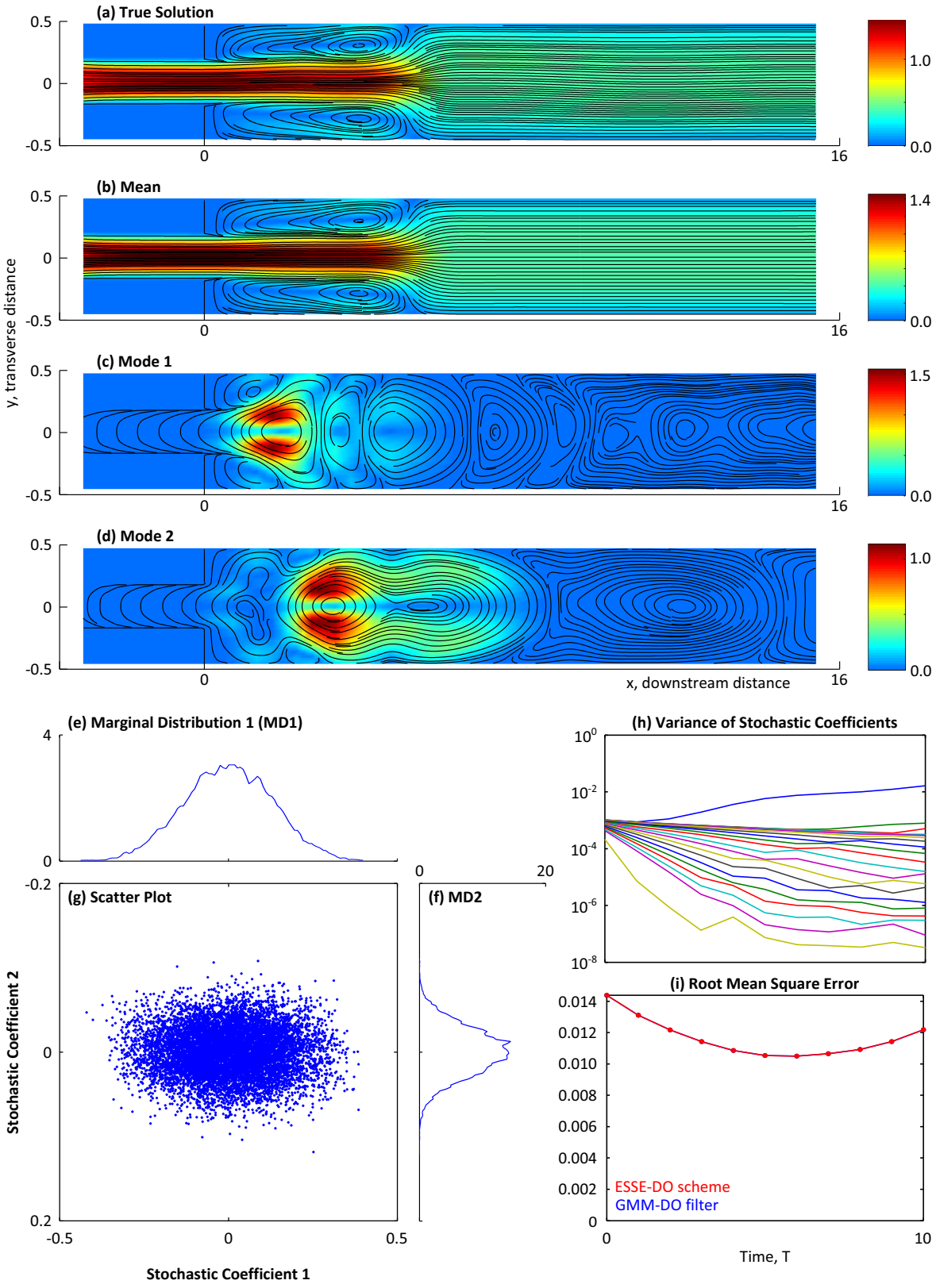

FIG. 10. Time $T = 10$: True solution; condensed representation of DO decomposition including the DO mean field, first two DO modes and their marginal pdfs, and the variance of the coefficients from time 0 to 10; and finally, root mean square errors (between the DO mean and the deterministic truth), also up to time $T = 10$.

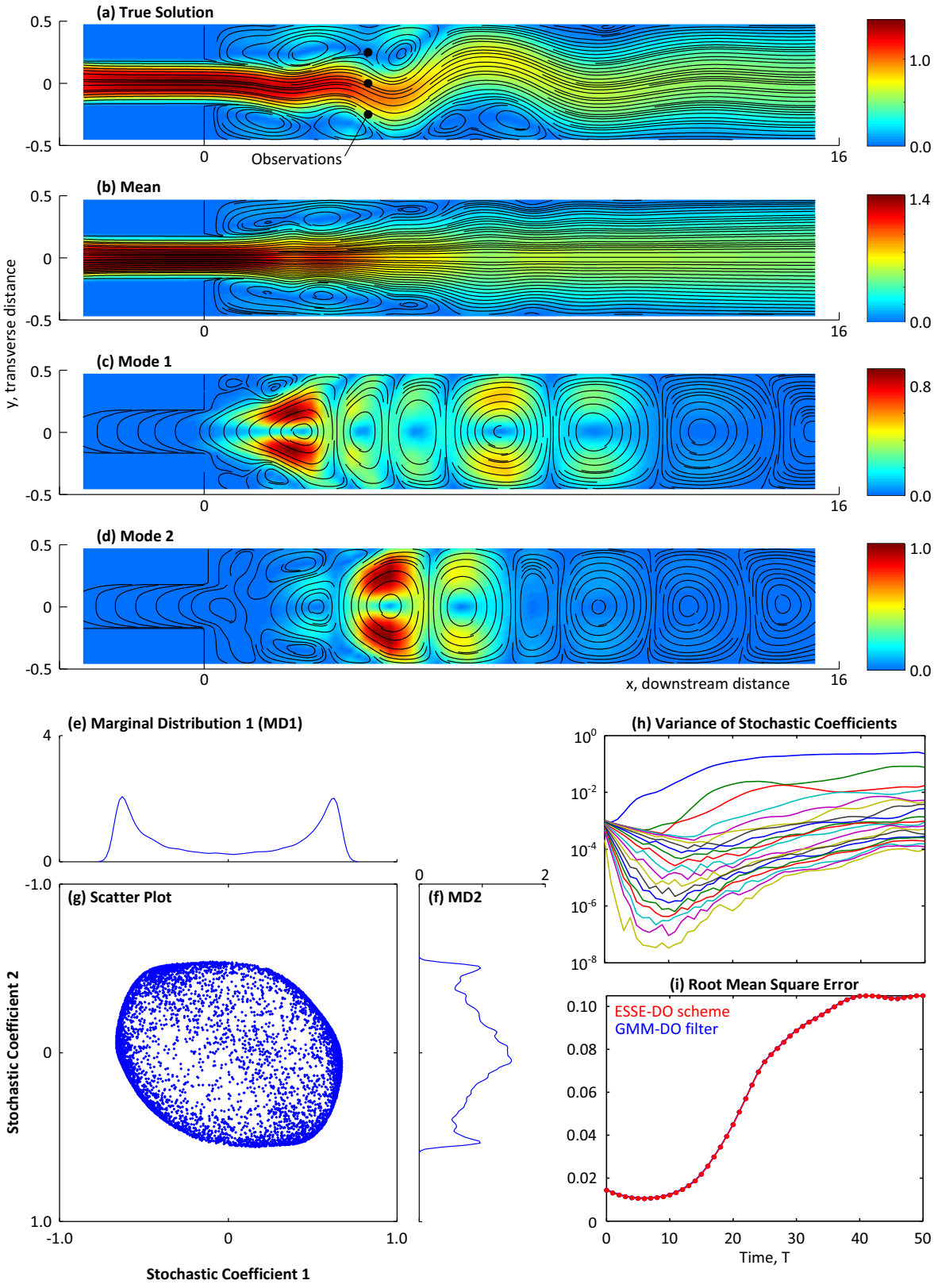

FIG. 11. As Fig. 10, but for the prior DO decomposition at time $T = 50$ which is the first assimilation time step.

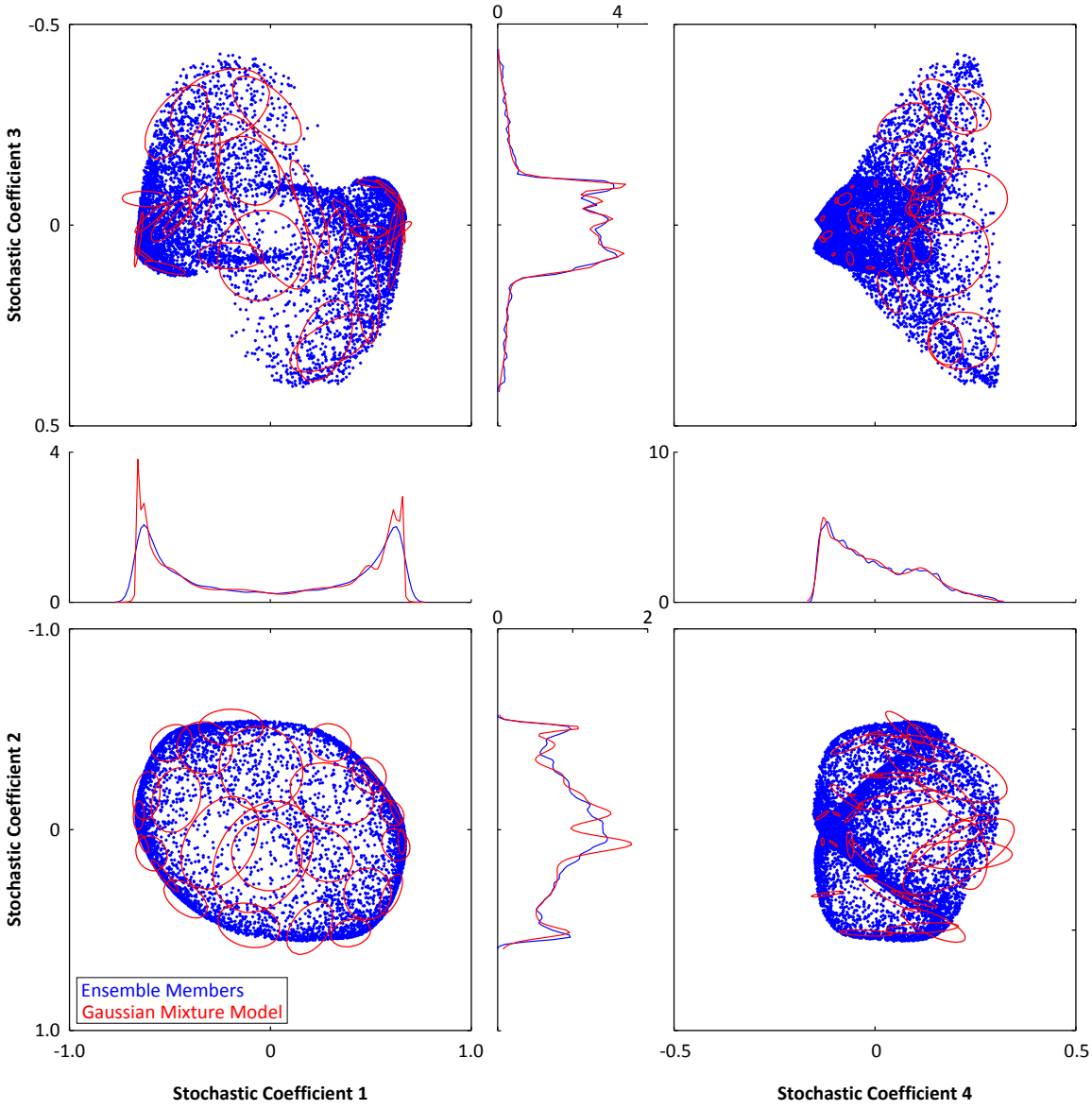

FIG. 12. Marginal prior distributions (in one and two dimensions) for the DO stochastic coefficients at time $T = 50$ (first assimilation time step) focusing on the first four DO modes only. The complete prior distribution is obtained by fitting the GMM using the EM algorithm and BIC to the ensemble of realizations in the DO subspace. This leads to an optimal complexity estimate of $M = 29$. The GMM scalar marginals (1d pdf) and planar marginals (2d joint pdfs) are illustrated by the red plain curves and red standard ellipses, respectively. The DO realizations are plotted in blue, using a kernel dressing scheme in 1d and a scatter plot in 2d. We note that the GMM marginals shown are only projections of a complex GMM distribution of 20d i.e. $s = 20$.

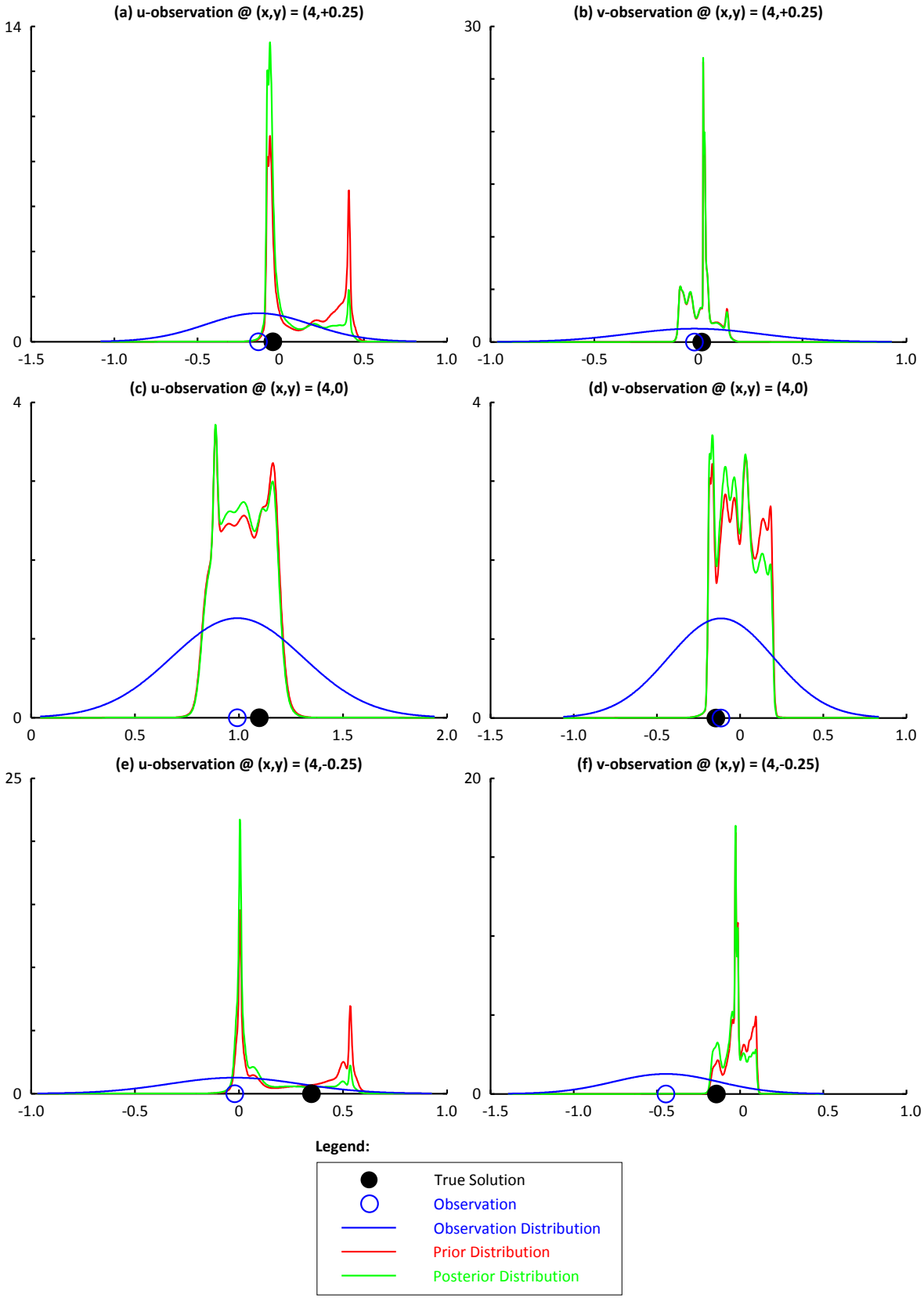

FIG. 13. Data space at time $T = 50$: True solution; observation and its associated Gaussian distribution; and the prior and posterior distributions at the observation locations, all at that time.

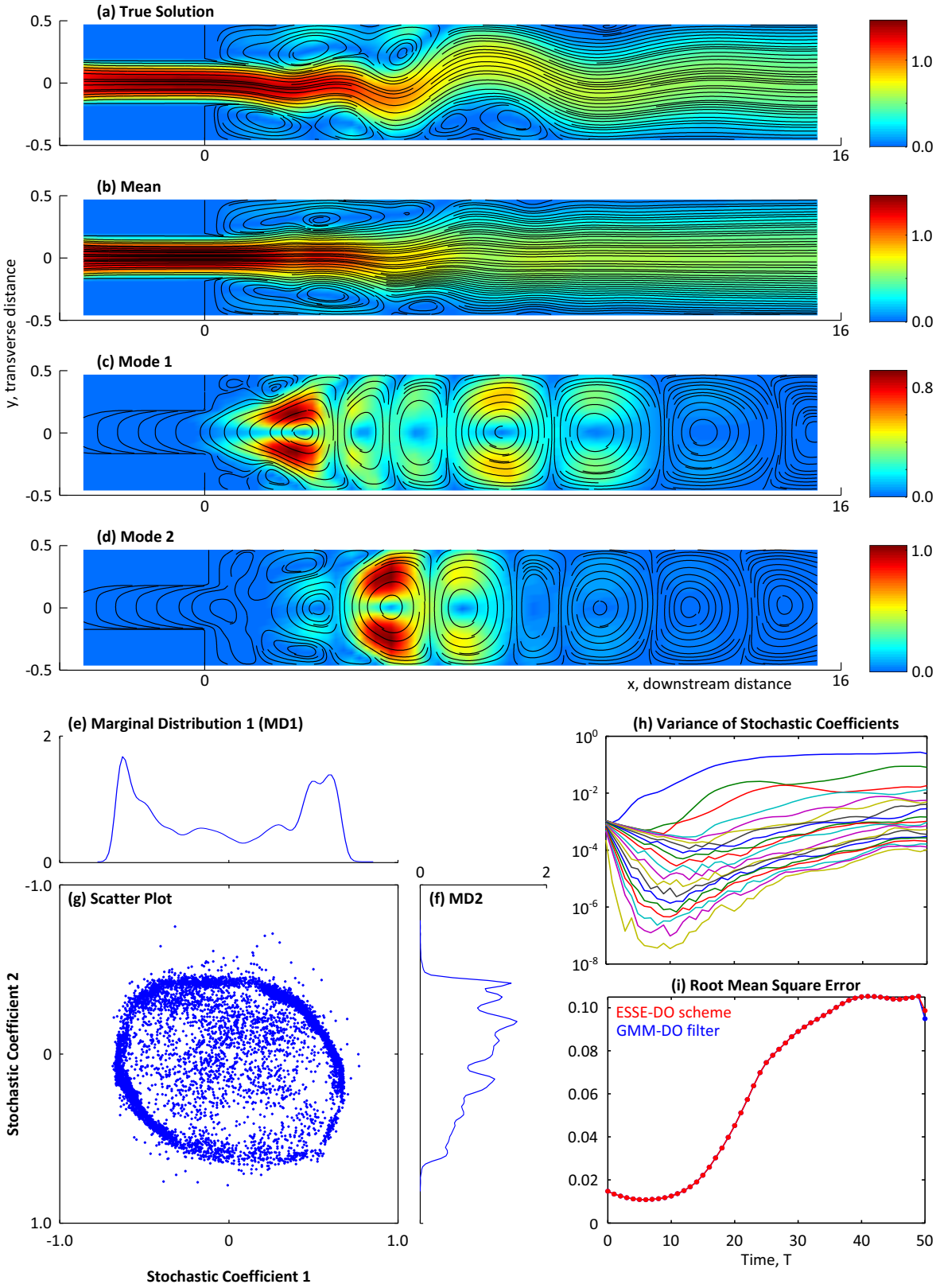

FIG. 14. As Fig. 11, but for the posterior GMM-DO estimates at time $T = 50$.

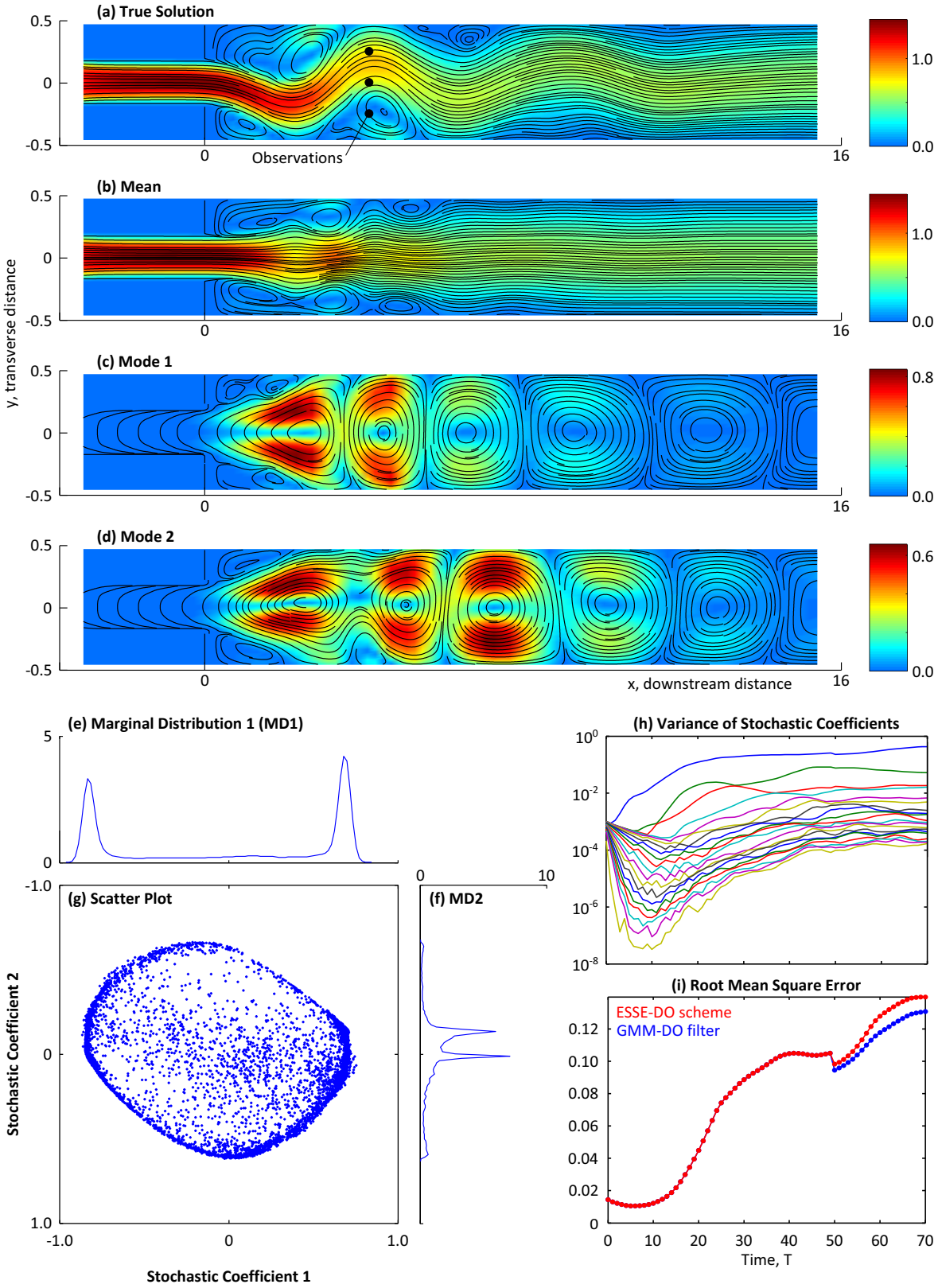

FIG. 15. As Fig. 11, but for the prior GMM-DO estimates at time $T = 70$.

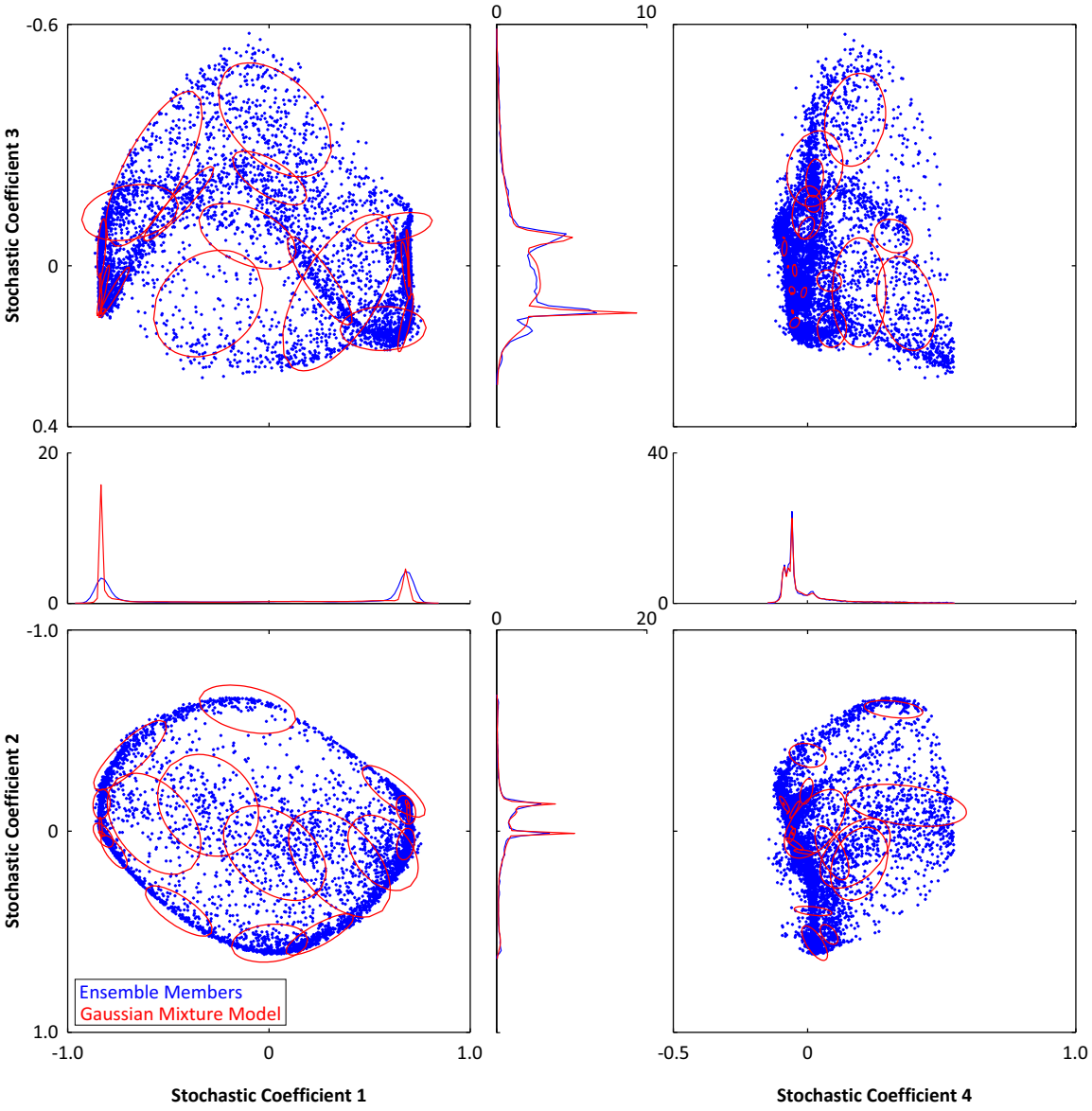

FIG. 16. As Fig. 12, but at the second GMM-DO assimilation time $T = 70$: when compared to the priors at $T = 50$, the optimal complexity is now found to be a bit smaller $M = 20$, the 2d marginals of the stochastic coefficients remain complex, while the 1d marginals are either more bimodal (coefs. 1 to 3) or relatively unimodal (coef. 4).

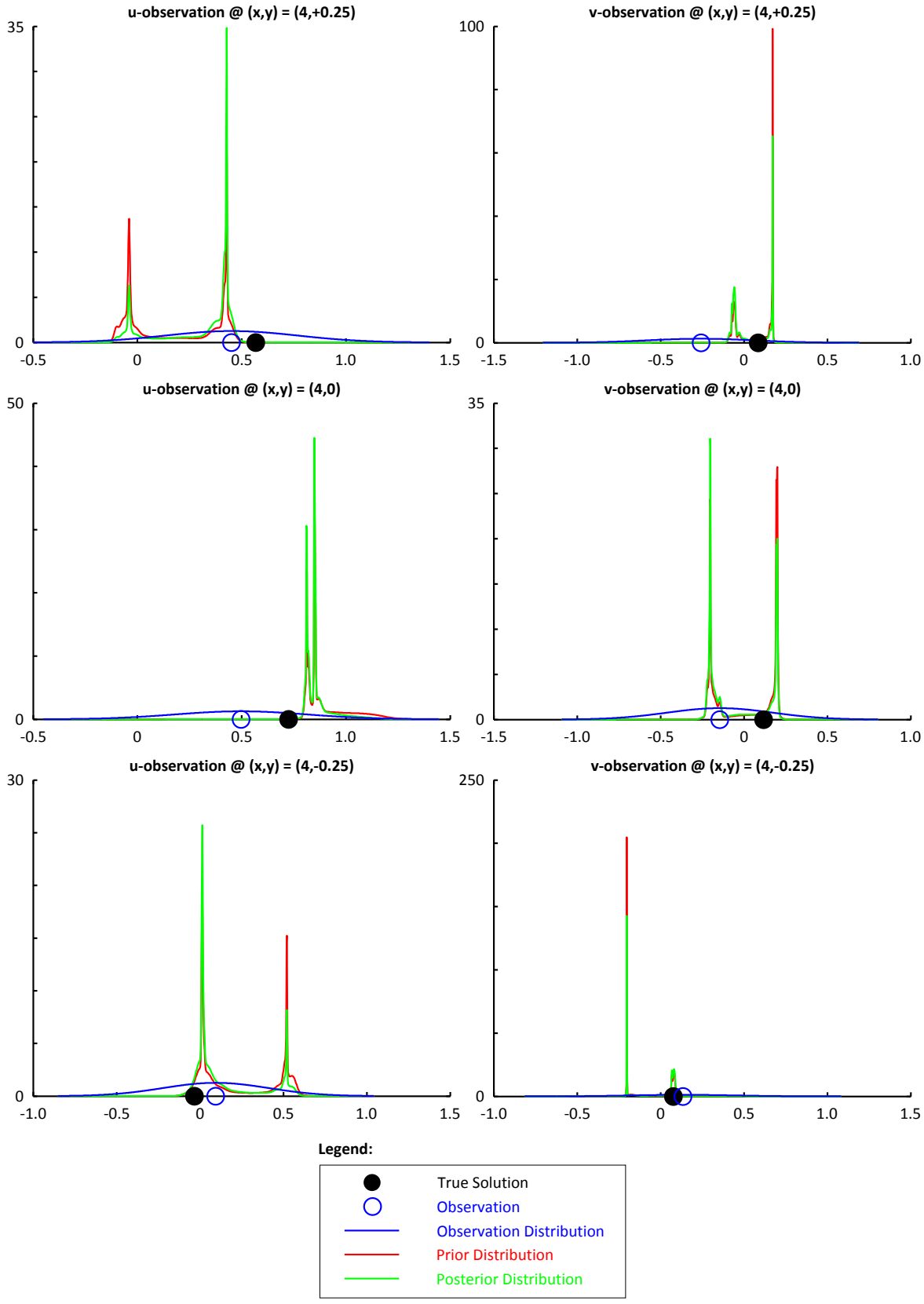

FIG. 17. As Fig. 13, but for the data space at time $T = 70$: when compared to the local updates at $T = 50$, the local GMM-DO pdfs are more bimodal.

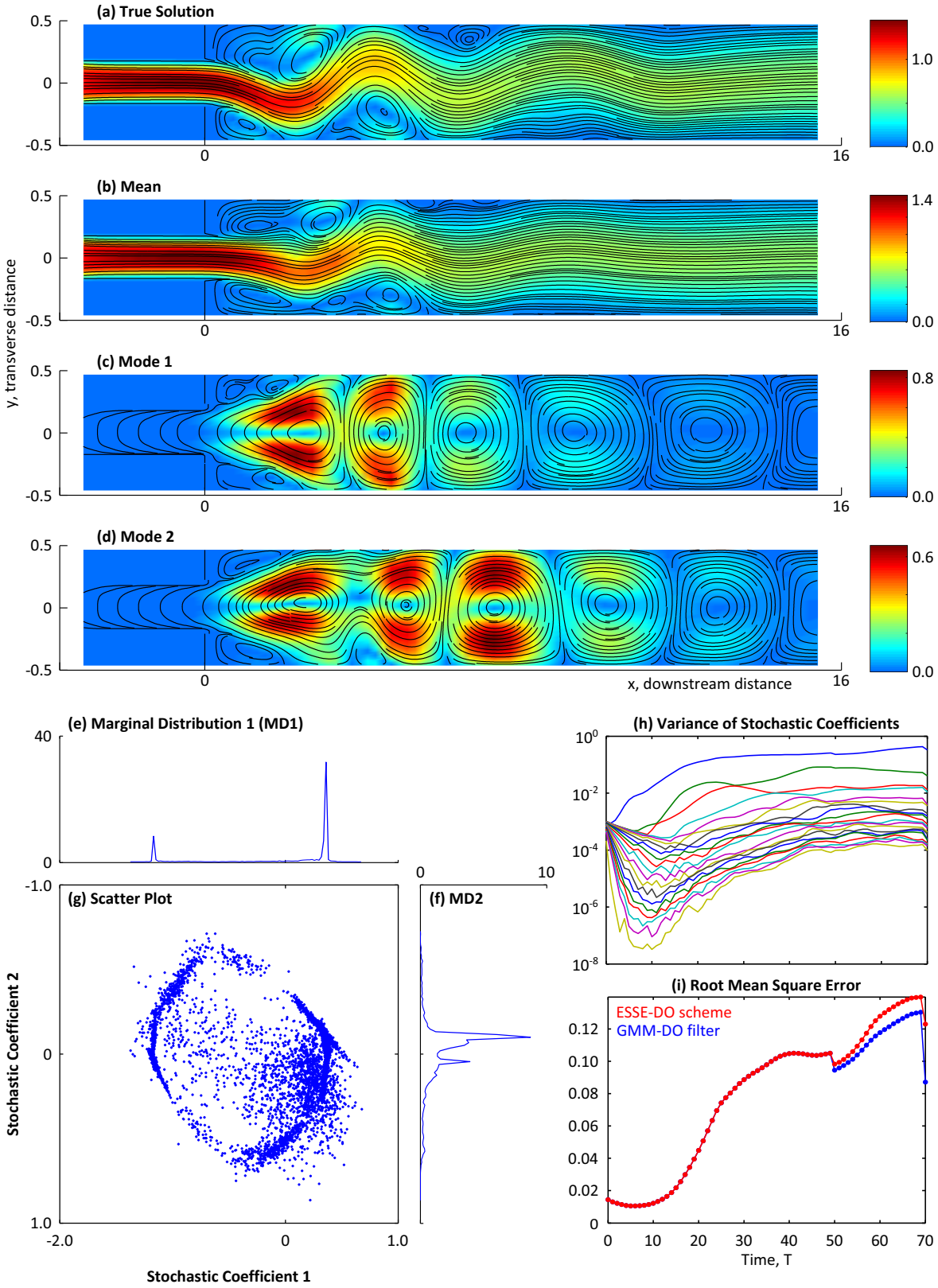

FIG. 18. As Fig. 15, but for the posterior GMM-DO estimates at time $T = 70$.

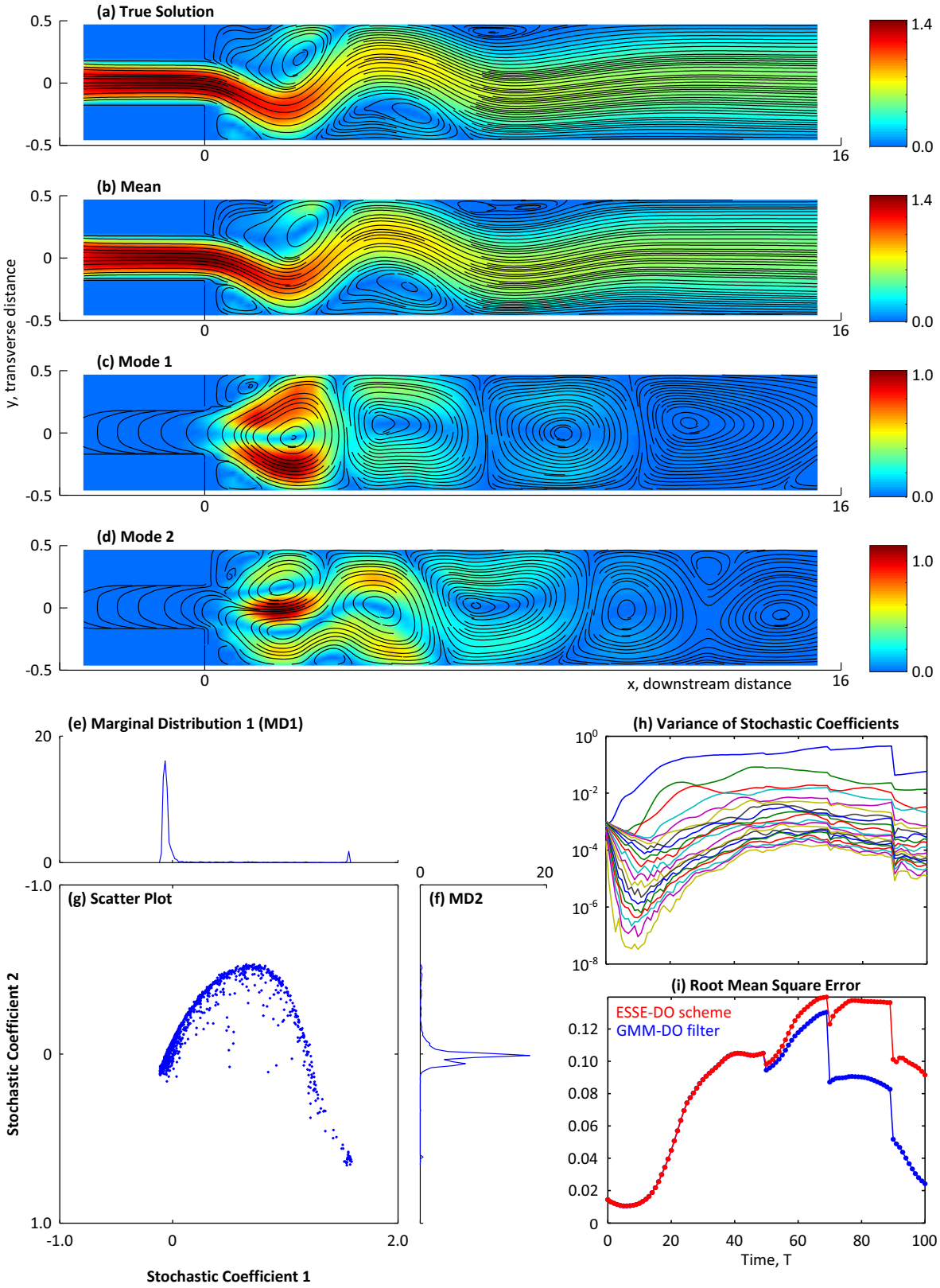

FIG. 19. As Fig. 15, but for the prior GMM-DO estimates at time $T = 100$.

TABLE 1. Notation relevant to the GMM-DO filter. (While we have primarily adopted notation specific to probability theory, information theory and estimation theory, where possible we also utilize the notation advocated by Ide et al. (1997).)

Descriptors		
$(\cdot)^f$		forecast
$(\cdot)^a$		analysis

Scalars		
i	$\in \mathbb{N}$	stochastic subspace index
j	$\in \mathbb{N}$	mixture component index
k	$\in \mathbb{N}$	discrete time index
n	$\in \mathbb{N}$	dimension of state vector
p	$\in \mathbb{N}$	dimension of observation vector
q	$\in \mathbb{N}$	dimension of dominant stochastic subspace
r	$\in \mathbb{N}$	realization index
s	$\in \mathbb{N}$	dimension of stochastic subspace
M	$\in \mathbb{N}$	complexity of Gaussian Mixture Model
N	$\in \mathbb{N}$	number of Monte Carlo members
Φ_i	$\in \mathbb{R}$	random variable describing the pdf for orthonormal mode $\tilde{\boldsymbol{x}}_i$

Vectors		
$\boldsymbol{X}$	$\in \mathbb{R}^n$	state (random) vector
$\boldsymbol{x}$	$\in \mathbb{R}^n$	state realization
$\tilde{\boldsymbol{x}}_i$	$\in \mathbb{R}^n$	DO mode i: dynamically orthonormal basis for stochastic subspace
$\bar{\boldsymbol{x}}$	$\in \mathbb{R}^n$	mean state vector
$\boldsymbol{Y}$	$\in \mathbb{R}^p$	observation (random) vector
$\boldsymbol{y}$	$\in \mathbb{R}^p$	observation realization
$\bar{\boldsymbol{x}}_j$	$\in \mathbb{R}^n$	mean vector of mixture component j in state space
$\boldsymbol{\mu}_j$	$\in \mathbb{R}^s$	mean vector of mixture component j in stochastic subspace
$\boldsymbol{\Phi}$	$\in \mathbb{R}^s$	multivariate random vector, $[\Phi_1 \ \dots \ \Phi_s]$
$\boldsymbol{\phi}$	$\in \mathbb{R}^s$	realization residing in stochastic subspace
$\boldsymbol{\Upsilon}$	$\in \mathbb{R}^p$	observation noise (random) vector
$\boldsymbol{\upsilon}$	$\in \mathbb{R}^p$	observation noise realization

Matrices		
$\boldsymbol{P}$	$\in \mathbb{R}^{n \times n}$	covariance matrix in state space
$\boldsymbol{\Sigma}_j$	$\in \mathbb{R}^{s \times s}$	covariance matrix of mixture component j in stochastic subspace
$\boldsymbol{P}_j$	$\in \mathbb{R}^{n \times n}$	covariance matrix of mixture component j in state space
$\boldsymbol{R}$	$\in \mathbb{R}^{p \times p}$	observation covariance matrix
$\boldsymbol{H}$	$\in \mathbb{R}^{m \times n}$	(linear) observation model
$\boldsymbol{\mathcal{X}}$	$\in \mathbb{R}^{n \times s}$	matrix of s DO modes, $[\tilde{\boldsymbol{x}}_1 \ \dots \ \tilde{\boldsymbol{x}}_s]$
$\{\boldsymbol{\phi}\}$	$\in \mathbb{R}^{s \times N}$	set of subspace ensemble realizations, $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N\}$
$\{\boldsymbol{x}\}$	$\in \mathbb{R}^{n \times N}$	set of state space ensemble realizations, $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$