Non-Gaussian Data Assimilation with Stochastic PDEs: Visualizing Probability Densities of Ocean Fields?

Pierre F.J. Lermusiaux
T. Sondergaard, T. Sapsis, M. Ueckermann, T. Lolla, P.J. Haley and W.G. Leslie

Mechanical Engineering, MIT

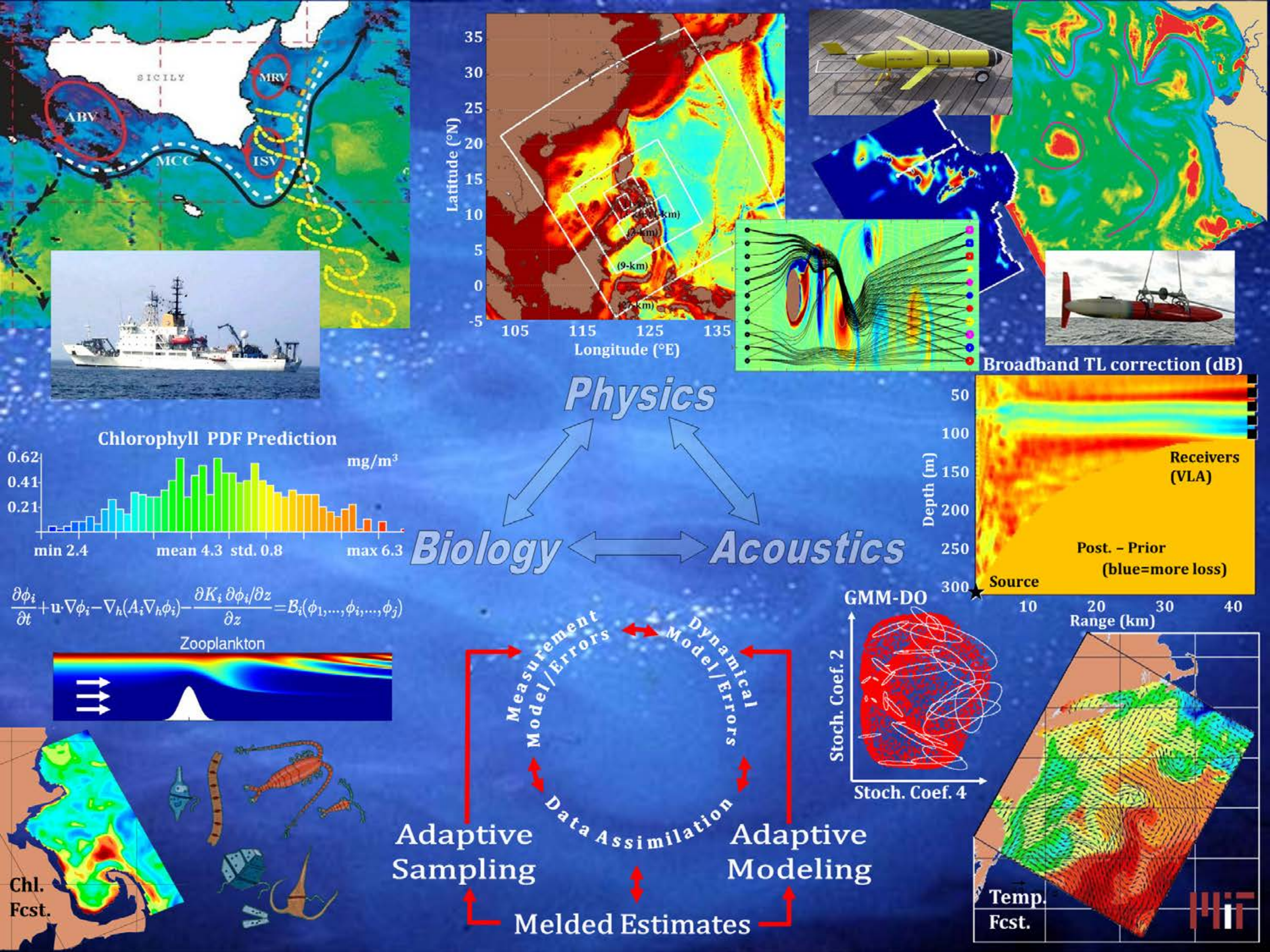
Multidisciplinary Simulation, Estimation and Assimilation Systems (MSEAS)
http://mseas.mit.edu/

- **Introduction**
- **Grand Challenges in Ocean/Earth-System Sciences & Engineering**
  - **Prognostic Equations for Stochastic Fields of Large-Dimension**
  - **Non-Gaussian Data Assimilation (here with DO eqns and GMM-algorithm)**
- **Conclusions**

Thanks to MIT, ONR and NSF

Peter Muller and Frank Henyey, 1997. Workshop Assesses Monte-Carlo Simulations in Oceanography

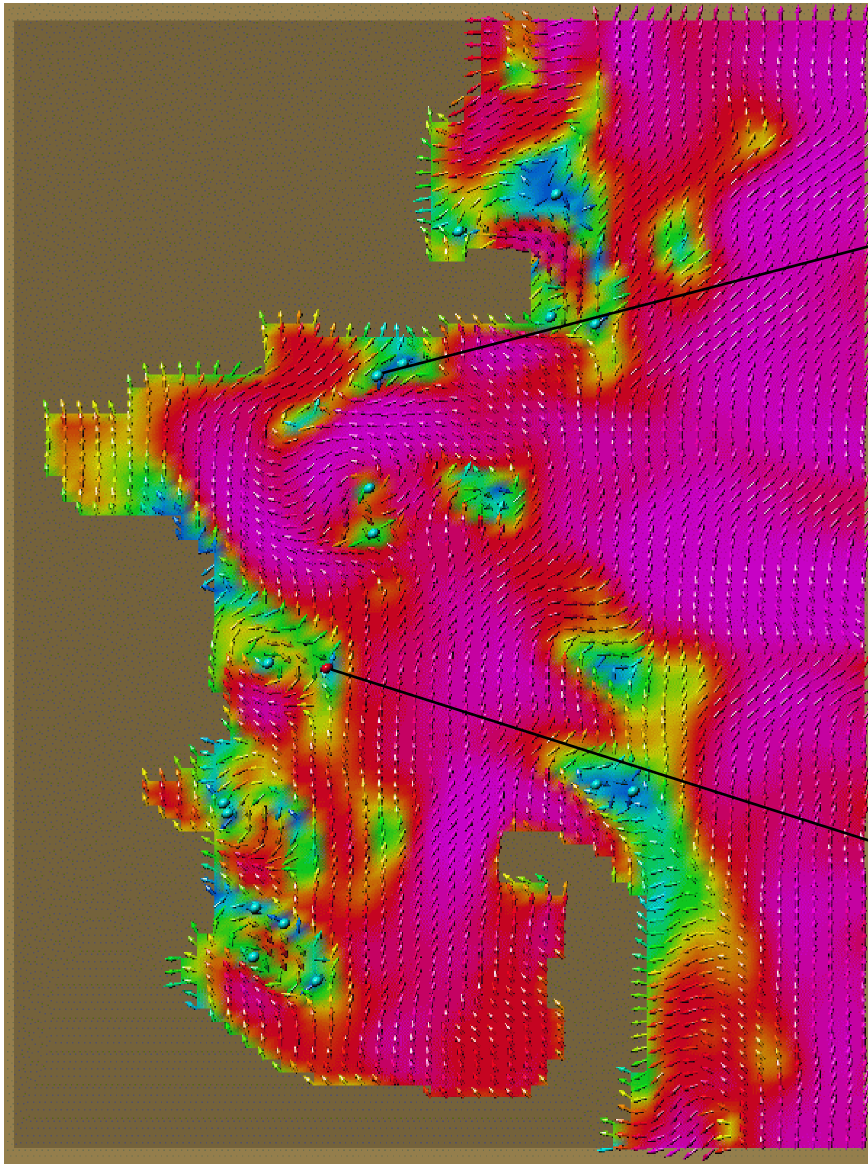
- "*Oceanographers enthusiastically integrate global ocean circulation models in conjunction with atmospheric models over periods of thousands of years in order to asses future climate states – without actually knowing the skill of their ocean models …*"
- "*The case was made at the workshop that … randomness be included in the dynamical equations …. "*
- *"… an oceanic circulation model is obtained by averaging and approximating the Navier-Stokes equations …. sub-grid-scales cannot be parameterized in terms of local mean flow quantities …Thus, the oceanic general circulation should be regarded as a stochastic problem described by a set of stochastic PDEs*."
- " *… the vast majority of the data assimilation schemes … were derived and validated for linear systems with Gaussian noise …. The nonlinearity might actually lessen the dimensionality problem since the motion of the system might become confined … to some lower-dimensional subset of the full state space …*"

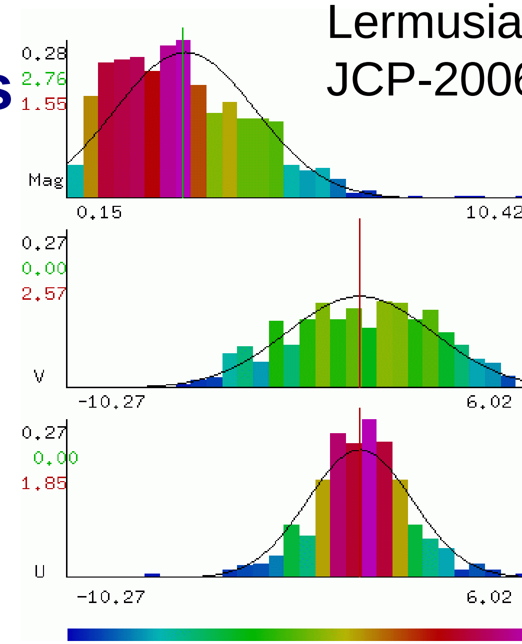


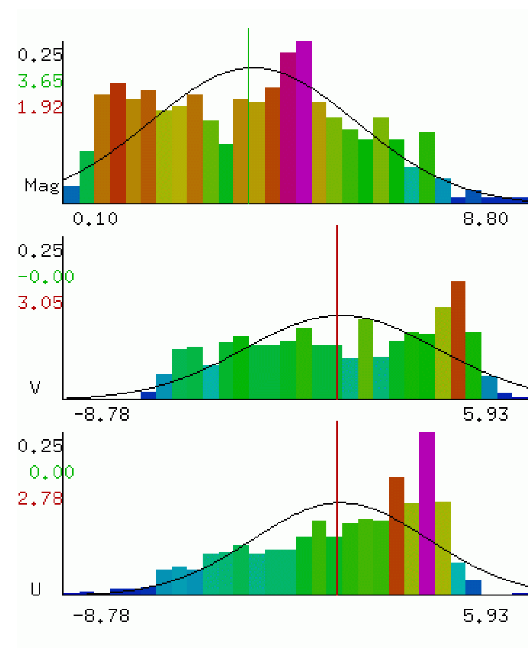
SICILY
MRV
ABV
ISV
MCC
35
30
25
20
15
10
5
0
-5
Latitude (°N)
105
115
125
135
Longitude (°E)
(9-km)
Broadband TL correction (dB)
Physics
Biology
Acoustics
Chlorophyll PDF Prediction
mg/m³
0.62
0.41
0.21
min 2.4
mean 4.3 std. 0.8
max 6.3
50
100
150
200
250
300
Depth (m)
Receivers
(VLA)
Post. – Prior
(blue=more loss)
Source
10
20
30
40
Range (km)
$$\frac{\partial \phi_i}{\partial t} + \mathbf{u}\cdot\nabla\phi_i - \nabla_h(A_i \nabla_h \phi_i) - \frac{\partial K_i\, \partial\phi_i/\partial z}{\partial z} = \mathcal{B}_i(\phi_1, \ldots, \phi_i, \ldots, \phi_j)$$
Zooplankton
GMM-DO
Stoch. Coef. 2
Stoch. Coef. 4
Measurement Model/Errors
Dynamical Model/Errors
Data Assimilation
Adaptive Sampling
Adaptive Modeling
Melded Estimates
Chl. Fcst.
Temp. Fcst.

Interactive Visualization and Targeting of pdf's – Time Dependent Fields

Lermusiaux, JCP-2006

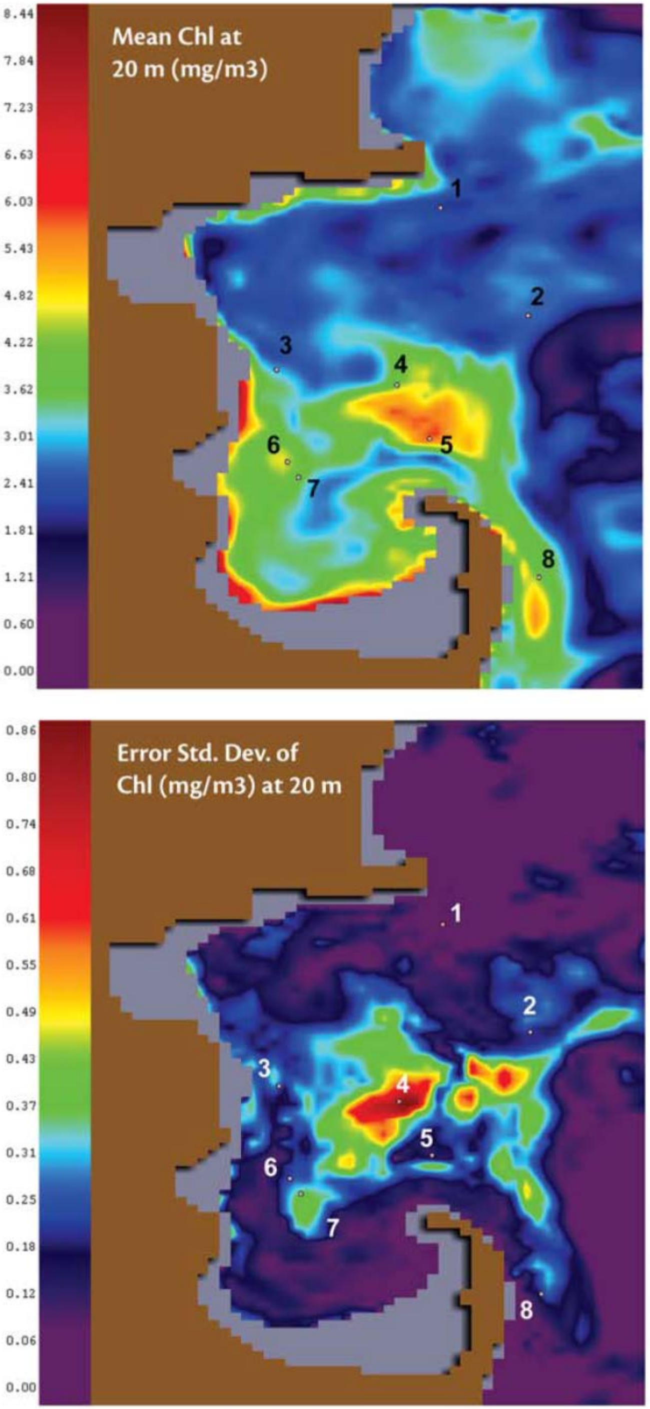






Advanced Visualization and Interactive Systems Lab: A. Love, W. Shen, A. Pang

Visualization of Uncertainties/pdf's: Multivariate Time-Dependent Fields



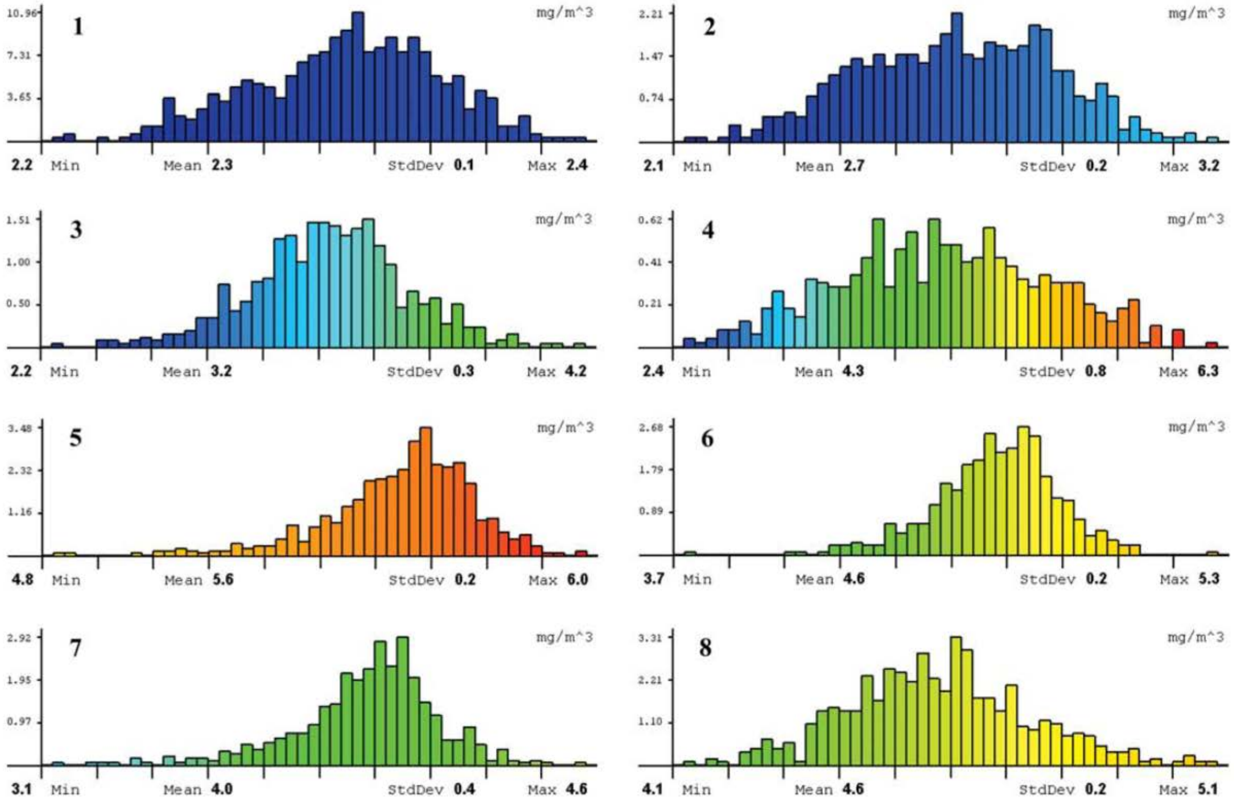
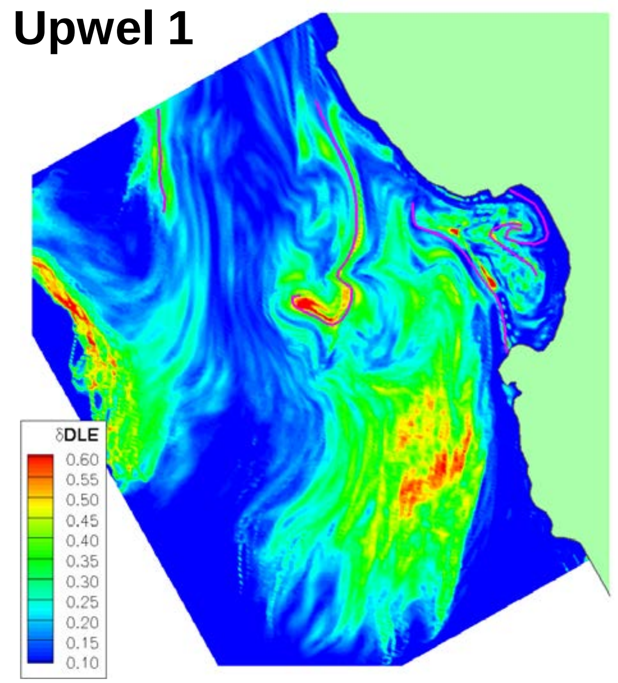


Figure 7. Chlorophyll *a* (Chl) mean and uncertainties at 20-m depth in the Mass Bay region on September 2, 1998, as hindcast by 600 Error Subspace Statistical Estimation (ESSE) ensemble members. ESSE was initialized on August 25, 1998. (Top left/right) Mean/Error Standard Deviation of Chl. (Bottom) Eight PDF estimates (normalized histograms, numbered 1 to 8) corresponding to the eight marked locations on the horizontal maps. Bars on the histograms are colored according to the center Chl value. The minimum, mean, standard deviation, and maximum values are given on each histogram (illustration by R.G. Hero, University of California, Santa Cruz).

Lermusiaux et al, Oceanography-2006

Flow Skeletons and Uncertainties:
Mean LCS overlaid on DLE error std estimate for 3 dynamical events

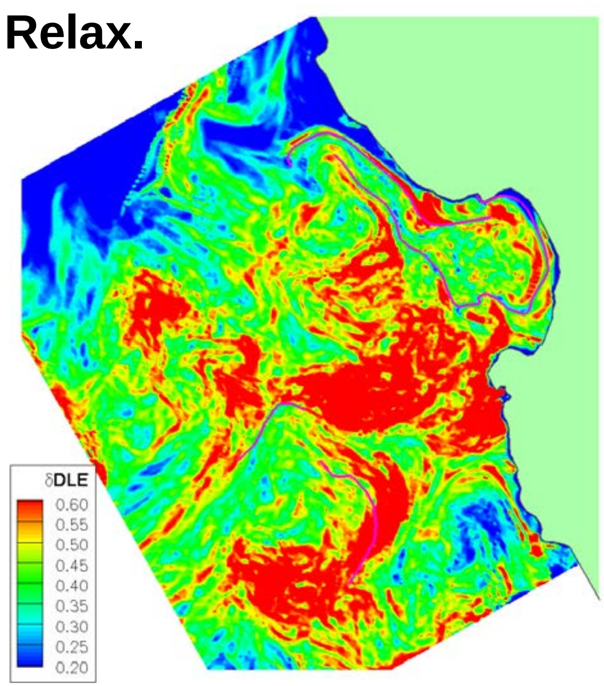
- Two upwellings and one relaxation (**about 1 week apart each**)
- Uncertainty estimates allow to identify most robust LCS (more intense DLE ridges are usually relatively more certain)
- Different oceanic regimes have different LCS uncertainty fields and properties

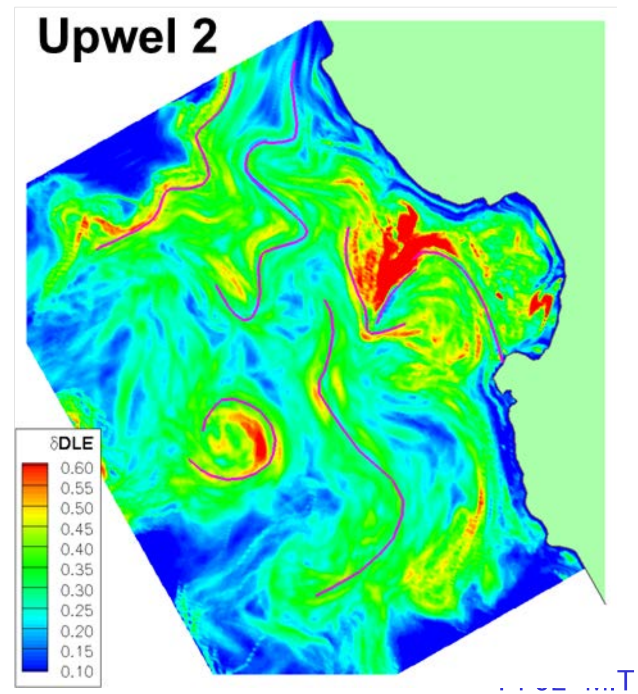


[Lermusiaux and Lekien, 2005. and In Prep, 2011

Lermusiaux, JCP-2006

Lermusiaux, Ocean.-2006]





A Grand challenge in Large Nonlinear Systems

Quantitatively estimate the accuracy of predictions

Computational challenges for the deterministic (ocean) problem

- Large dimensionality of the problem, un-stationary statistics
- Wide range of temporal and spatial scales (turbulent to climate)
- Multiple instabilities internal to the system
- Very limited observations

Need for stochastic modeling …

- Approximations in deterministic models including parametric uncertainties
- Initial and Boundary conditions uncertainties
- Measurement models

Need for data assimilation …

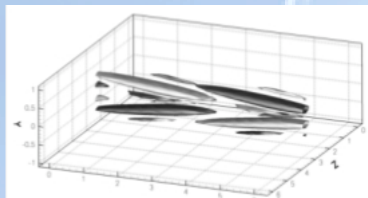
- Evolve the nonlinear, i.e. non-Gaussian, correlation structures
- Nonlinear Bayesian Estimation

Overview of Uncertainty Predictions Schemes

$$\boldsymbol{u}(\boldsymbol{x},t;\omega)=\bar{\boldsymbol{u}}(\boldsymbol{x},t)+\sum_{i=1}^{s}Y_i(t;\omega)\,\boldsymbol{u}_i(\boldsymbol{x},t)$$

Uncertainty propagation via POD method

[C. Rowley, Oberwolfach, 2008]



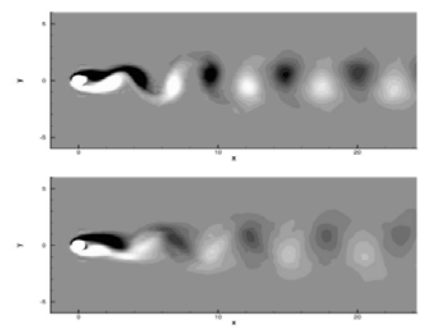
According to Lumley (Stochastic tools in Turbulence, 1971) it was introduced independently by numerous people at different times, including Kosambi (1943), Loeve (1945), Karhunen (1946), Pougachev (1953), Obukhov (1954).

Uncertainty propagation via generalized Polynomial-Chaos Method

Xiu & Karniadakis, J. Comp. Physics, 2002

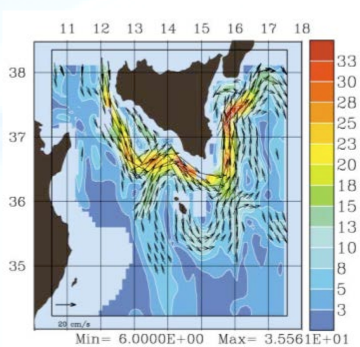
Knio & Le Maitre, Fluid Dyn. Research, 2006

Meecham & Siegel, Phys. Fluids, 1964



[Xiu & Karniadakis, J. Comp. Physics, 2002]

[Lermusiaux & Robinson, Deep Sea Research, 2001]



Uncertainty propagation via Monte Carlo method restricted to an "evolving uncertainty subspace"

(Error Subspace Statistical Estimation - ESSE)

Lermusiaux & Robinson, MWR-1999, Deep Sea Research-2001

Lermusiaux, J. Comp. Phys., 2006

B. Ganapathysubramanian & N. Zabaras, J. Comp. Phys., 2009

Problem Setup: Derive equations for UQ

Statement of the problem: A Stochastic PDE

$$\frac{\partial \boldsymbol{u}(\boldsymbol{x},t;\omega)}{\partial t} = \mathcal{L}\left[\boldsymbol{u}(\boldsymbol{x},t;\omega);\omega\right] \qquad \boldsymbol{x} \in \boldsymbol{D}$$

$$\boldsymbol{u}(\boldsymbol{x},t_0;\omega) = \boldsymbol{u}_0(\boldsymbol{x};\omega) \qquad \boldsymbol{x} \in \boldsymbol{D} \qquad \mathcal{B}\left[\boldsymbol{u}|_{\partial \boldsymbol{D}}\right] = h\left[\partial \boldsymbol{D};\omega\right]$$

$\mathcal{L}[\bullet;\omega]$ Nonlinear differential operator (possibly with stochastic coefficients)

$\boldsymbol{u}_0(\boldsymbol{x};\omega)$ Stochastic initial conditions (given full probabilistic information)

$h[\partial \boldsymbol{D};\omega]$ Stochastic boundary conditions (given full probabilistic information)

Goal: Evolve the full probabilistic information describing $\boldsymbol{u}(\boldsymbol{x},t;\omega)$

An important representation property for the solution: Compactness

$$\boldsymbol{u}(\boldsymbol{x},t;\omega) = \bar{\boldsymbol{u}}(\boldsymbol{x},t) + \sum_{i=1}^{s} Y_i(t;\omega)\,\boldsymbol{u}_i(\boldsymbol{x},t)$$

Advantage: Finite Dimension Evolving Subspace

Disadvantage: Redundancy of representation

Evolving the full representation

Major Challenge : Redundancy

$$\boldsymbol{u}(\boldsymbol{x},t;\omega)=\bar{\boldsymbol{u}}(\boldsymbol{x},t)+\sum_{i=1}^{s}Y_i(t;\omega)\boldsymbol{u}_i(\boldsymbol{x},t)$$

First Step (easy): Separate deterministic from stochastic/error subspace

Commonly used approach: Assume that $\overline{Y_i(t;\omega)}=0$

Second step (tricky): Evolving the finite dimensional subspace $\mathcal{V}_S$

A separation of roles: What can $\dfrac{dY_i(t;\omega)}{dt}$ tell us ?

Only *how the stochasticity evolves inside* $\mathcal{V}_S$

A separation of roles: What can $\dfrac{\partial \boldsymbol{u}_i(\boldsymbol{x},t)}{\partial t}$ tell us ?

How the stochasticity evolves ***both*** *inside and normal to* $\mathcal{V}_S$

source of redundancy

Natural constraint to overcome redundancy

Restrict "evolution of $\mathcal{V}_S$" to be "normal to $\mathcal{V}_S$" i.e.

$$\int \frac{\partial \boldsymbol{u}_i(\boldsymbol{x},t)}{\partial t}\boldsymbol{u}_j(\boldsymbol{x},t)\,d\boldsymbol{x}=0 \qquad \text{for all} \quad i=1,\dots,s \quad \text{and} \quad j=1,\dots,s$$

Dynamically Orthogonal Evolution Equations

Theorem 1: For a stochastic field described by the evolution equation

$$\frac{\partial \boldsymbol{u}(\boldsymbol{x},t;\omega)}{\partial t} = \mathcal{L}\left[\boldsymbol{u}(\boldsymbol{x},t;\omega);\omega\right] \quad , \quad \boldsymbol{x} \in \boldsymbol{D}$$

$$\boldsymbol{u}(\boldsymbol{x},t_0;\omega) = \boldsymbol{u}_0(\boldsymbol{x};\omega) \quad , \quad \boldsymbol{x} \in \boldsymbol{D} \qquad \mathcal{B}\left[\boldsymbol{u}(\boldsymbol{\xi},t;\omega)\right] = h(\boldsymbol{\xi},t;\omega) \quad , \quad \boldsymbol{\xi} \in \partial\boldsymbol{D}$$

assuming a response of the form $\boldsymbol{u}(\boldsymbol{x},t;\omega) = \bar{\boldsymbol{u}}(\boldsymbol{x},t) + \sum_{i=1}^{s} Y_i(t;\omega)\boldsymbol{u}_i(\boldsymbol{x},t)$

we obtain the following evolution equations

PDE describing evolution of mean field

$$\frac{\partial \bar{\boldsymbol{u}}(\boldsymbol{x},t)}{\partial t} = E^{\omega}\left[\mathcal{L}\left[\boldsymbol{u}(\boldsymbol{x},t;\omega)\right]\right], \ \boldsymbol{x} \in \boldsymbol{D} \qquad \mathcal{B}\left[\bar{\boldsymbol{u}}(\boldsymbol{\xi},t;\omega)\right] = E^{\omega}\left[h(\boldsymbol{\xi},t;\omega)\right], \ \boldsymbol{\xi} \in \partial\boldsymbol{D}$$

SDE describing evolution of stochasticity inside V_s

$$\frac{dY_j(t;\omega)}{dt} = \left\langle \mathcal{L}[\boldsymbol{u};\omega] - E^{\omega}\left[\mathcal{L}[\boldsymbol{u};\omega]\right], \boldsymbol{u}_j \right\rangle$$

Family of PDEs describing evolution of stochastic subspace V_s

$$\frac{\partial \boldsymbol{u}_j(\boldsymbol{x},t)}{\partial t} = E^{\omega}\left[Y_i\mathcal{L}[\boldsymbol{u};\omega]\right]\boldsymbol{C}^{-1}_{Y_iY_j} - \left\langle E^{\omega}\left[Y_i\mathcal{L}[\boldsymbol{u};\omega]\right], \boldsymbol{u}_k \right\rangle \boldsymbol{u}_k \boldsymbol{C}^{-1}_{Y_iY_j}$$

$$\mathcal{B}\left[\boldsymbol{u}_j(\boldsymbol{\xi},t;\omega)\right] = E^{\omega}\left[Y_i(t;\omega)h(\boldsymbol{\xi},t_0;\omega)\right]\boldsymbol{C}^{-1}_{Y_iY_j}, \ \boldsymbol{\xi} \in \partial\boldsymbol{D}$$

Sapsis and Lermusiaux, Physica D (2009, 2011)

POD & PC methods from DO equations

SDE describing evolution of stochasticity inside V_s

$$\frac{dY_j(t;\omega)}{dt} = \int_D \left\{ \mathcal{L}\left[\mathbf{u}(\mathbf{y},t;\omega)\right] - E^{\omega}\left[\mathcal{L}\left[\mathbf{u}(\mathbf{y},t;\omega)\right]\right]\right\} \mathbf{u}_j(\mathbf{y},t)\, d\mathbf{y}$$

Family of PDEs describing evolution of stochastic subspace V_s

$$\frac{\partial \mathbf{u}_j(\mathbf{x},t)}{\partial t} = E^{\omega}\left[Y_i(t;\omega)\mathcal{L}\left[\mathbf{u}(\mathbf{x},t;\omega)\right]\right]\mathbf{C}^{-1}_{Y_iY_j} - E^{\omega}\left[\int_D \mathbf{u}_k(\mathbf{y},t)Y_i(t;\omega)\mathcal{L}\left[\mathbf{u}(\mathbf{y},t;\omega)\right]d\mathbf{y}\right]\mathbf{C}^{-1}_{Y_iY_j}\mathbf{u}_k(\mathbf{x},t)$$

$$\mathcal{B}\left[\mathbf{u}_j(\boldsymbol{\xi},t;\omega)\right] = E^{\omega}\left[Y_i(t;\omega)h(\boldsymbol{\xi},t_0;\omega)\right]\mathbf{C}^{-1}_{Y_iY_j},\ \boldsymbol{\xi} \in \partial \mathbf{D}$$

PDE describing evolution of mean field

$$\frac{\partial \bar{\mathbf{u}}(\mathbf{x},t)}{\partial t} = E^{\omega}\left[\mathcal{L}\left[\mathbf{u}(\mathbf{x},t;\omega)\right]\right],\ \mathbf{x} \in \mathbf{D} \qquad \mathcal{B}\left[\bar{\mathbf{u}}(\boldsymbol{\xi},t;\omega)\right] = E^{\omega}\left[h(\boldsymbol{\xi},t;\omega)\right],\ \boldsymbol{\xi} \in \partial \mathbf{D}$$

Choosing a priori the stochastic subspace V_s using POD methodology we recover POD equations.

Choosing a priori the statistical characteristics of the stochastic coefficients $Y_j(t;\omega)$ we recover the PC equations.

Application I : Navier-Stokes in a cavity

2D viscous flow with stochastic initial conditions and no stochastic excitation

$$\frac{\partial u}{\partial t}+\frac{\partial P}{\partial x}=\frac{1}{\mathrm{Re}}\Delta u-\frac{\partial\left(u^2\right)}{\partial x}-\frac{\partial\left(uv\right)}{\partial y}$$

$$\frac{\partial v}{\partial t}+\frac{\partial P}{\partial y}=\frac{1}{\mathrm{Re}}\Delta v-\frac{\partial\left(uv\right)}{\partial x}-\frac{\partial\left(v^2\right)}{\partial y}$$

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0$$

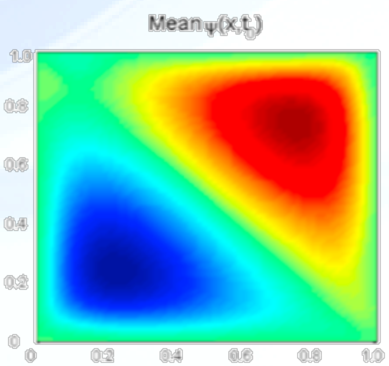
$u=U, v=0$

$u=0$
$v=0$

$u=0$
$v=0$

$u=0, v=0$

Initial mean flow



Streamfunction

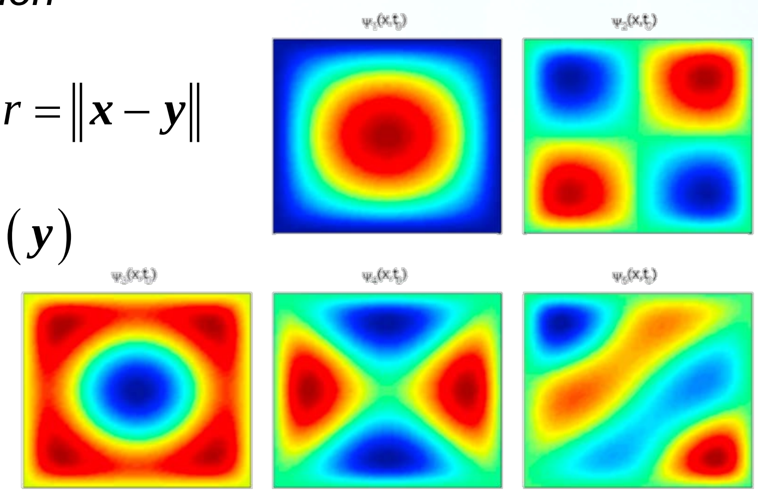
Initial Covariance function

$$C(r)=\left(1+br+\frac{b^2r^2}{3}\right)e^{-br} \quad r=\left\|\boldsymbol{x}-\boldsymbol{y}\right\|$$

$$\int \boldsymbol{C}\left(\left\|\boldsymbol{x}-\boldsymbol{y}\right\|\right)\hat{\boldsymbol{u}}_i\left(\boldsymbol{x}\right)d\boldsymbol{x}=\lambda_i^2\hat{\boldsymbol{u}}_i\left(\boldsymbol{y}\right)$$

$$\boldsymbol{u}_{0,i}\left(\boldsymbol{x}\right)=\hat{\boldsymbol{u}}_i\left(\boldsymbol{x}\right)$$

$$Y_i\left(t_0;\omega\right)\sim\mathcal{N}\left(0,\lambda_i\right)$$



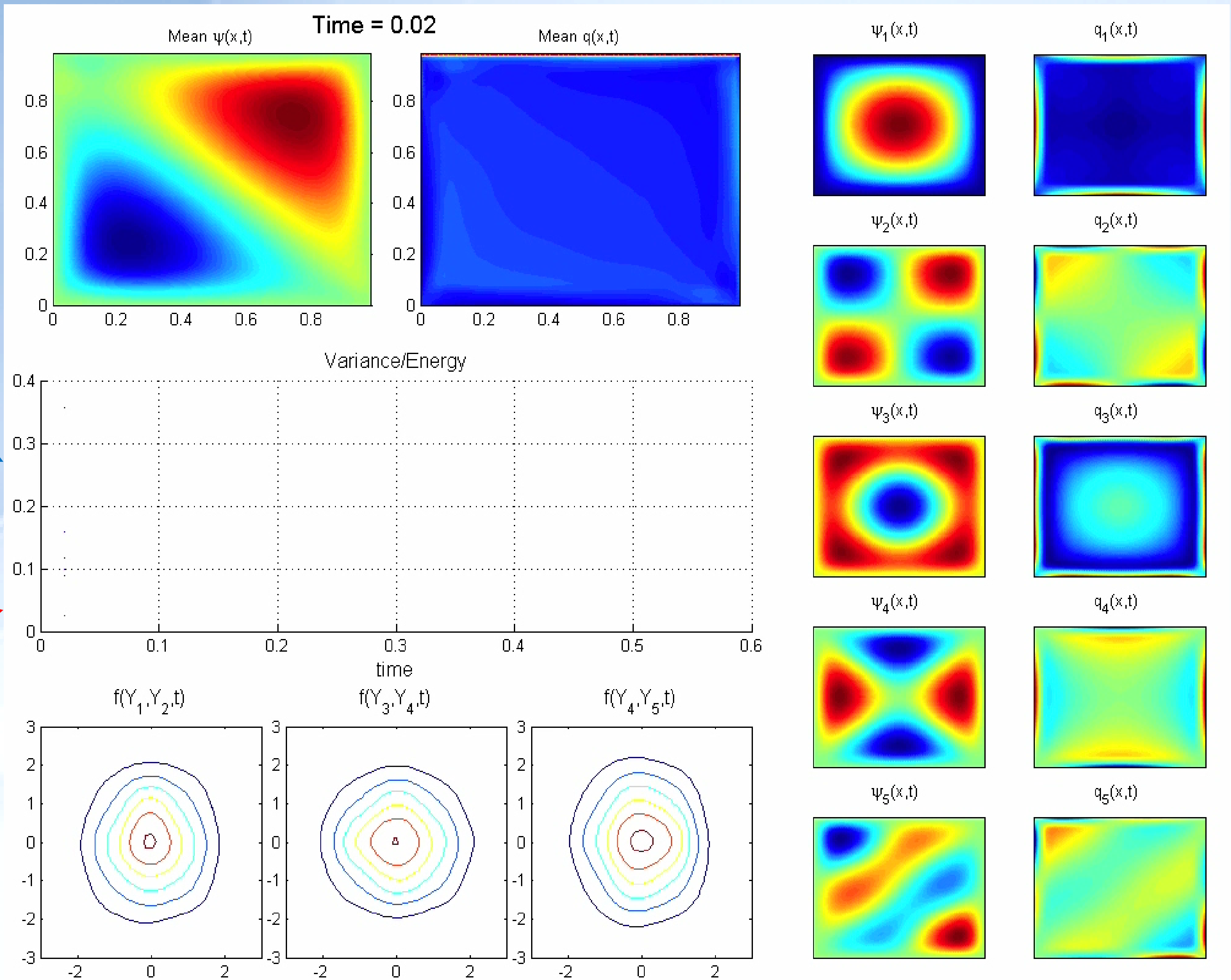
PDE Numerics: C-grid, upwind [M. Griebel et al., 1998] SDE Numerics: here, s-dimensional Monte-Carlo

Application I : Navier-Stokes in a cavity

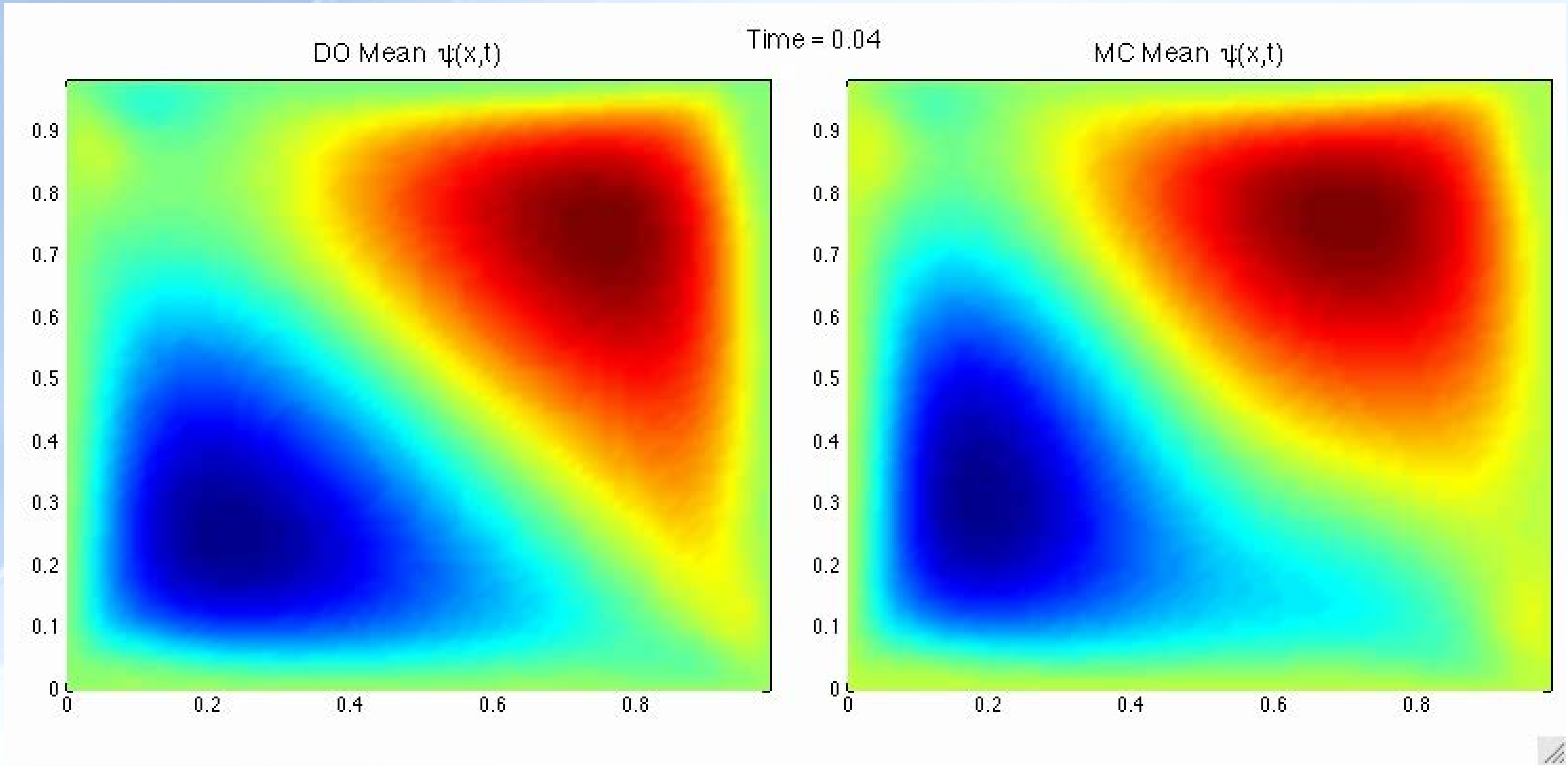
Re = 1000

Variances of each mode

Energy of mean flow

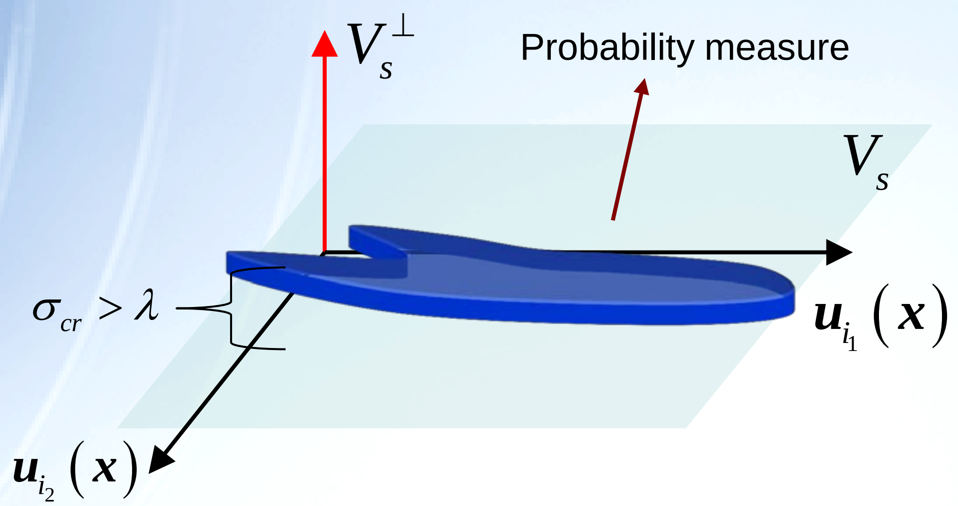


Comparison with Monte-Carlo



Comp. time: 11min (4000 samples or analytical Y_i) 12,3h (300 samples)

Adapt the stochastic subspace dimension



- In the context of DO equations so far the size of the stochastic subspace V_s remained invariant.
- For intermittent or transient phenomena the dimension of the stochastic subspace may vary significantly with time. This is accounted for by ESSE.

We need criteria to evolve the dimensionality of the stochastic subspace

This is a particularly important issue for stochastic systems with deterministic initial conditions

Criteria for dimension reduction / increase

Dimension Reduction

Comparison of the minimum eigenvalue of the correlation matrix $\boldsymbol{C}_{Y_iY_j}$.

$$\lambda_{\min}\left[\boldsymbol{C}_{Y_iY_j}\right] < \sigma_{cr} \longrightarrow \text{pre-defined value}$$

Removal of the corresponding direction from the stochastic subspace.

Dimension Increase

Comparison of the minimum eigenvalue of the correlation matrix $\boldsymbol{C}_{Y_iY_j}$.

$$\lambda_{\min}\left[\boldsymbol{C}_{Y_iY_j}\right] > \Sigma_{cr} \longrightarrow \text{pre-defined value}$$

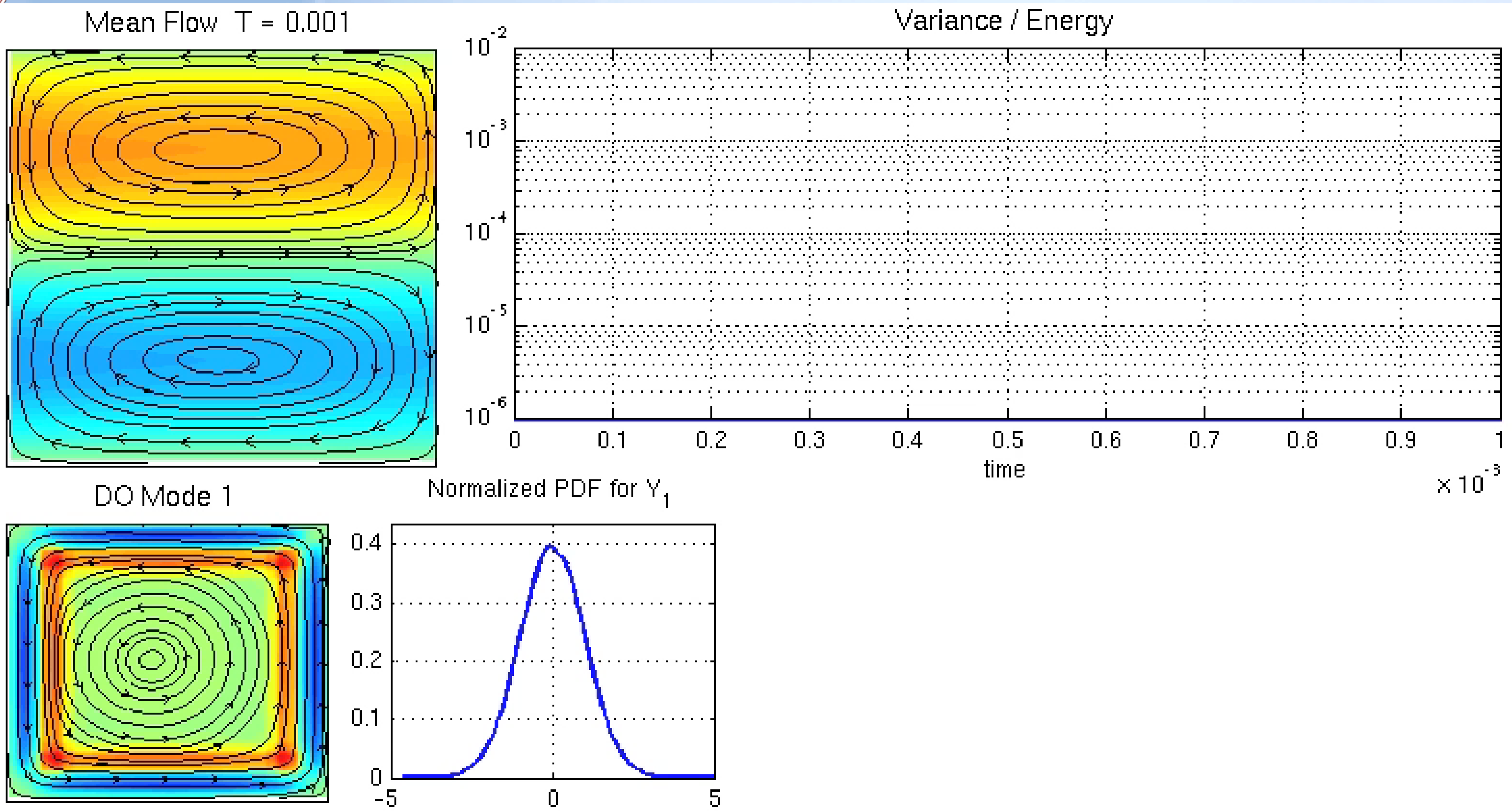
Addition of a new direction $\boldsymbol{u}_i(\boldsymbol{x},t)$ in the stochastic subspace V_s.

How do we choose this new direction?

By breeding in the orthogonal complement $V_s^{\perp}$

Same problem when we start with deterministic initial condition (dimension of stochastic subspace is zero)

Example: Double Gyre, Re=10,000

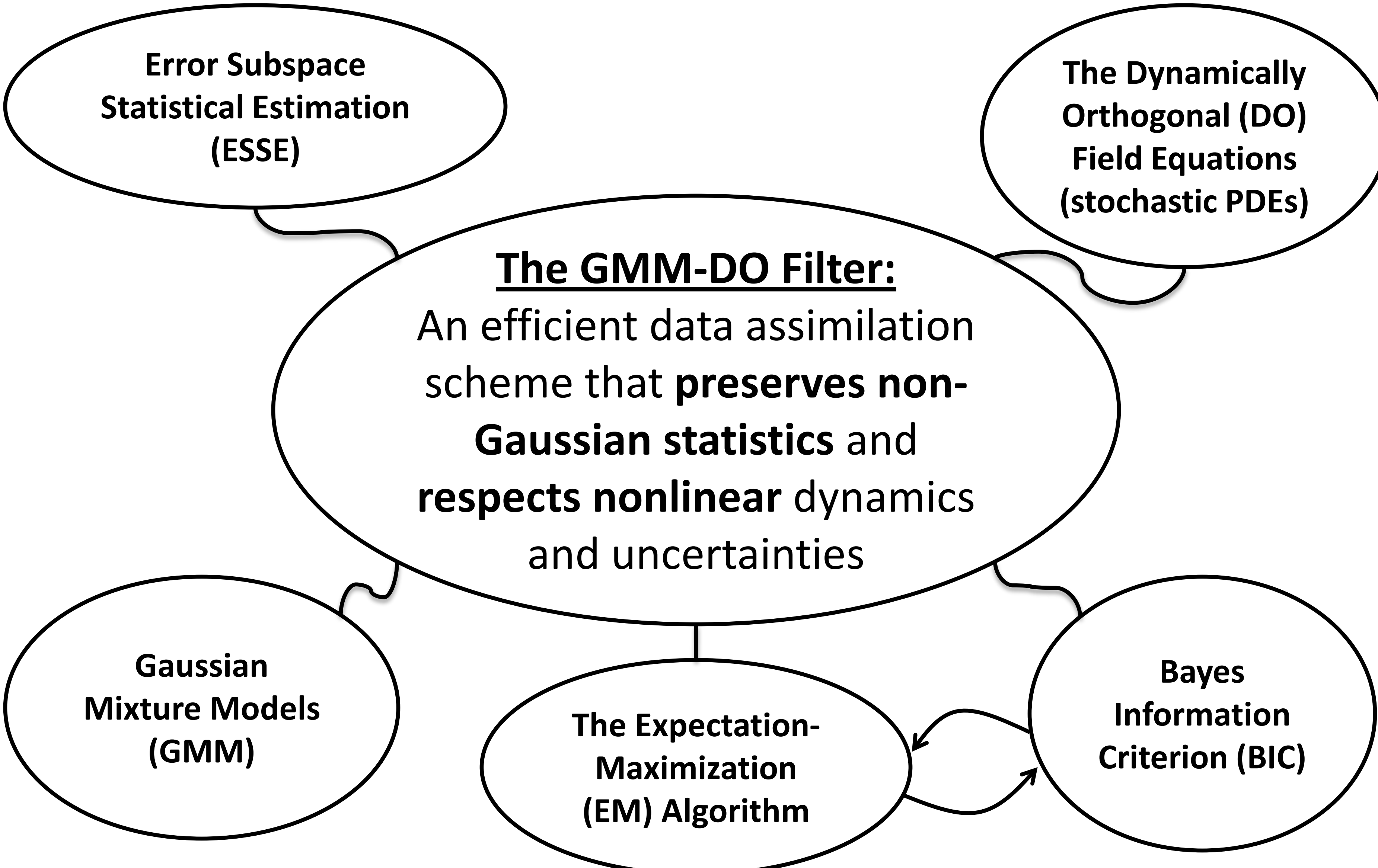


The GMM-DO Filter:

Data Assimilation and Adaptive Sampling
with
Gaussian Mixture Models
using the
Dynamically Orthogonal field equations

(Sondergaard, 2011; Sondergaard and Lermusiaux, MWR-to-be-submitted , Parts I and II)

Overview of the GMM-DO Filter



Gaussian Mixture Models (with Bayesian update)

The probability density function for a random vector, **x**, distributed according to a multivariate Gaussian mixture model is given by

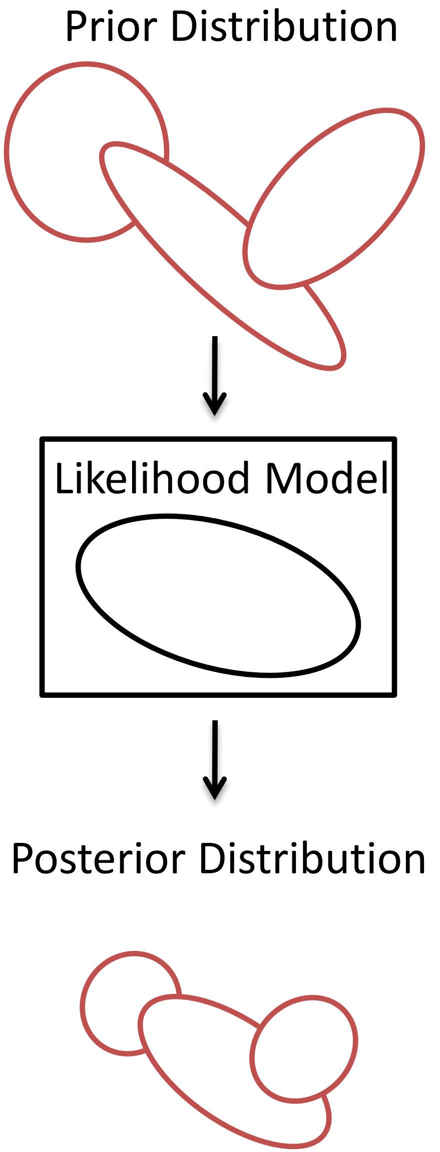
$$p_{\mathbf{X}}(\boldsymbol{x}) = \sum_{j=1}^{M} \pi_j \times N(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

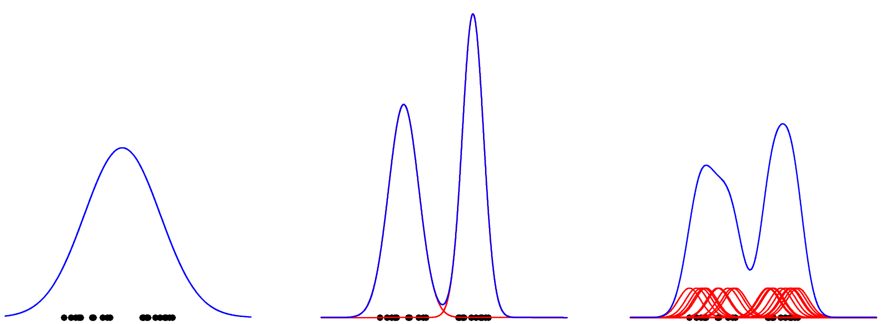
subject to the constraint that

$$\sum_{j=1}^{M} \pi_j = 1.$$

We refer to M as the mixture complexity and π_j as the mixture weights. The multivariate Gaussian density function takes the form:

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}.$$





Parametric Distribution | Gaussian Mixture Model | Kernel Density Approximation

Overview of the GMM-DO Filter

Initial Conditions: *Initialize the state vector in a decomposed form that accords with the Dynamically Orthogonal field equations:*

$$\boldsymbol{x}_{r,0} = \bar{\boldsymbol{x}}_0 + \boldsymbol{\mathcal{X}}_0 \boldsymbol{\phi}_{r,0}, \quad r = \{1, \ldots, N\}.$$

$\bar{\boldsymbol{x}} \in \mathbb{R}^n$ is the mean vector, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times s}$ defines the matrix of modes describing an orthonormal basis for the stochastic subspace, and the $\boldsymbol{\phi}_r \in \mathbb{R}^s$ represent N realizations drawn from the multivariate random vector described by $\{\Phi_{1,0}(\omega), \ldots, \Phi_{s,0}(\omega)\}$ that reside in the stochastic subspace of dimension s.

Forecast: *Using either the initial DO conditions or the posterior state description following the assimilation of data at time $k-1$,*

$$\boldsymbol{x}^a_{r,k-1} = \bar{\boldsymbol{x}}^a_{k-1} + \boldsymbol{\mathcal{X}}_{k-1} \boldsymbol{\phi}^a_{r,k-1}, \quad r = \{1, \ldots, N\},$$

apply the DO equations to efficiently evolve the probabilistic description of the state vector in time, arriving at a forecast for observation time k:

$$\boldsymbol{x}^f_{r,k} = \bar{\boldsymbol{x}}^f_k + \boldsymbol{\mathcal{X}}_k \boldsymbol{\phi}^f_{r,k}, \quad r = \{1, \ldots, N\}.$$

Observation: *Employ a linear (or linearized) observation model,*

$$\boldsymbol{Y}_k = \boldsymbol{H}\boldsymbol{X}_k + \boldsymbol{\Upsilon}_k, \quad \boldsymbol{\Upsilon}_k \sim \mathcal{N}(\boldsymbol{\upsilon}_k; \boldsymbol{0}, \boldsymbol{R}).$$

where $\boldsymbol{Y}_k \in \mathbb{R}^p$ is the observation random vector at discrete time k; $\boldsymbol{H} \in \mathbb{R}^{p \times n}$ is the linear observation model; and $\boldsymbol{\Upsilon}_k \in \mathbb{R}^p$ the corresponding random noise vector, assumed to be of a Gaussian distribution. The realized observation vector is denoted by $\boldsymbol{y}_k \in \mathbb{R}^p$.

(Sondergaard, 2011; Sondergaard and Lermusiaux, MWR-to-be-submitted , Parts I and II)

GMM Filter Example:
Flow Exiting a Strait or “Sudden Expansion Flow”

Time t = 50, **True solution**

Mean field prior

Modes 1 to 4

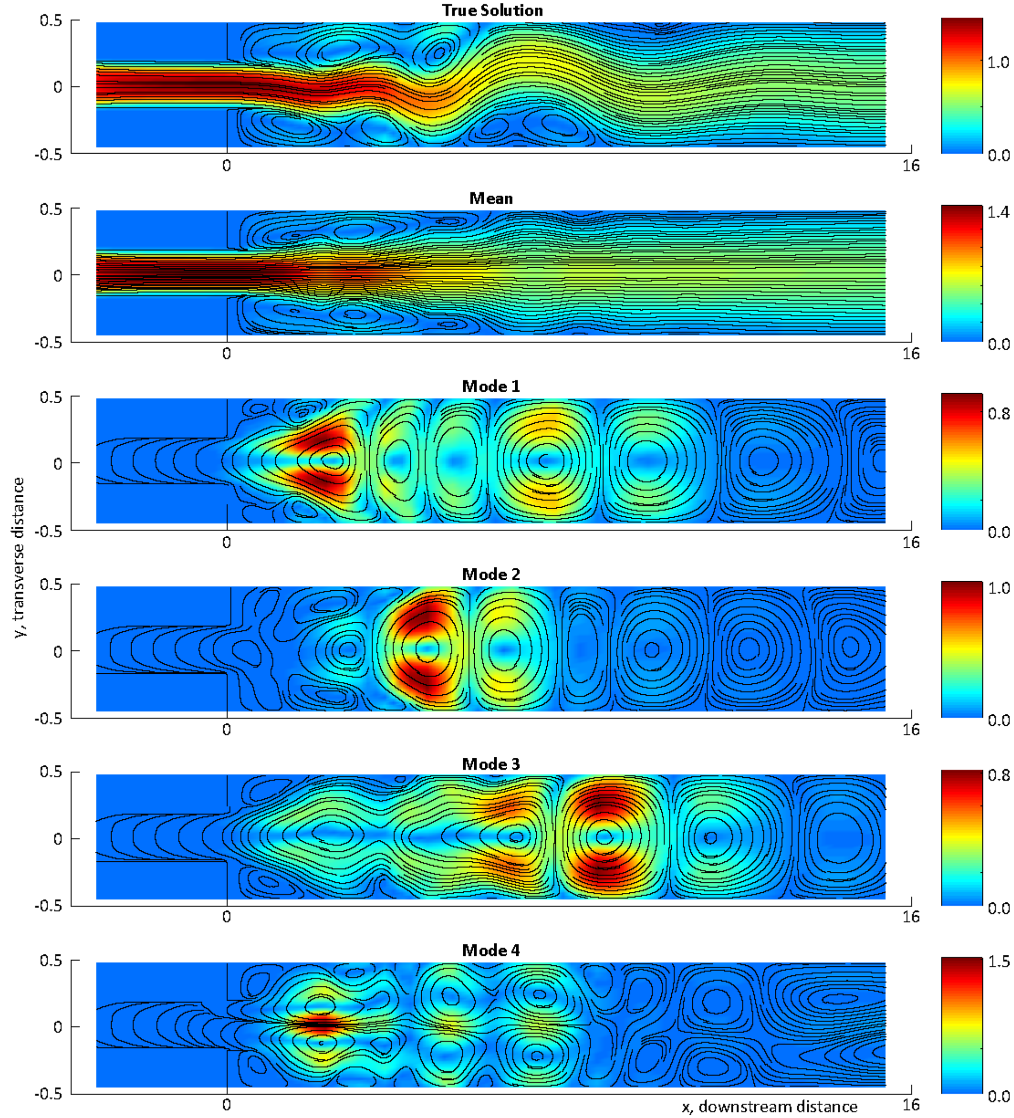


Figure 5-15: True solution; DO mean field; and first four DO modes at the first assimilation step, time $T = 50$.

Overview of the GMM-DO Filter

GMM fit in DO stochastic subspace

Forecast: *Using either the initial DO conditions or the posterior state description following the assimilation of data at time* $k-1$,

$$\boldsymbol{x}^a_{r,k-1} = \bar{\boldsymbol{x}}^a_{k-1} + \boldsymbol{\mathcal{X}}_{k-1}\boldsymbol{\phi}^a_{r,k-1}, \quad r = \{1,\ldots,N\},$$

apply the DO equations to efficiently evolve the probabilistic description of the state vector in time, arriving at a forecast for observation time k:

$$\boldsymbol{x}^f_{r,k} = \bar{\boldsymbol{x}}^f_k + \boldsymbol{\mathcal{X}}_k\boldsymbol{\phi}^f_{r,k}, \quad r = \{1,\ldots,N\}.$$

Fitting of GMM: *For Gaussian mixture models of increasing complexity (i.e.* $M = 1,2,3,\ldots$*), repeat until a minimum of the BIC is met:*

i. Use the EM algorithm to obtain the prior mixture parameters

$$\pi^f_{j,k}, \boldsymbol{\mu}^f_{j,k}, \boldsymbol{\Sigma}^f_{j,k}, \qquad j = 1,\ldots,M$$

within the stochastic subspace based on the set of ensemble realizations, $\{\boldsymbol{\phi}^f_k\} = \{\boldsymbol{\phi}^f_{1,k},\ldots,\boldsymbol{\phi}^f_{N,k}\}$.

ii. Use the Bayesian Information Criterion to evaluate the fit of the Gaussian mixture model for the given M.

(Sondergaard, 2011; Sondergaard and Lermusiaux, MWR-to-be-submitted , Parts I and II)

The EM algorithm with GMM

Based on the data at hand, the Expectation-Maximization algorithm describes an iterative procedure for obtaining the **Maximum Likelihood** estimate for the unknown set of parameters, θ, here of our prior Gaussian mixture model:

$$\{\pi_1, \ldots, \pi_M, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M, \Sigma_1, \ldots, \Sigma_M\}.$$

Procedure. Given the n data, x, and initial parameter estimate θ(0), repeat until convergence:

(1) Expectation: Using the current set of parameters, θ(k), form

$$\tau_j\left(\boldsymbol{x}_i; \boldsymbol{\theta}^{(k)}\right) = \frac{\pi_j^{(k)} N\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j^{(k)}, \Sigma_j^{(k)}\right)}{\sum_{m=1}^{M} \pi_k^{(k)} N\left(\boldsymbol{x}_i; \boldsymbol{\mu}_m^{(k)}, \Sigma_m^{(k)}\right)}.$$

(2) Minimization: Update the estimate for the set of parameters, θ(k+1), according to

$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_j\left(\boldsymbol{x}_i; \boldsymbol{\theta}^{(k)}\right)}{n} = \frac{n_j^{(k)}}{n}$$

$$\boldsymbol{\mu}_j^{(k+1)} = \frac{1}{n_j^{(k)}} \sum_{i=1}^{n} \tau_j\left(\boldsymbol{x}_i; \boldsymbol{\theta}^{(k)}\right) \boldsymbol{x}_i$$

$$\Sigma_j^{(k+1)} = \frac{1}{n_j^{(k)}} \sum_{i=1}^{n} \tau_j\left(\boldsymbol{x}_i; \boldsymbol{\theta}^{(k)}\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)}\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(k)}\right)^T.$$

Bayes Information Criterion

Determining the complexity of a Gaussian mixture model can be put in the context of **model selection**: based on the data at hand, **x**, we wish to select the model complexity that maximizes the likelihood of this data:

$$p_{\mathbf{X}}(x; M) = \int p_{\mathbf{X}|\Theta}(x|\theta; M) p_{\Theta}(\theta; M) d\theta$$

We use **Bayes Information Criterion** -- we select the simplest hypothesis consistent with the data, i.e. maximize the log-likelihood of the data around the EM-ML estimate of the parameters:

$$L_x^N(\hat{\theta}_{ML}, M) = \sum_{i=1}^{N} \log p_{\mathbf{X}|\Theta}(x_i|\hat{\theta}_{ML}; M)$$

Number of parameters

$$\frac{1}{N} L_x^N(M) = \frac{1}{N} L_x(\hat{\theta}_{ML}, M) - \frac{K_m}{2N} \log N.$$

log-likelihood of the data:

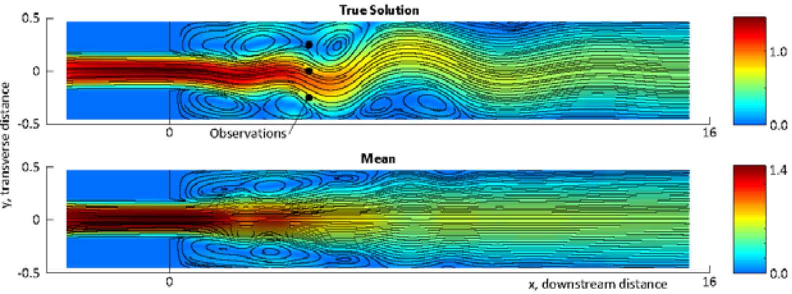
ML estimate of parameter vector

Number of data points

$$L_x^N(M) = \sum_{i=1}^{N} \log p_{\mathbf{X}}(x_i; M)$$

GMM Filter Example:
Flow Exiting a Strait or "Sudden Expansion Flow"

Time t = 50   Prior Distribution



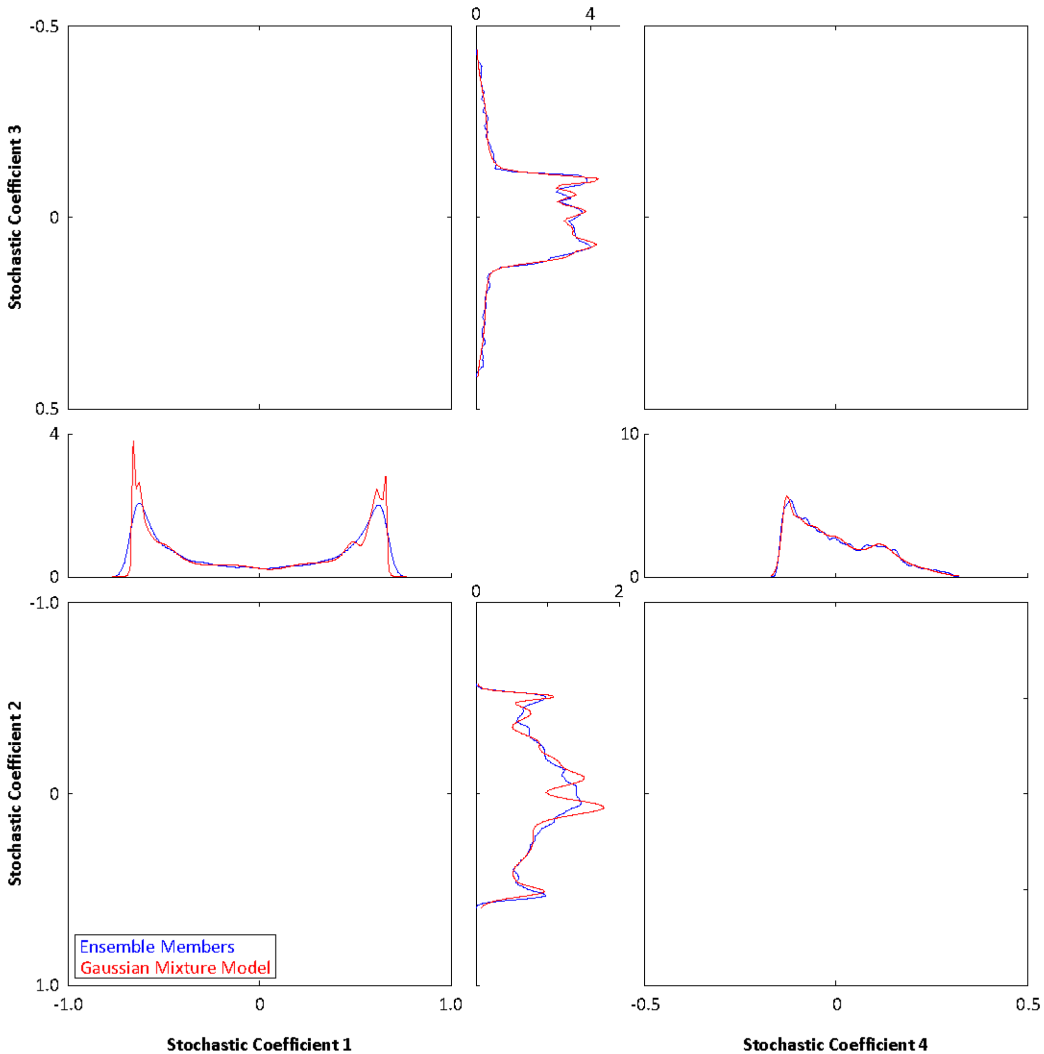
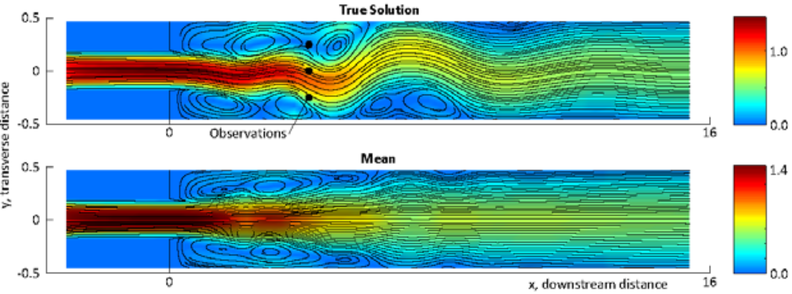


Figure 5-16: True solution; DO mean field; and joint and marginal prior distribution, identified by the Gaussian mixture model of complexity 29, and associated ensembles of the first four modes at the first assimilation step, time $T = 50$.

GMM Filter Example:
Flow Exiting a Strait or "Sudden Expansion Flow"

Time t = 50   Prior Distribution



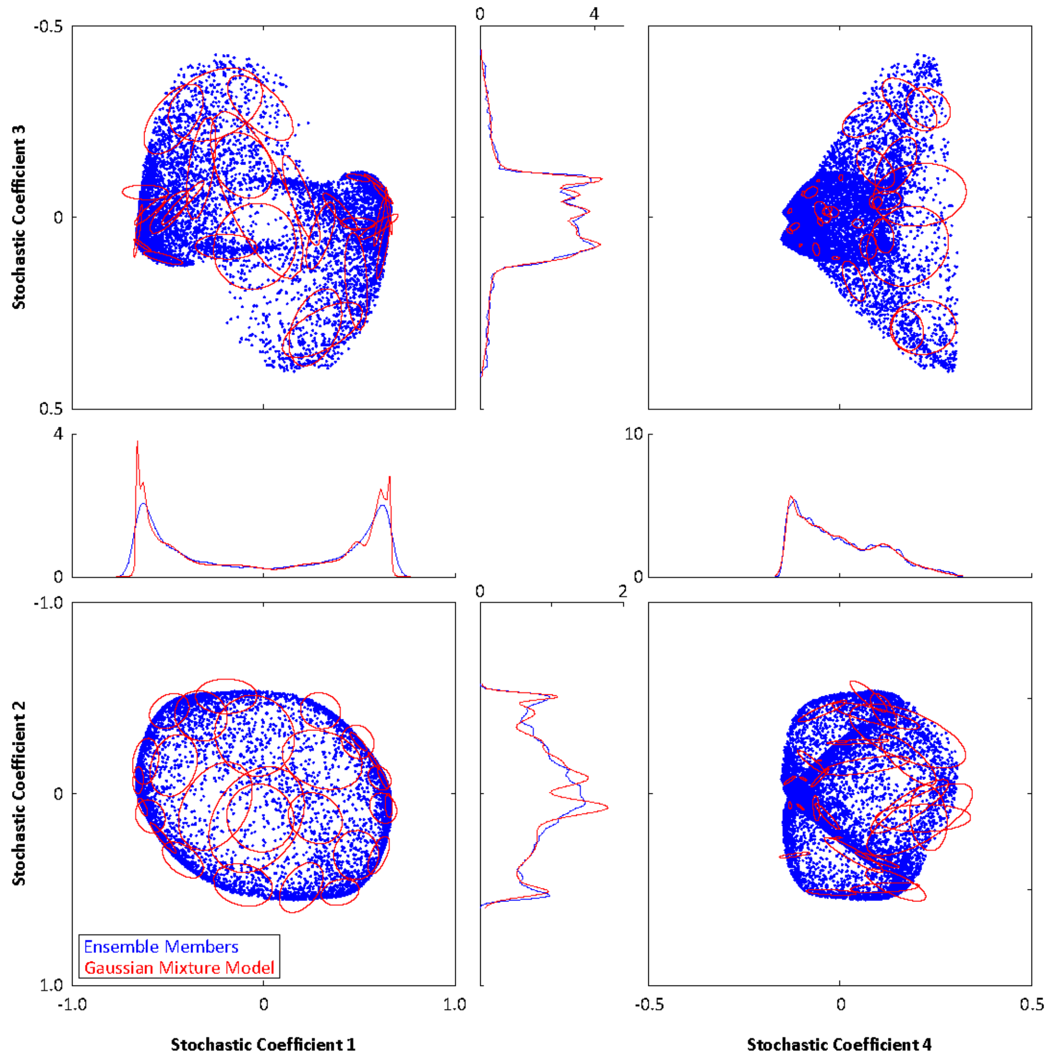


Figure 5-16: True solution; DO mean field; and joint and marginal prior distribution, identified by the Gaussian mixture model of complexity 29, and associated ensembles of the first four modes at the first assimilation step, time $T = 50$.

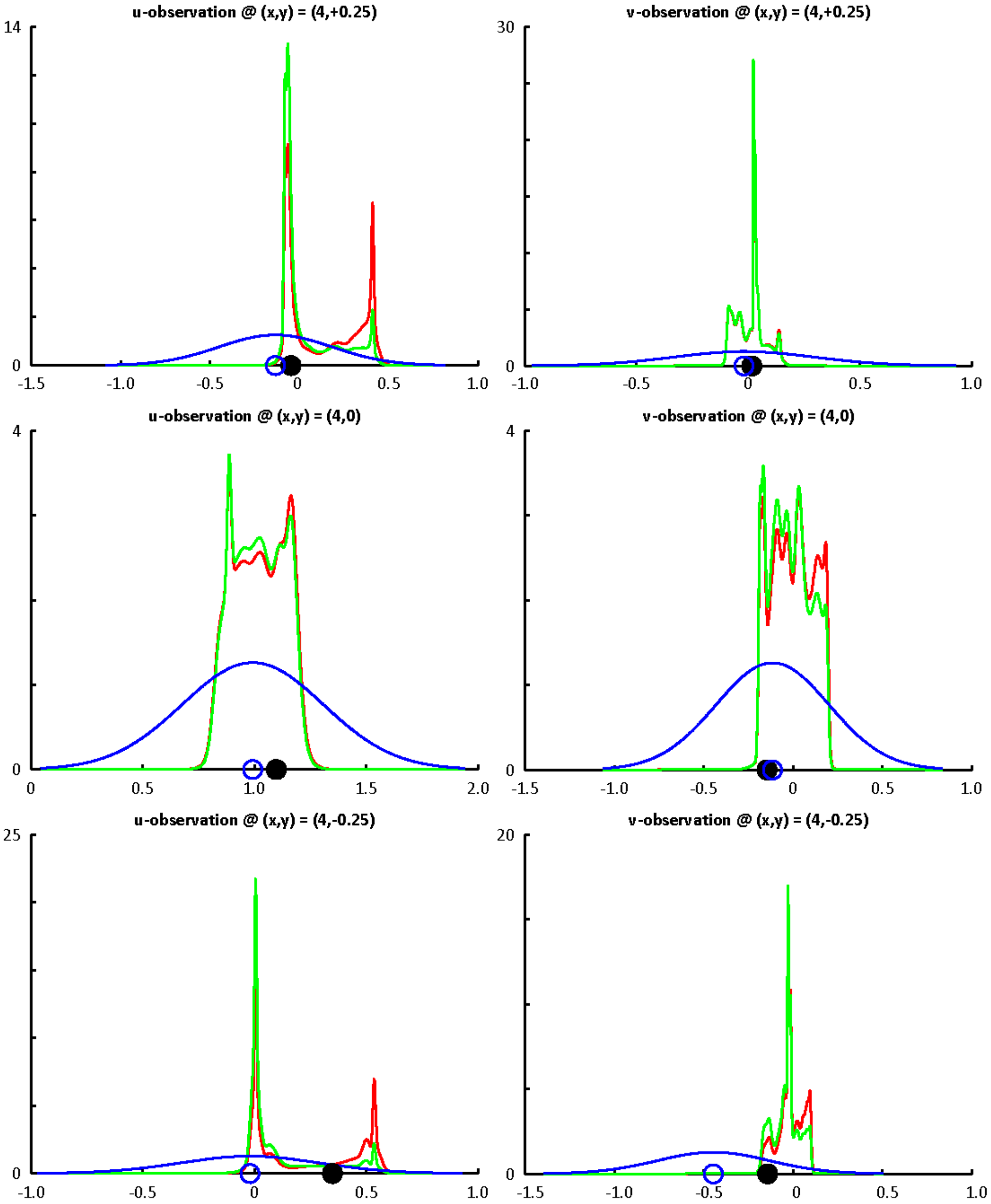
GMM Filter Example: Flow Exiting a Strait or "Sudden Expansion Flow"

Time t = 50

Observations and their pdf

Prior Distributions at these data points

Posterior Distributions at these data points



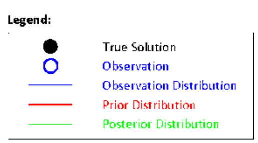


Figure 5-17: True solution; observation and its associated Gaussian distribution; and the prior and posterior distributions at the observation locations at time $T = 50$.

Overview of the GMM-DO Filter

Observation: *Employ a linear (or linearized) observation model,*

$$\boldsymbol{Y}_k = \boldsymbol{H}\boldsymbol{X}_k + \boldsymbol{\Upsilon}_k, \quad \boldsymbol{\Upsilon}_k \sim \mathcal{N}(\boldsymbol{\upsilon}_k; \boldsymbol{0}, \boldsymbol{R}).$$

where $\boldsymbol{Y}_k \in \mathbb{R}^p$ *is the observation random vector at discrete time* k*;* $\boldsymbol{H} \in \mathbb{R}^{p\times n}$ *is the linear observation model; and* $\boldsymbol{\Upsilon}_k \in \mathbb{R}^p$ *the corresponding random noise vector, assumed to be of a Gaussian distribution. The realized observation vector is denoted by* $\boldsymbol{y}_k \in \mathbb{R}^p$*.*

Bayesian Update of GMM in DO stochastic subspace

For GMM-DO Update Theorem, see:

(Sondergaard and Lermusiaux, MWR -to-be-submitted -2011, Parts I and II)

Update:

i. *Compute:*

$$\tilde{\boldsymbol{H}}_k \equiv \boldsymbol{H}\boldsymbol{\mathcal{X}}_k$$
$$\tilde{\boldsymbol{y}}_k \equiv \boldsymbol{y}_k - \boldsymbol{H}\bar{\boldsymbol{x}}^f_k$$

and determine the individual Kalman gain matrices:

$$\tilde{\boldsymbol{K}}_{j,k} = \boldsymbol{\Sigma}^f_{j,k}\tilde{\boldsymbol{H}}^T_k(\tilde{\boldsymbol{H}}_k\boldsymbol{\Sigma}^f_{j,k}\tilde{\boldsymbol{H}}^T_k + \boldsymbol{R})^{-1}, \quad j = 1,\ldots,M.$$

ii. *Assimilate the measurement,* $\boldsymbol{y}_k$*, by calculating the 'intermediate' mixture means in the stochastic subspace,*

$$\hat{\boldsymbol{\mu}}^a_{j,k} = \boldsymbol{\mu}^f_{j,k} + \tilde{\boldsymbol{K}}_{j,k}(\tilde{\boldsymbol{y}}_k - \tilde{\boldsymbol{H}}_k\boldsymbol{\mu}^f_{j,k}),$$

and further compute the posterior mixture weights:

$$\pi^a_{j,k} = \frac{\pi^f_{j,k}\times\mathcal{N}(\tilde{\boldsymbol{y}}_k; \tilde{\boldsymbol{H}}_k\boldsymbol{\mu}^f_{j,k}, \tilde{\boldsymbol{H}}_k\boldsymbol{\Sigma}^f_{j,k}\tilde{\boldsymbol{H}}^T_k + \boldsymbol{R})}{\sum_{l=1}^{M}\pi^f_{l,k}\times\mathcal{N}(\tilde{\boldsymbol{y}}_k; \tilde{\boldsymbol{H}}_k\boldsymbol{\mu}^f_{l,k}, \tilde{\boldsymbol{H}}_k\boldsymbol{\Sigma}^f_{l,k}\tilde{\boldsymbol{H}}^T_k + \boldsymbol{R})}.$$

iii. *Update the DO mean field,*

$$\bar{\boldsymbol{x}}^a_k = \bar{\boldsymbol{x}}^f_k + \boldsymbol{\mathcal{X}}_k\sum_{j=1}^{M}\pi^a_{j,k}\times\hat{\boldsymbol{\mu}}^a_{j,k},$$

as well as the mixture parameters within the stochastic subspace:

$$\boldsymbol{\mu}^a_{j,k} = \hat{\boldsymbol{\mu}}^a_{j,k} - \sum_{j=1}^{M}\pi^a_{j,k}\times\hat{\boldsymbol{\mu}}^a_{j,k}$$
$$\boldsymbol{\Sigma}^a_{j,k} = (\boldsymbol{I} - \tilde{\boldsymbol{K}}_{j,k}\tilde{\boldsymbol{H}}_k)\boldsymbol{\Sigma}^f_{j,k}.$$

iv. *Generate the posterior set of subspace ensemble realizations,* $\{\boldsymbol{\phi}^a_k\} = \{\boldsymbol{\phi}^a_{1,k},\ldots,\boldsymbol{\phi}^a_{N,k}\}$*, according to the multivariate Gaussian mixture model with posterior parameter values:*

$$\pi^a_{j,k}, \boldsymbol{\mu}^a_{j,k}, \boldsymbol{\Sigma}^a_{j,k}, \quad j = 1,\ldots,M.$$

GMM Filter Example:
Flow Exiting a Strait or “Sudden Expansion Flow”

Time t = 50

Posterior Distribution

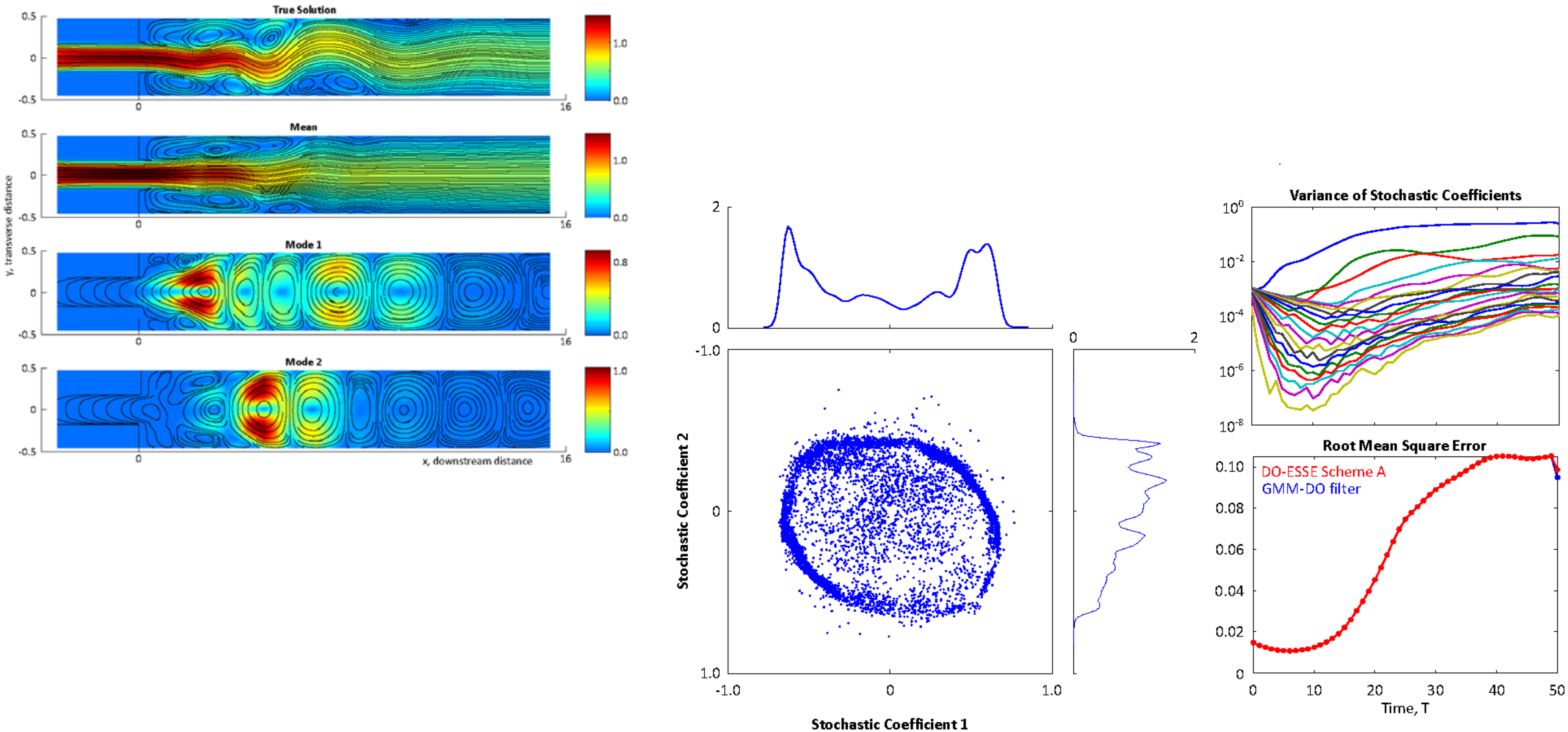


Figure 5-18: True solution; condensed representation of the posterior DO decomposition; and root mean square errors at time $T = 50$.

GMM-DO Filter:
DO equations and Non-Gaussian Data Assimilation

“Flow exiting a Strait” Test Case: Results show that our new DO equations and Non-Gaussian assimilation leads to optimal error reduction

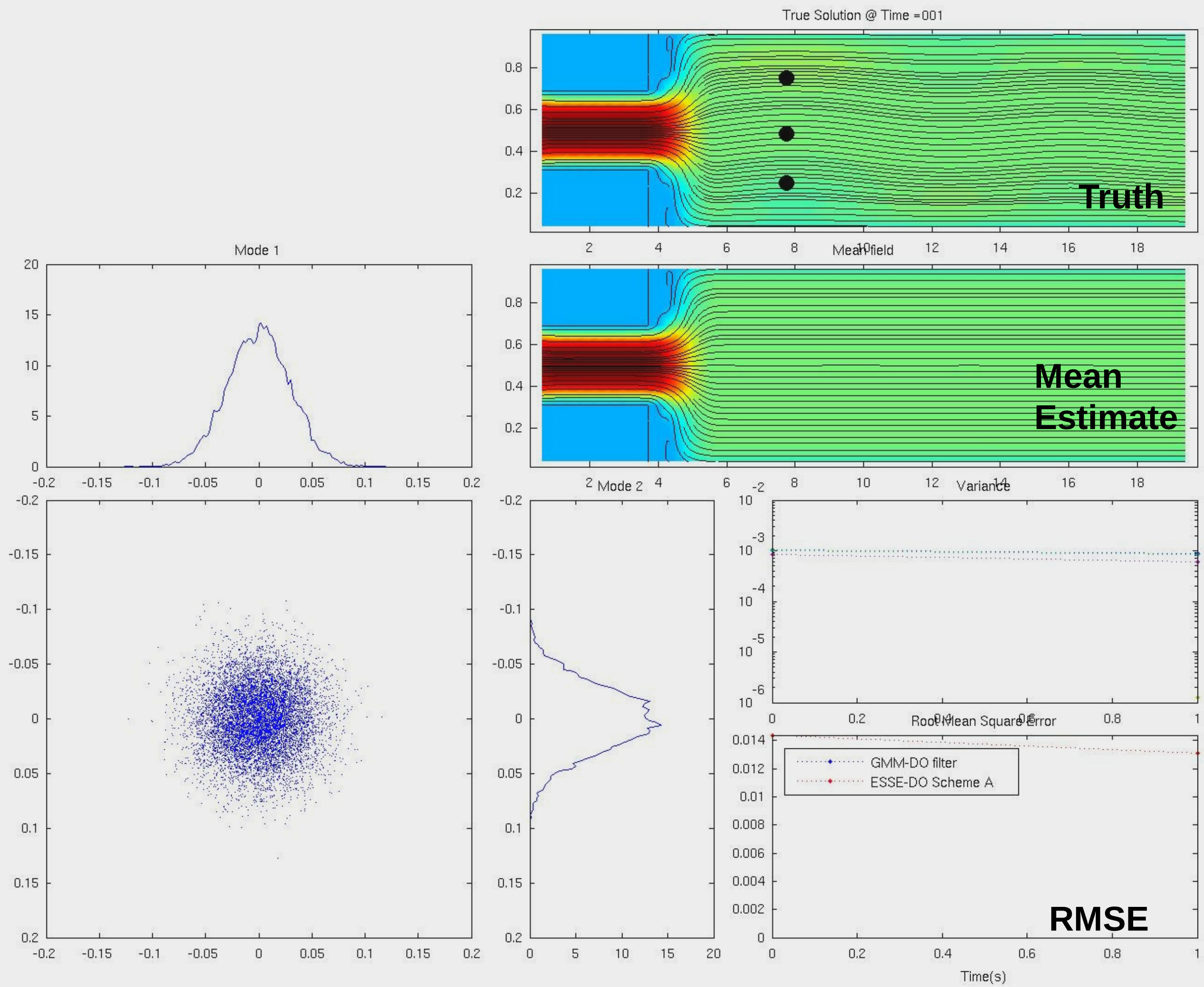
(Top Right): True solution mean flow-field streamlines overlaid on vorticity with sampling positions as circles

(Left): DO marginal pdfs represented as samples with the single 1st and 2nd DO marginal pdfs on each sides, clearly showing non-Gaussian behavior.

(Middle Right): Mean estimate mean using GMM-DO filter.

(Bottom Right): Variance of 10 DO modes as a function of time

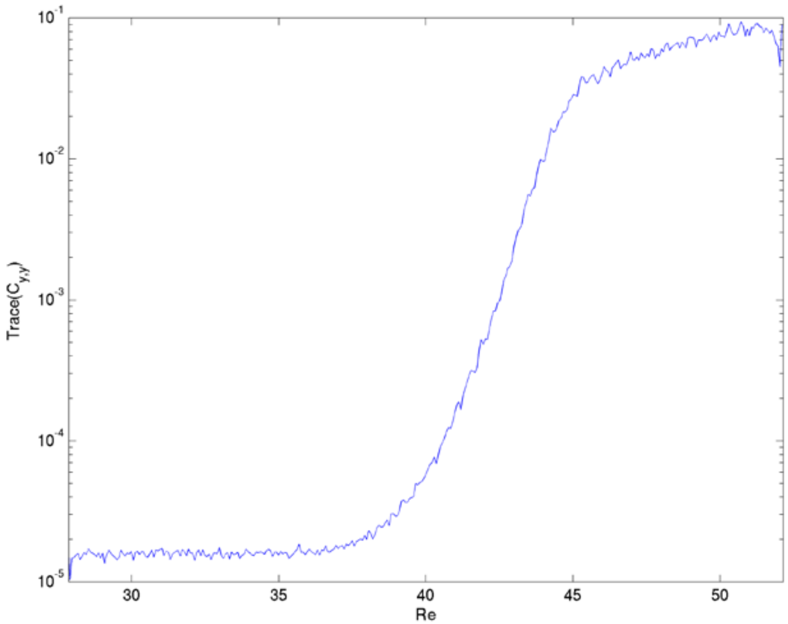
(Bottom Right): Comparisons of the root-mean-square-error (truth minus mean) as a function of time, clearly showing superior performance of GMM-DO filter

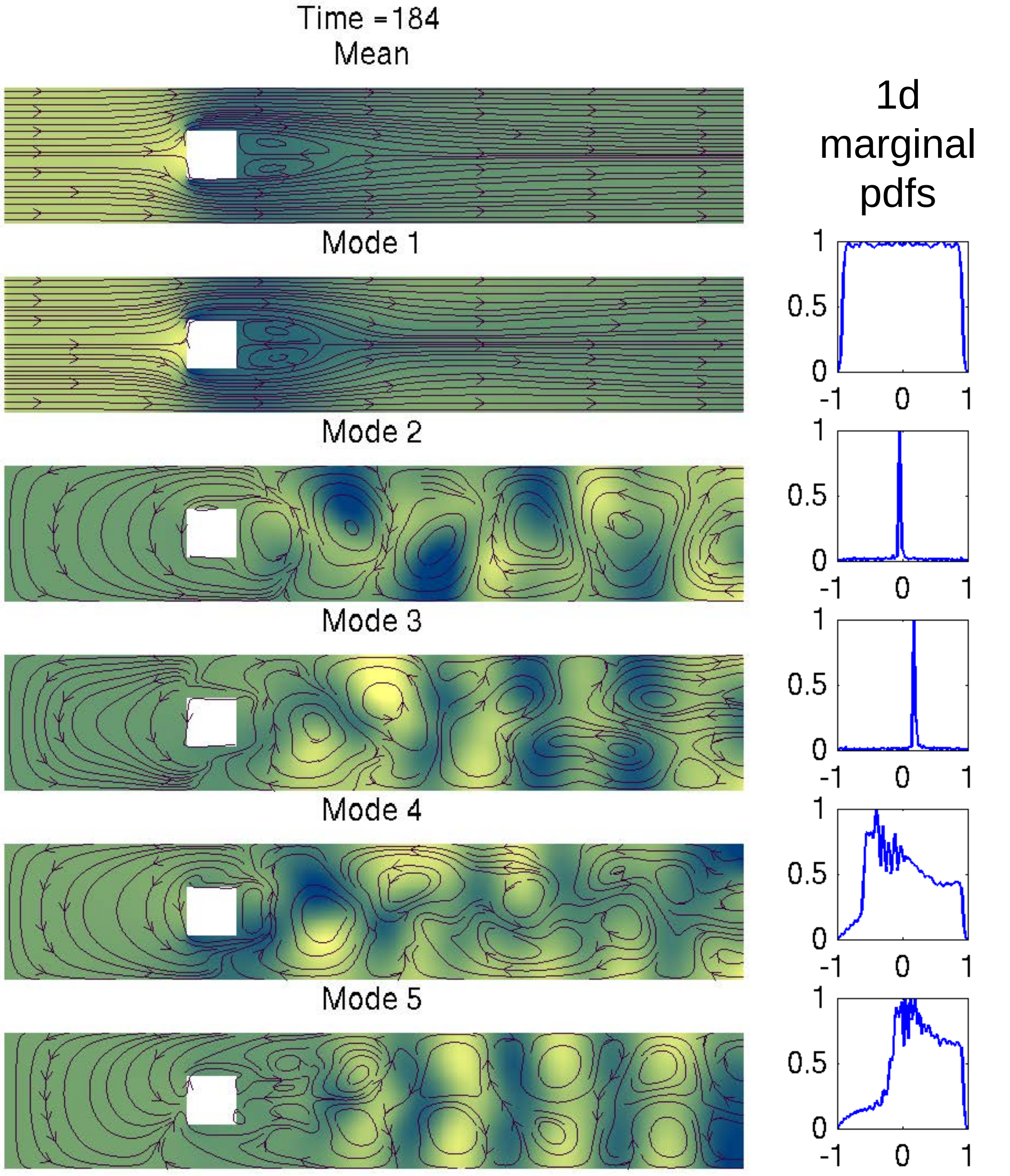


Visualizing Uncertainty in Fluid and Ocean Flows

Stochastic Flow behind a square cylinder

- Uncertain initial and boundary conditions
- Range of Reynolds number modeled with a single DO simulation
- Equivalent to 10^5 deterministic runs





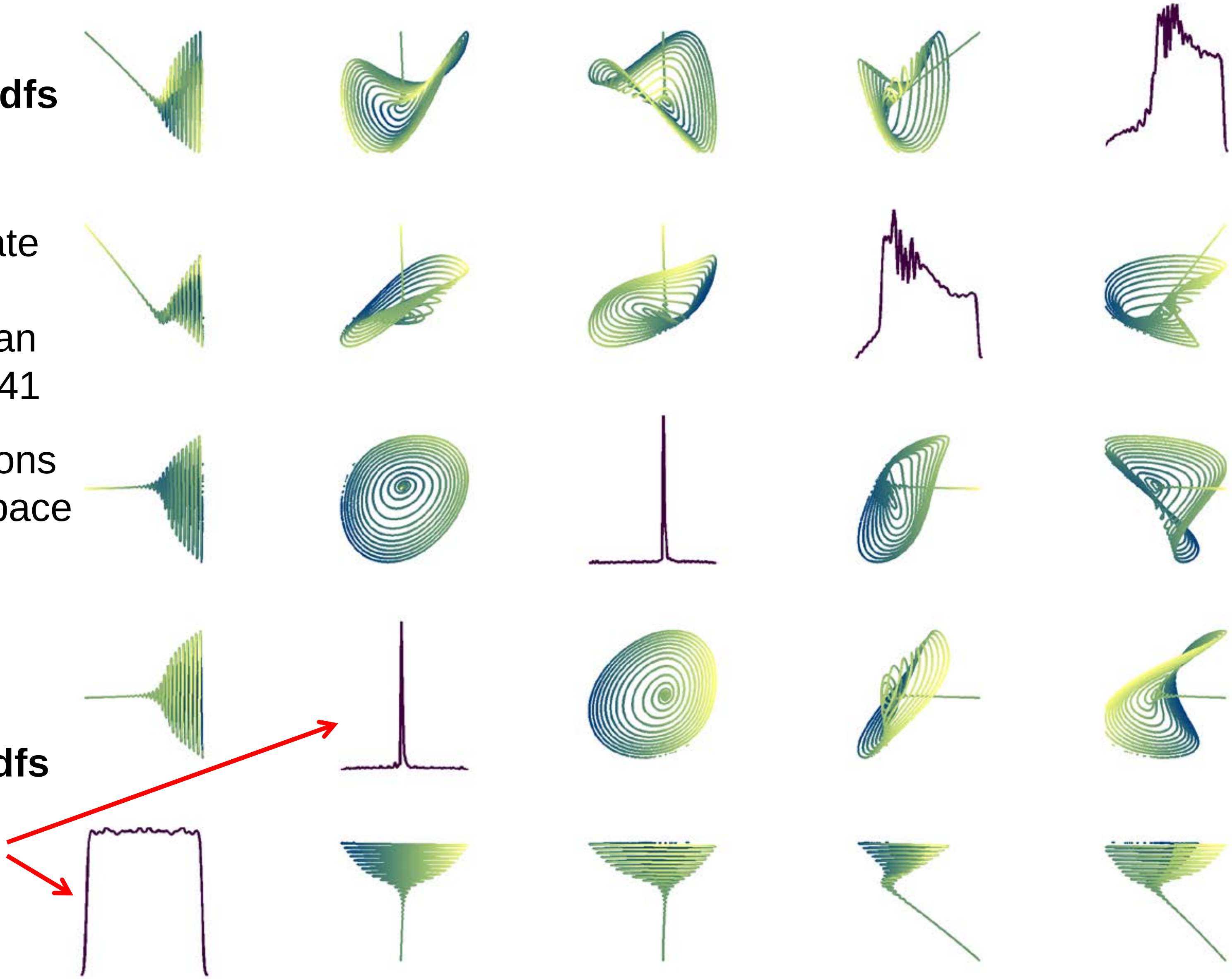
Visualizing Uncertainty in Fluid and Ocean Flows

2d marginal pdfs

- off-diagonal
- here, illustrate transition to non-Gaussian pdf at Re ~ 41
- 10^5 realizations in DO subspace

1d marginal pdfs

- on diagonal

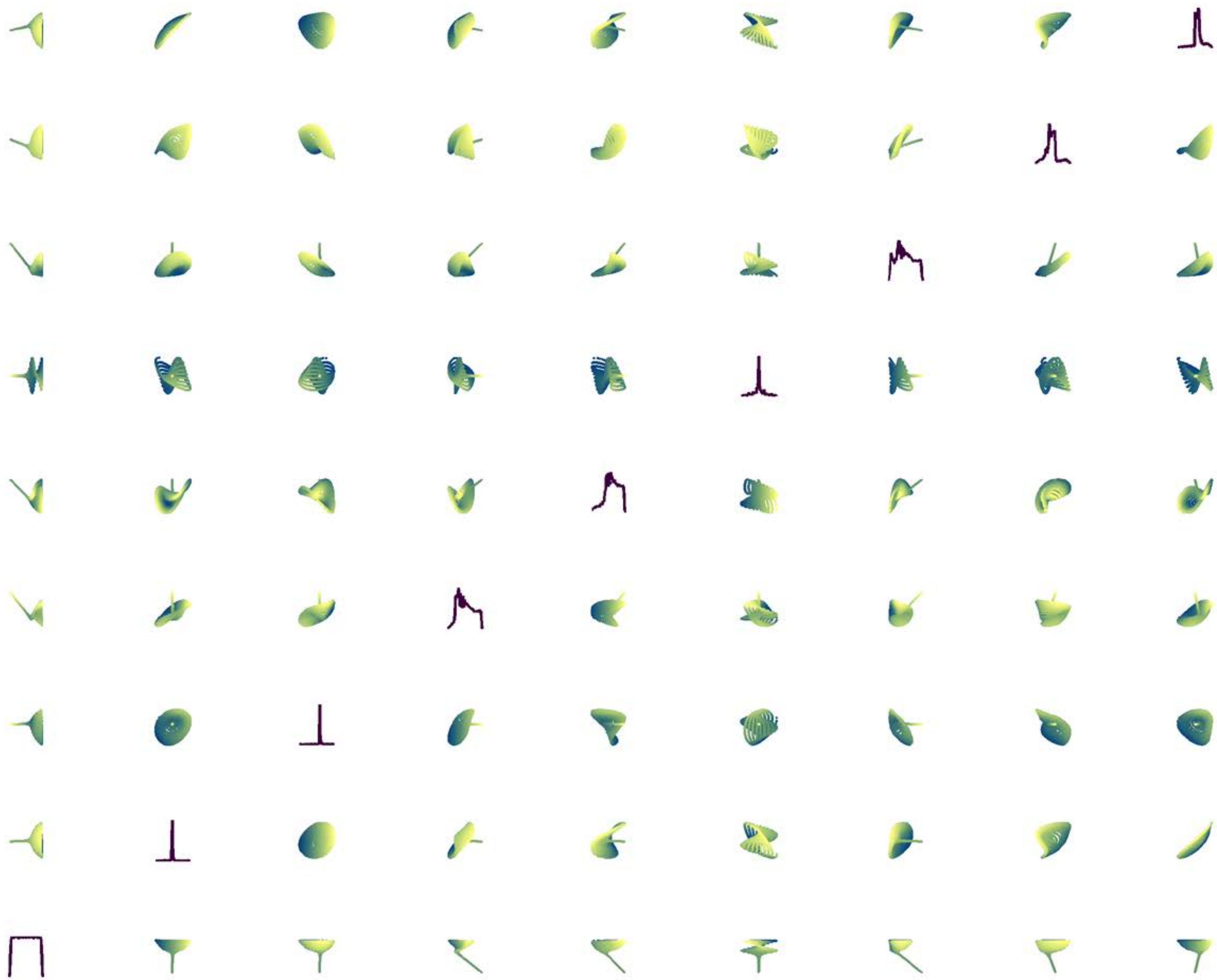


Visualizing Uncertainty in Fluid and Ocean Flows

2d marginal pdfs

- 9 DO modes
- Still 10^5 realizations in DO subspace

How to visualize 3d marginal pdfs to full 9d pdfs?



CONCLUSIONS

- **Prognostic DO Equations for Stochastic Fields**

- **GMM-DO Data Assimilation**

- **Visualizing Probability Densities of Ocean Fields?**
  - Scientific Visualization of Uncertainty
    - Overlays (pseud-color, contours, etc)
    - Histograms at each point in physical space, time-dependent
    - Key question: how to visualize pdfs in DO subspace?, but then in physical space?
  - Societal Visualization of Uncertainty
    - Overlays
    - Direct Volume rendering, Transparency
    - Glyphs, etc