Statistical Data Assimilation: Exact Formulation and Variational Approximations

Henry D. I. Abarbanel

Department of Physics

and

Marine Physical Laboratory (Scripps Institution of Oceanography)

Center for Theoretical Biological Physics

University of California, San Diego

habarbanel@ucsd.edu

==========

In collaboration with Philip Gill, Elizabeth Wong, Dan Creveling, Mark Kostuk, Jack Quinn, Bryan Toth, William Whartenby, Chris Knowlton

Data assimilation is a unidirectional communications system: data = transmitter → model = receiver. Synchronize the model and the data to achieve communication of information from data to model.

We present an exact formula for the probability of the performance of the communications channel. The exact formula is an integral along a path in the space of states and parameters of the receiver = model.

We analyze approximations to the path integral: variational and direct evaluations.

We present an example from a core climate model element: shallow water equations.

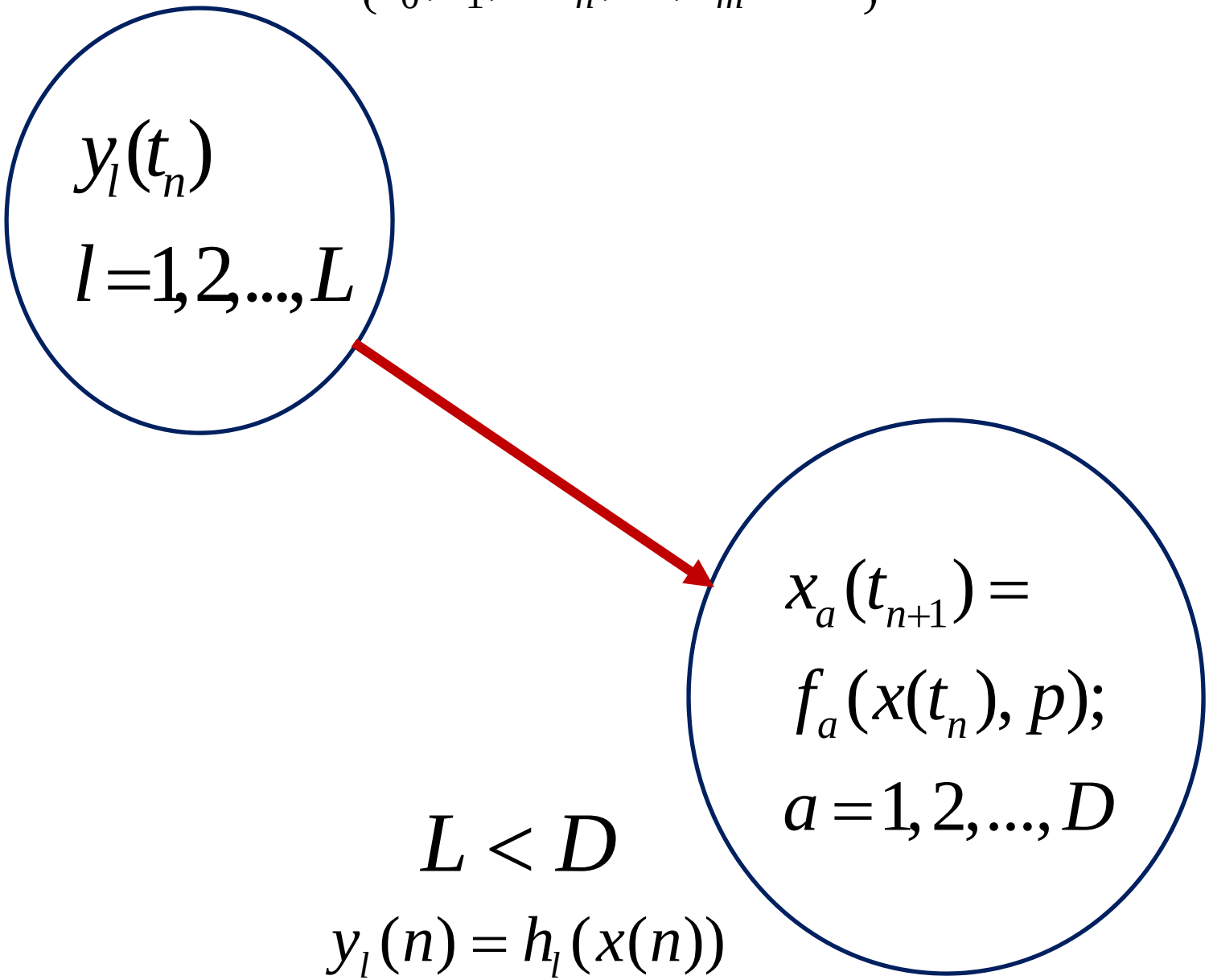
Common features in developing predictive models of observed nonlinear systems:

I Few of the many state variables are observable.

II We must estimate the ***unobserved*** *state variables* and the fixed parameters to make predictions, using $x(t_{n+1}) = f(x(t_n),p)$.

III Using observations to guide the dynamics to the right sector of phase space can allow prediction in chaos

$$t = \{t_0, t_1, \ldots t_n, \ldots, t_m = T\}$$



$$L < D$$

$$y_l(n) = h_l(x(n))$$

Our general interest is in making models of observed physical systems:

These models often have unknown parameters---conductivity, transport coefficients, reaction rates, coupling strengths, ….

These models often have unobservable state variables---gating variables for ion channels, population inversion in lasers, …

$$\frac{dx(t)}{dt} = F(x(t), p); \text{ need all x(0) and } p$$

$$x(t_{n+1}) = f(x(t_n), p); \text{ need all x}(t_0) \text{ and } p$$

Our discussion starts with the model and the data, and an interest in determining the unknown parameters and unobserved state variables.

Data Assimilation—transfer of information from observations to models:

Noisy Data, Errors in the Model—low resolution, …., unknown initial states when data acquisition begins

Exact Formulation as a Path Integral

$P(x(m)|Y(m))$ probability distribution for the model state at time t_m, given measurements $\{y(0), y(1), ..., y(m)\} = Y(m)$

Path Integral for $P(x(m)|Y(m))$:

$$P(x(m)|Y(m)) =$$

$$= \prod_{n=0}^{m-1} d^D x(n)\; e^{TMI(X,Y(m))} P(x(n+1)|x(n)) P(x(0))$$

$$= \prod_{n=0}^{m-1} d^D x(n)\; e^{-A_0(X,Y(m))} \qquad X = \{x(m), x(m-1), \ldots, x(0)\}$$

$$TMI(Y, Z(m)) = \sum_{n=0}^{m} CMI(y(n), z(n)|Z(n-1))$$

Total Conditional Mutual Information between path X and observations $Y(m)$

Path

What is different about this path integral: nonlinear "propagators" dissipative dynamics, orbits on strange attractors.

Action for State and Parameter Estimation

$$A_0(X \mid Y(m)) = -\sum_{n=0}^{m} \log\{MI(x(n), y(n) \mid Y(n-1))\}$$

$$-\sum_{n=0}^{m-1} \log\{P(x(n+1) \mid x(n))\} - \log\{P(x(0))\}$$

With the density of paths $\exp[-A_0(X\,|Y(m))]$,

we are able to evaluate any conditional expectation value of a function

$F(X) = F(x(m), x(m-1), ..., x(1), x(0), p)$ as

$$E[F(X)|Y] = \langle F(X) \rangle = \frac{\int dX e^{-A_0(X|Y(m))} F(X)}{\int dX e^{-A_0(X|Y(m))}}$$

Path is through state variable+parameter space

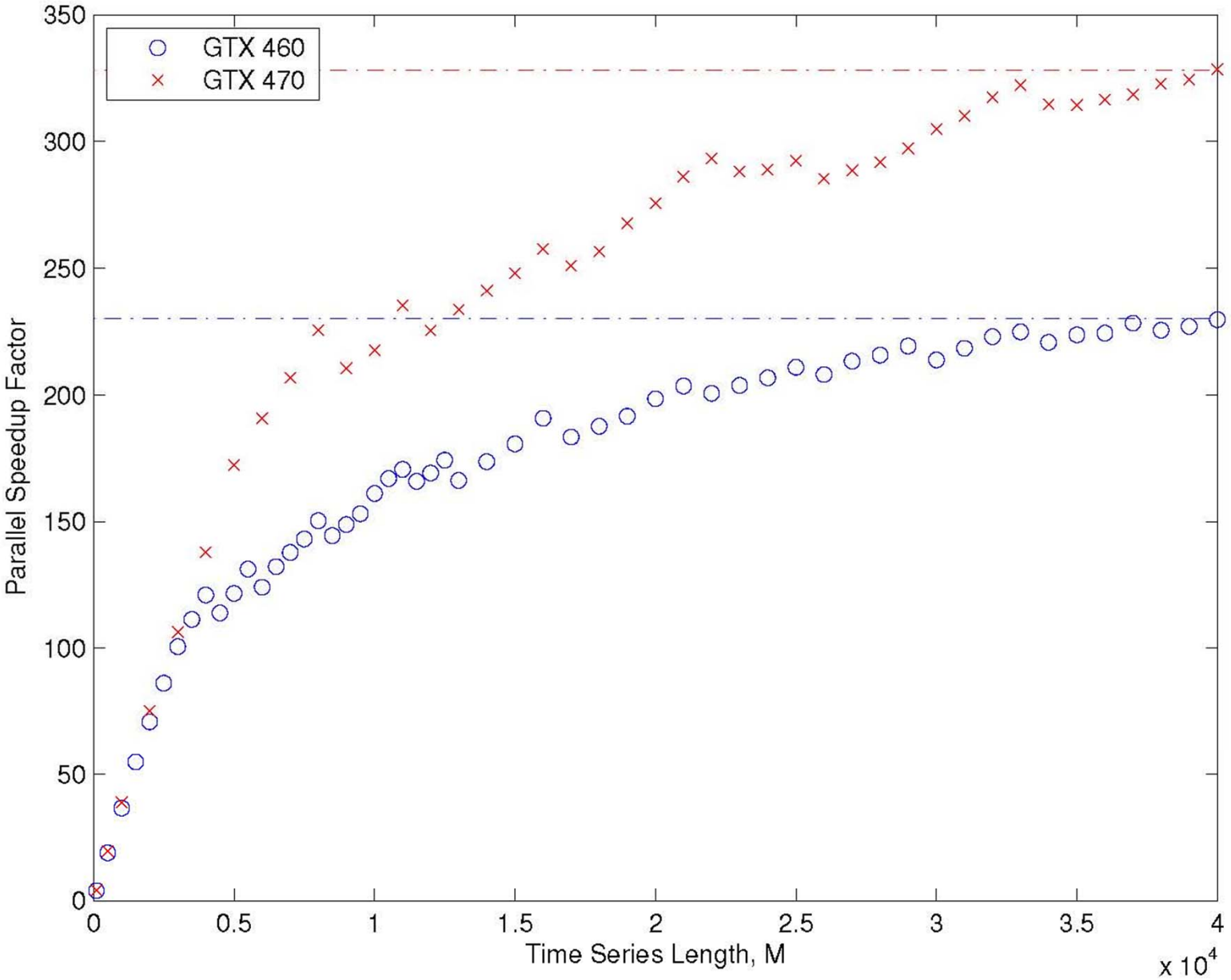
X={x(m),x(m-1),...x(0),p}

$$E[F(X)|Y] = \langle F(X) \rangle = \frac{\int dX e^{-A_0(X|Y(m))} F(X)}{\int dX e^{-A_0(X|Y(m))}}$$

We have explored Monte Carlo numerical evaluation of the integral representation of the data assimulation task. We produced the results here using single CPU machines.

This is eminently parallelizable. Same problem on GPU machines runs 60-300 times faster! Bigger problems utilize more GPU threads.

This is the source of numerical optimism.


GTX 460
GTX 470
Parallel Speedup Factor
Time Series Length, M
x 10^4
0
50
100
150
200
250
300
350
0
0.5
1
1.5
2
2.5
3
3.5
4

$$E[F(X)|Y] = \langle F(X) \rangle = \frac{\int dX e^{-A_0(X|Y(m))} F(X)}{\int dX e^{-A_0(X|Y(m))}}$$

Another approach is expansion about a saddle path

$$\frac{\partial A_0(X)}{\partial X_\alpha}|_{X=S} = 0 \qquad \alpha = 1, 2..., (m+1)D + K$$

This is 4DVar.

This is a numerical optimization problem--see Gill tomorrow.

Significant problems with this when trajectories are chaotic.

Path integral representation allows corrections to 4DVar:

do Gaussian integral about $X = S$.

This requires $\frac{\partial^2 A_0(X)}{\partial X_\alpha \partial X_\beta}|_{X=S}$

Lorenz96 Model

Dynamical variables--longitudinal `activity'--no vertical levels no latitude variations.

$y_a(t) \quad a = 1, 2, ..., D$

$$y_{-1}(t) = y_{D-1}(t) \qquad y_0(t) = y_D(t) \quad y_{D+1}(t) = y_1(t)$$

$$\frac{dy_a(t)}{dt} = y_{a-1}(t)(y_{a+1}(t) - y_{a-2}(t)) - y_a(t) + \text{Forcing}$$

We explored D = 5. Chaotic at Forcing = 7.9; we used Forcing = 8.17

$$\frac{dy_1(t)}{dt} = y_5(t)(y_2(t) - y_4(t)) - y_1(t) + F + u_1(t)(x_1(t) - y_1(t))$$

$$\frac{dy_2(t)}{dt} = y_1(t)(y_3(t) - y_5(t)) - y_2(t) + F$$

$$\frac{dy_3(t)}{dt} = y_2(t)(y_4(t) - y_1(t)) - y_3(t) + F + u_2(t)(x_3(t) - y_3(t))$$

$$\frac{dy_4(t)}{dt} = y_3(t)(y_5(t) - y_2(t)) - y_4(t) + F$$

$$\frac{dy_2(t)}{dt} = y_4(t)(y_1(t) - y_3(t)) - y_5(t) + F$$

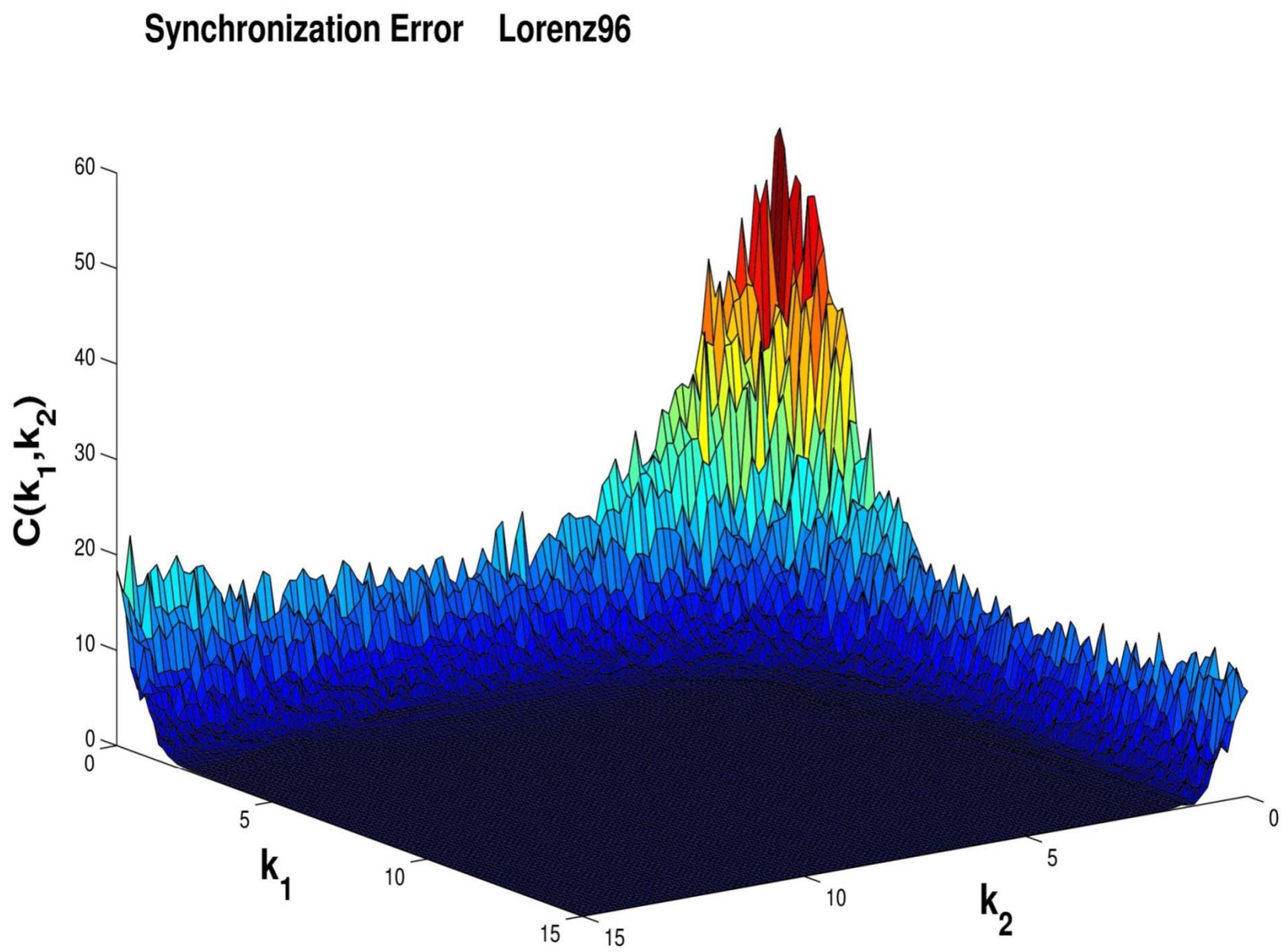
Model does not synchronize with 'data' (*x(t)*) with only one coupling: two positive Conditional Lyapunov Exponents → two different couplings for data are required

Calculate the synchronization error or cost function

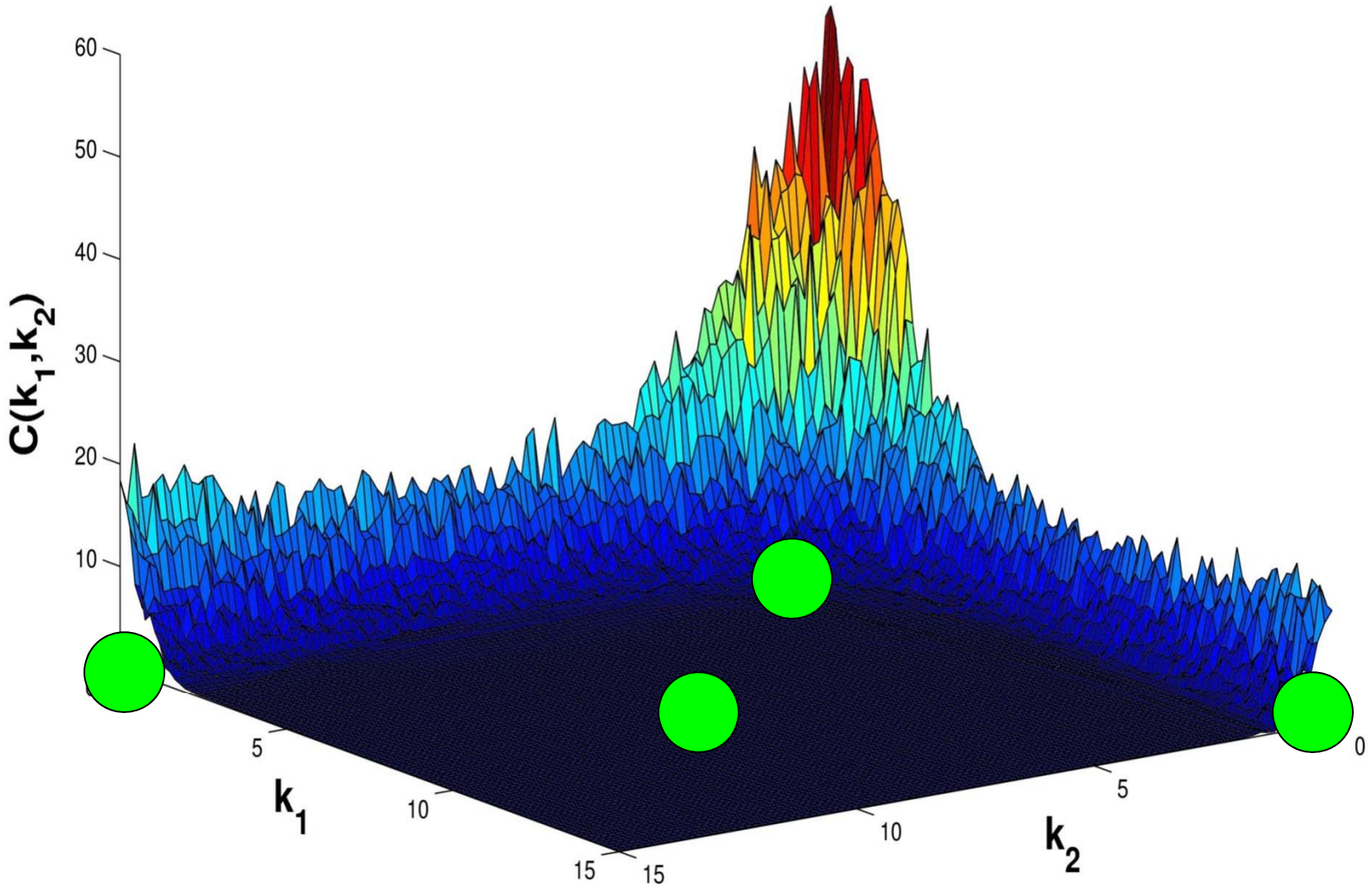
$$C(k_1, k_2) = \sum_{n=0}^{N} \sum_{l=1}^{L} (x_l(t_n) - y_l(t_n))^2$$

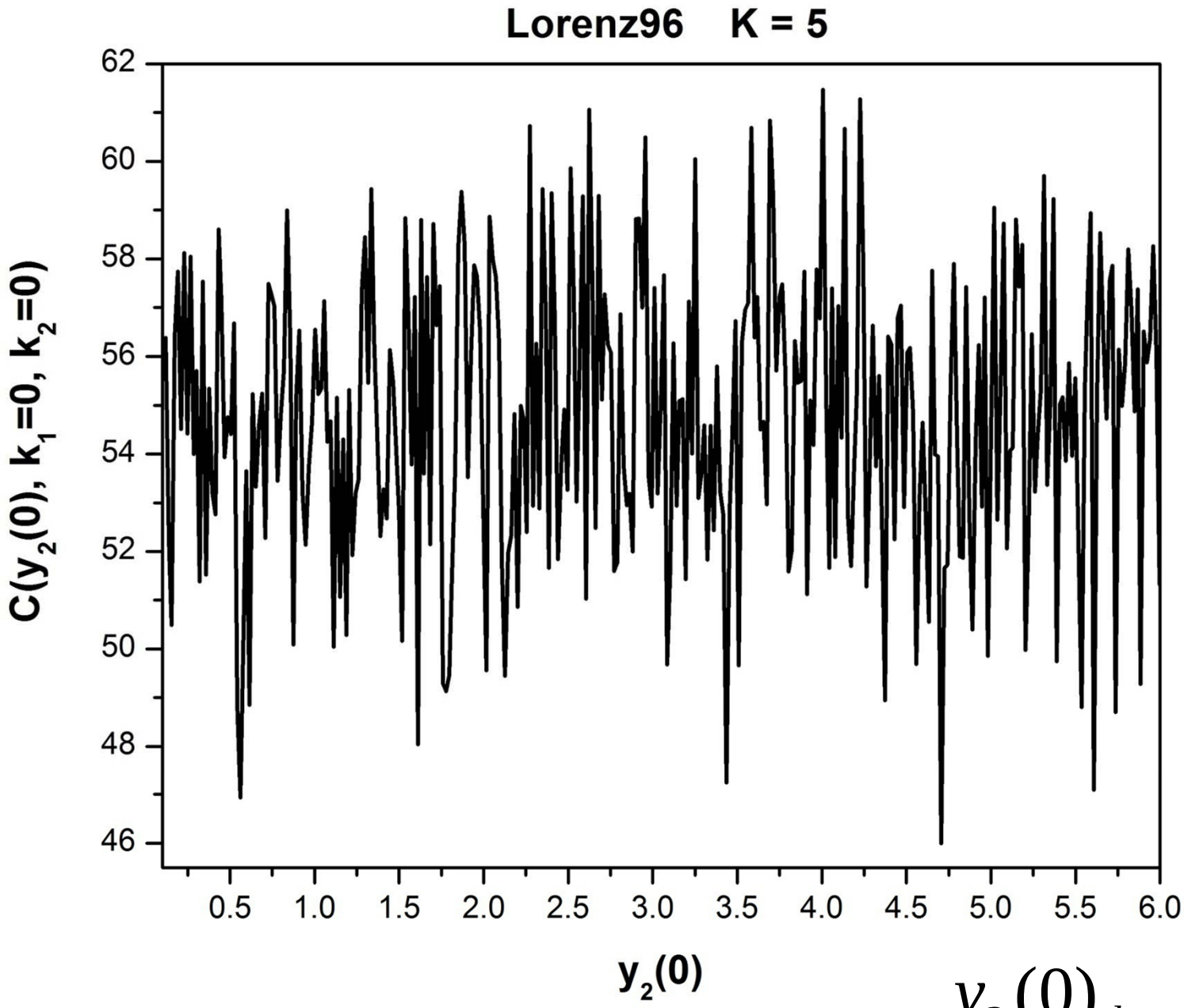
as a function of the couplings ($u_1(t)=k_1, u_2(t) = k_2$).

Examine the dependence on an initial condition, here $y_2(0)$ for various regions of (k_1, k_2).

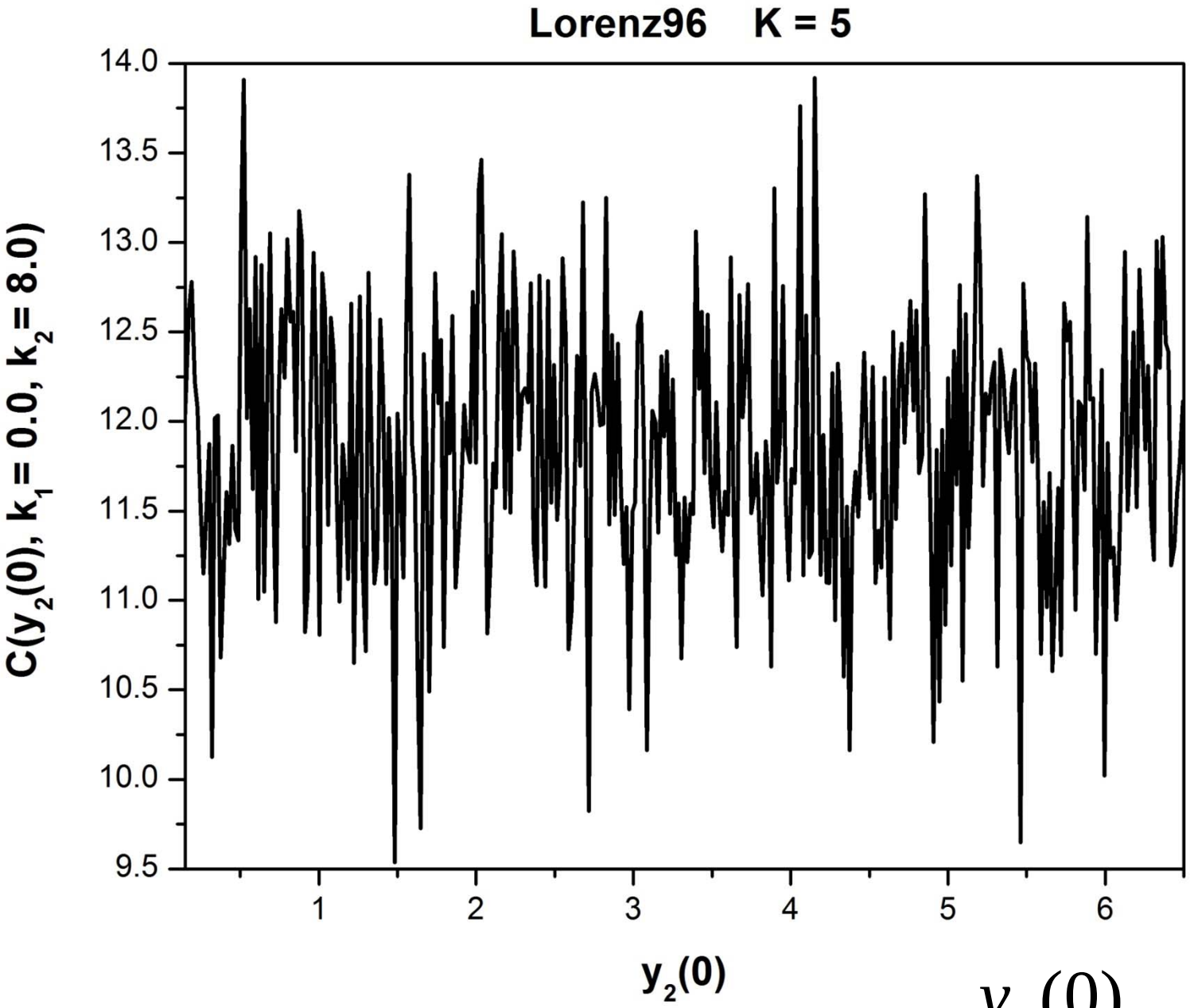


Synchronization Error Lorenz96

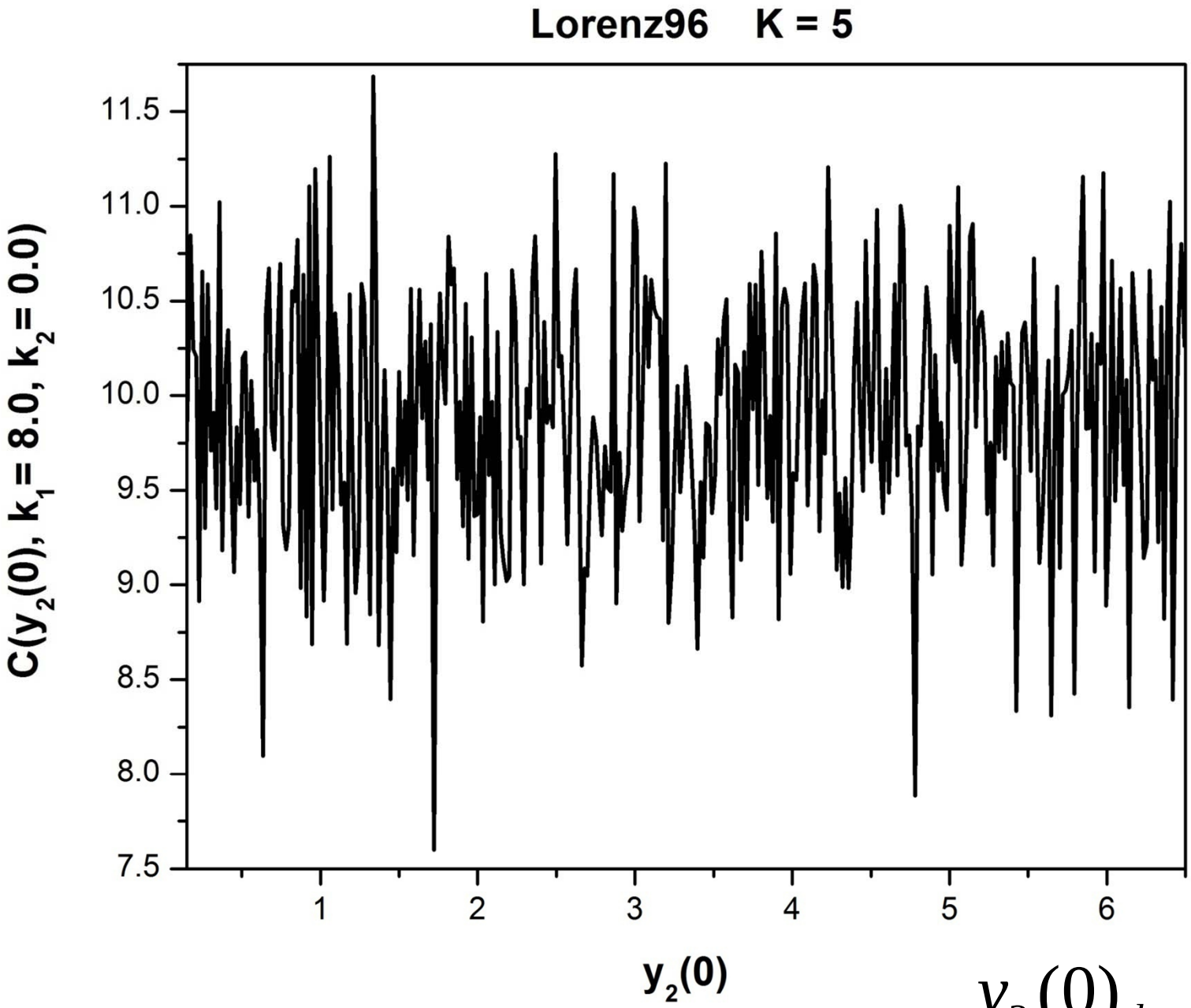




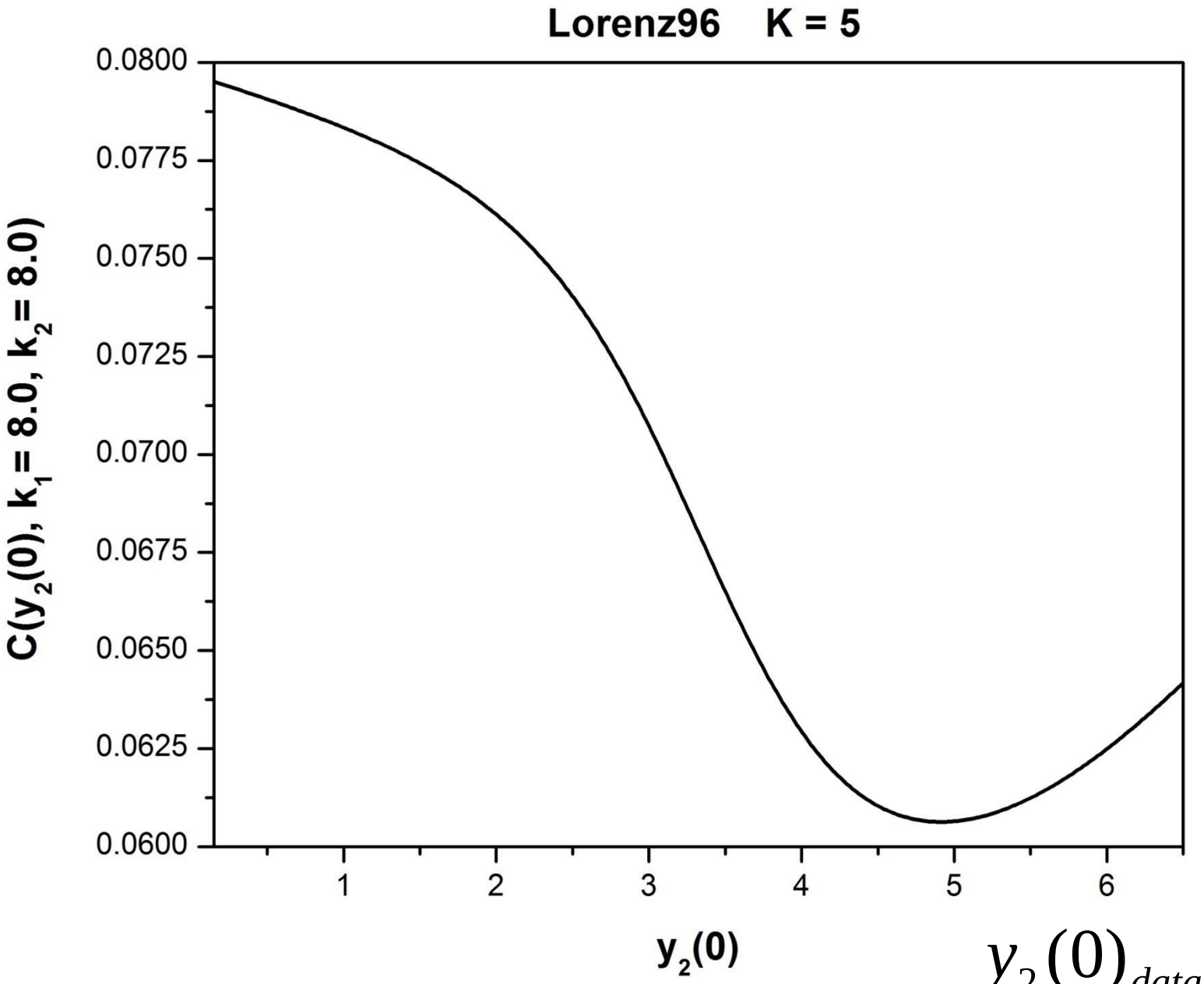
$$y_2(0)_{data} \approx 4.9$$



$y_2(0)_{data} \approx 4.9$



$$y_2(0)_{data} \approx 4.9$$



$$y_2(0)_{data} \approx 4.9$$

Dynamical State and Parameter Estimation (DSPE)

Minimize:

$$C(y,u,p)=\frac{1}{2T}\int_0^T\{(x_1(t)-y_1(t))^2+u(t)^2\}$$

Subject to the equality constraints:

$$\frac{dy_1(t)}{dt}=F_1(y_1(t),y_R(t),p)+u(t)(x_1(t)-y_1(t))$$

$$\frac{dy_R(t)}{dt}=F_R(y_1(t),y_R(t),p)$$

Use variational answer, saddle path, from optimization

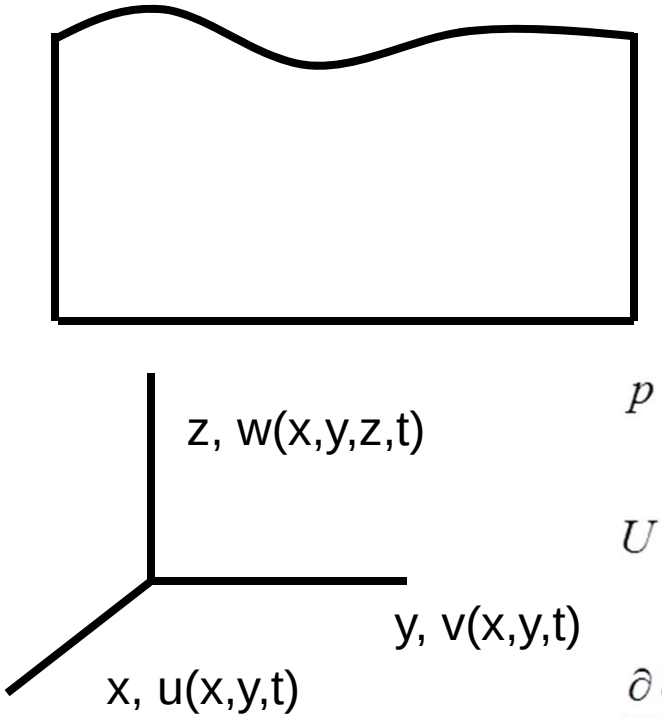
Saddle path

$$\frac{\partial A_0(X)}{\partial X_\alpha}\Big|_{X=S} = 0 \qquad \alpha = 1, 2..., (m+1)D + K$$

As a first approximation and use path integral to evaluate fluctuations about this optimal path. We use IPOPT to solve optimization problem via direct method.

Shallow Water Equations

$$H(x,y,t) = H_0 + h(x,y,t)$$



$$p(x,y,z,t) = \rho_0 g(H(x,y,t) - z)$$

$$U(x,y,t) = \{u(x,y,t), v(x,y,t)\}$$

$$\frac{\partial U(x,y,t)}{\partial t} + U(x,y,t) \bullet \nabla U(x,y,t) = -g\nabla H(x,y,t) +$$

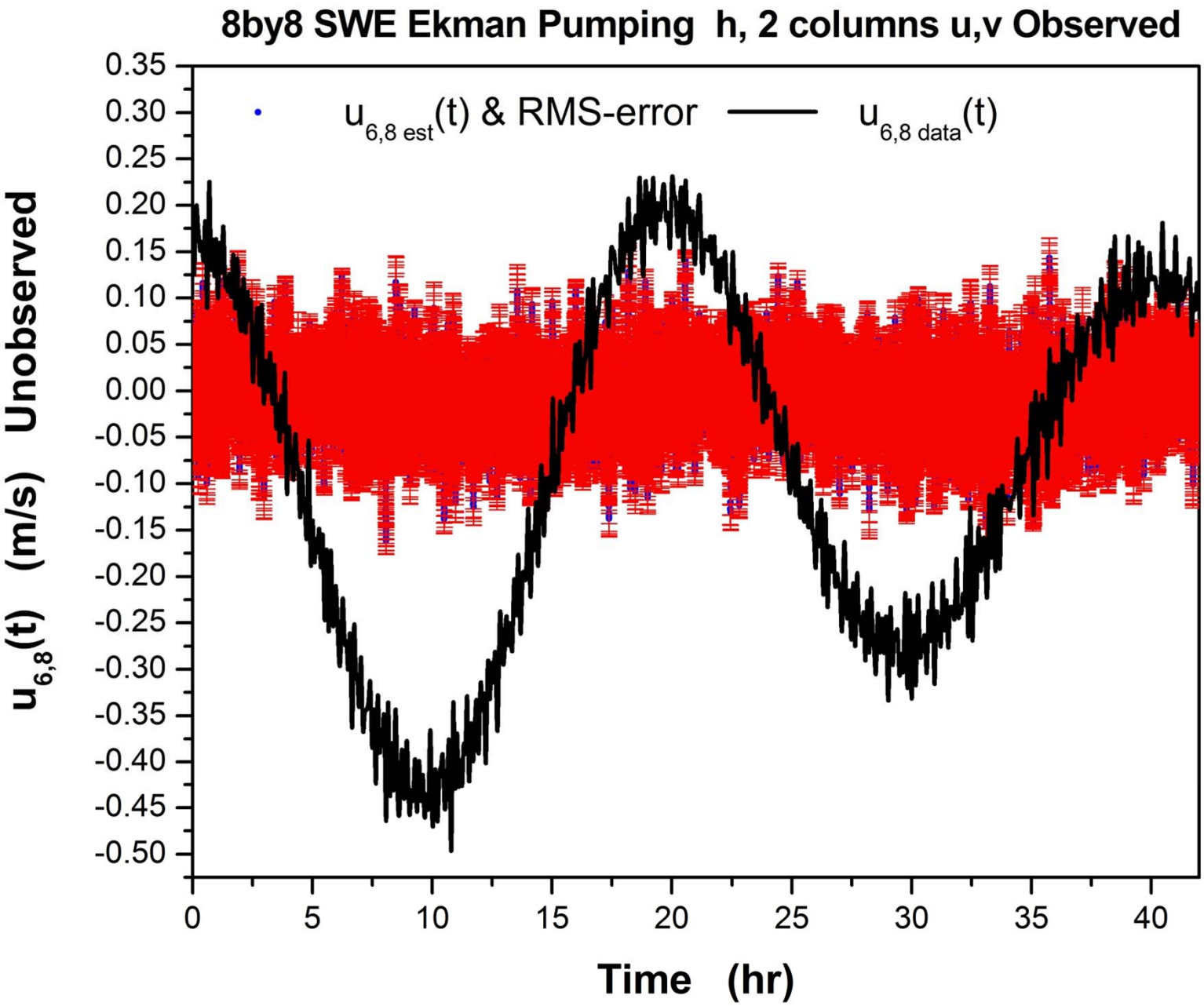
$$-RU(x,y,t) + \nu\nabla^2 U(x,y,t) + f(y)\hat{z} \times U(x,y,t)$$

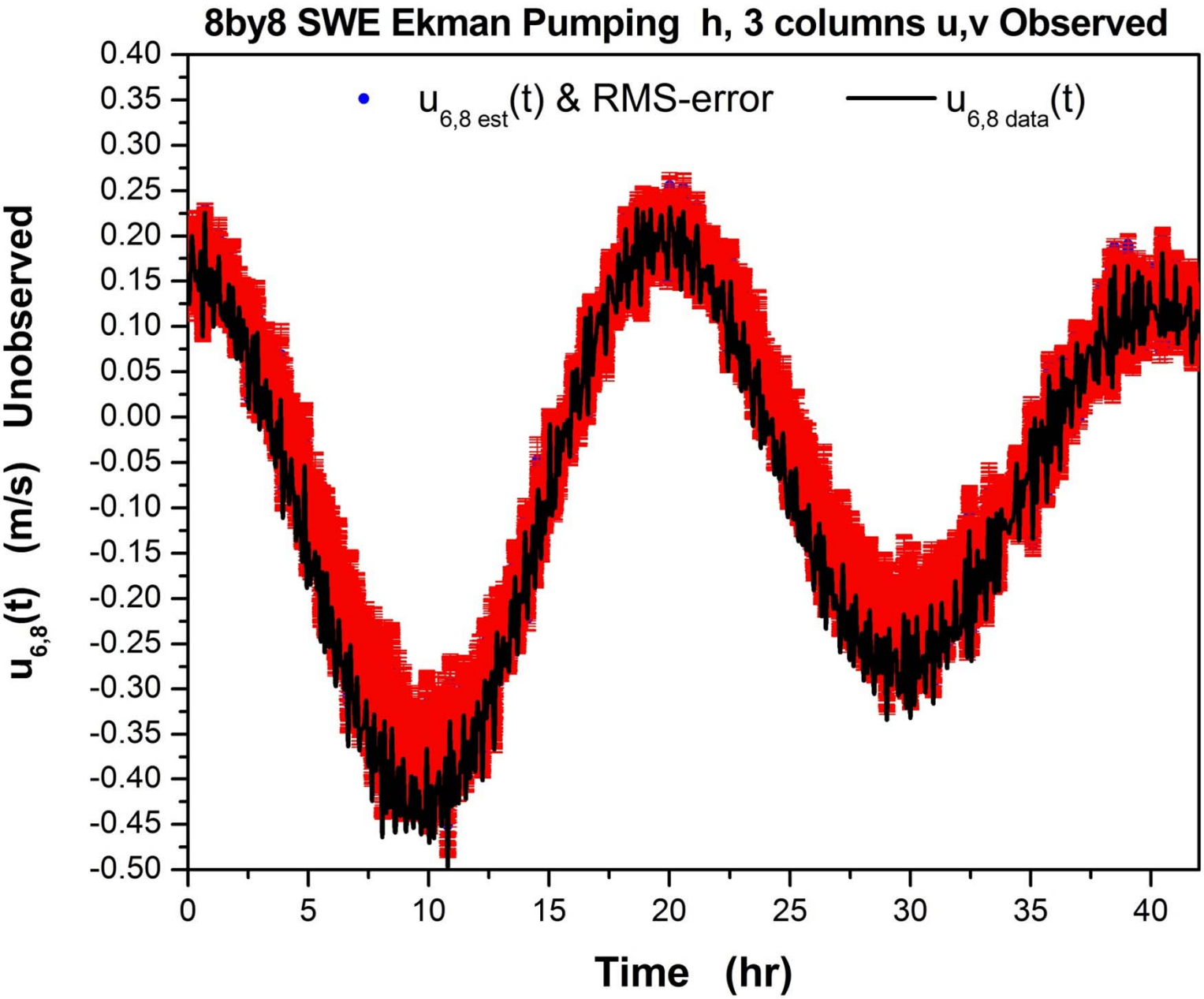
$$\frac{\partial H(x,y,t)}{\partial t} + \nabla \bullet [U(x,y,t)H(x,y,t)] = w_{\text{Eckman}}$$

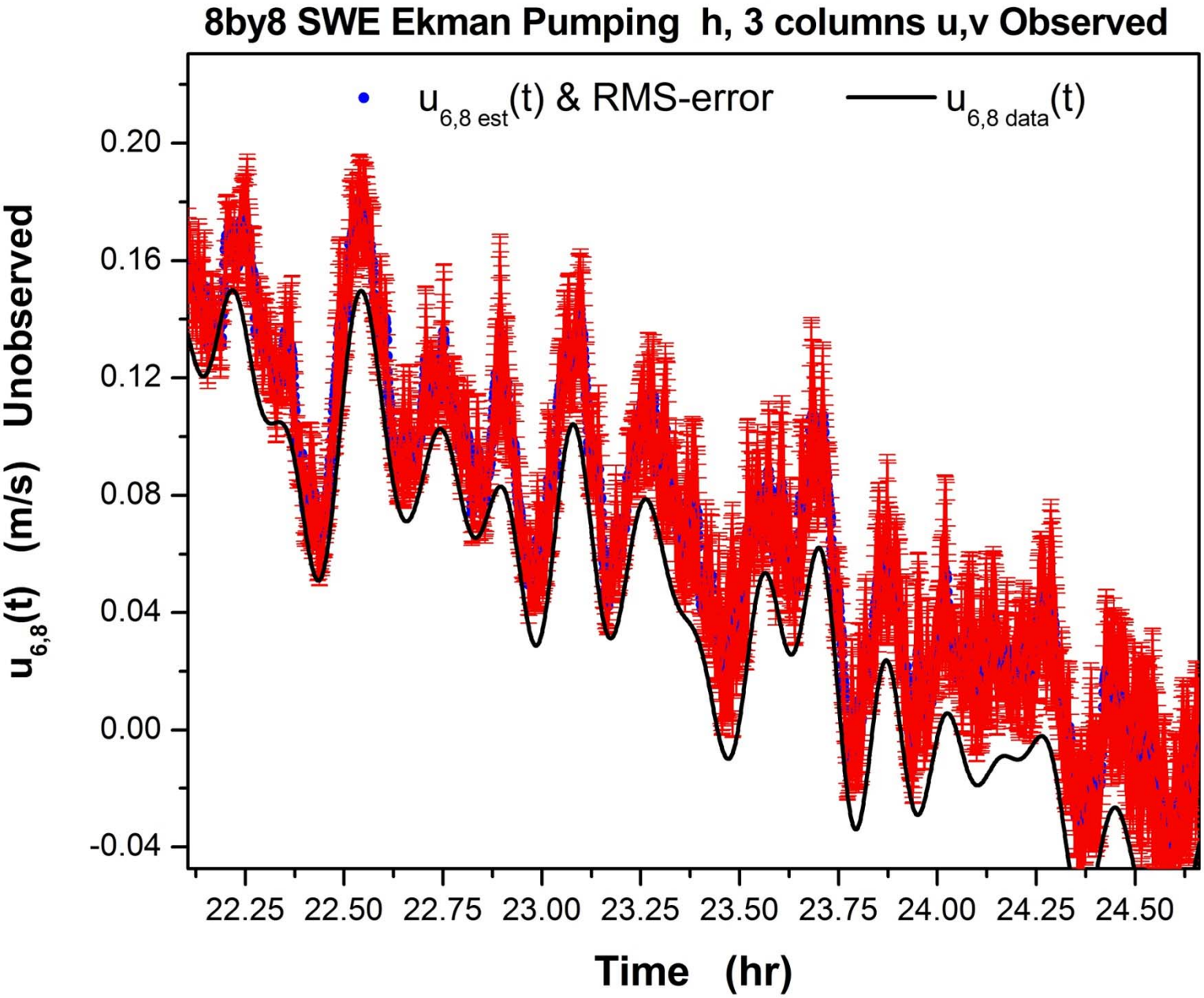
Twin Experiment: Generate ‘data’ from NbyN shallow water equation. Present $L \leq N^2$ observed variables: u, v, h. How big must L be to allow estimation of other unobserved state variables? How many measurements must one make to accurately estimate unobserved states?

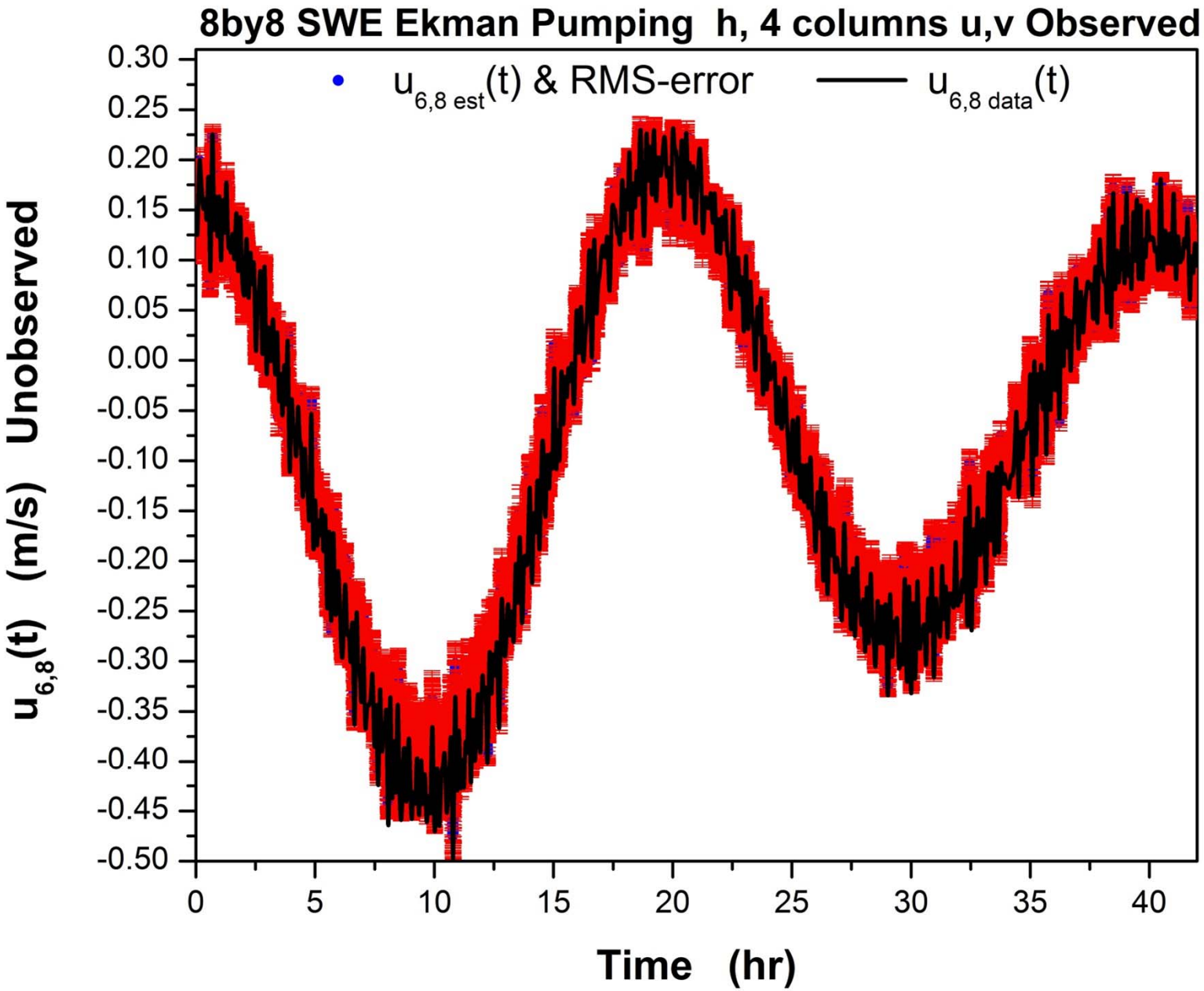
For 8 by 8 example with 192 state variables, the number is 112 out of 192, about 60%.

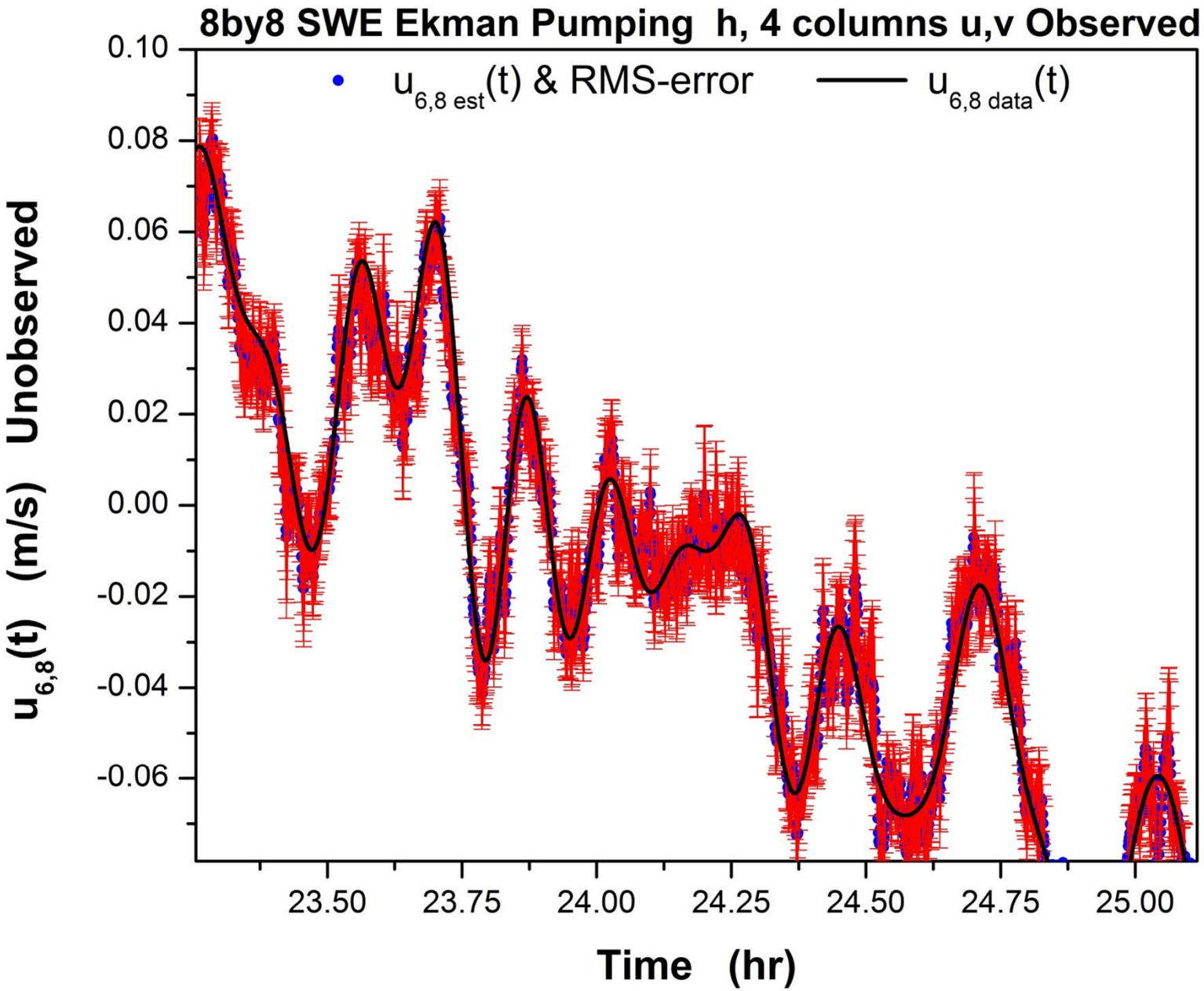
In 8 by 8 shallow water equations, present all measurements of h(x,y,t) as well as 2, 3, 4, .. Columns of data from wind velocities. Columns contain data on shears in the wind fields.


8by8 SWE Ekman Pumping h, 2 columns u,v Observed
u6,8 est(t) & RMS-error
u6,8 data(t)
u6,8(t) (m/s) Unobserved
Time (hr)
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
-0.40
-0.45
-0.50
0
5
10
15
20
25
30
35
40


8by8 SWE Ekman Pumping h, 3 columns u,v Observed
u6,8 est(t) & RMS-error
u6,8 data(t)
u6,8(t) (m/s) Unobserved
Time (hr)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
-0.40
-0.45
-0.50
0
5
10
15
20
25
30
35
40


8by8 SWE Ekman Pumping h, 3 columns u,v Observed
u6,8 est(t) & RMS-error
u6,8 data(t)
u6,8(t) (m/s) Unobserved
0.20
0.16
0.12
0.08
0.04
0.00
-0.04
22.25
22.50
22.75
23.00
23.25
23.50
23.75
24.00
24.25
24.50
Time (hr)


8by8 SWE Ekman Pumping h, 4 columns u,v Observed
u6,8 est(t) & RMS-error
u6,8 data(t)
u6,8(t) (m/s) Unobserved
Time (hr)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
-0.40
-0.45
-0.50
0
5
10
15
20
25
30
35
40


8by8 SWE Ekman Pumping h, 4 columns u,v Observed
u6,8 est(t) & RMS-error
u6,8 data(t)
u6,8(t) (m/s) Unobserved
Time (hr)
0.10
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
23.50
23.75
24.00
24.25
24.50
24.75
25.00

Effective Actions for Path Integrals

$$< F(X) >= \int dX F(X) e^{-A_0(X)}$$

stationary path approximation

$$F(X) = F(S) + (X - S) \bullet \frac{\partial F(X)}{\partial X}|_S + \frac{1}{2}(X - S)\frac{\partial^2 F(X)}{\partial X \ \partial X}(X - S) + \ldots$$

4DVar is $\frac{\partial F(X)}{\partial X}|_S = 0,$

$$F(X) = F(S) + \frac{1}{2}(X - S)\frac{\partial^2 F(X)}{\partial X \ \partial X}(X - S) + \ldots$$

Performing an integral involves an optimization problem

$$< F(X) > \approx F(S)\sqrt{\frac{(2\pi)^D}{\det(F''(S)}}$$

Effective Actions for Path Integrals Generalize 4Dvar to another, exact variational principle

Generating function for moments along the path

$$e^{C(K)} = \int dX \; e^{-A_0(X)+K\bullet X}$$

$$\frac{\partial C(K)}{\partial K_\alpha}\Big|_{K=0} = E[X] = \frac{\int dX \; \mathrm{X}_\alpha \; e^{-A_0(X)}}{\int dX \; e^{-A_0(X)}}$$

$$\frac{\partial^2 C(K)}{\partial X_\alpha \partial X_\beta} = \frac{\int dX \; \mathrm{X}_\alpha \mathrm{X}_\beta \; e^{-A_0(X)} - \left\{\int dX \; \mathrm{X}_\alpha \; e^{-A_0(X)}\right\}\left\{\int dX \; \mathrm{X}_\beta \; e^{-A_0(X)}\right\}}{\int dX \; e^{-A_0(X)}}$$

and similarly for higher moments

Now trade in the `current' K_α for the mean field φ_α via

$$\frac{A(\varphi)}{\mu} = \frac{-C(K) + K \bullet \varphi}{\mu}$$

$$\frac{\partial C(K)}{\partial K_\alpha} = \varphi_\alpha \quad \text{implies} \quad \frac{\partial A(\varphi)}{\partial \varphi_\alpha} = K_\alpha$$

$$e^{-\frac{A(\varphi)}{\mu}} = \int dX \; e^{\frac{-A_0(X) + K \bullet (X - \varphi)}{\mu}}$$

$$= \int dX \; e^{-\frac{A_0(X) + \frac{\partial A(\varphi)}{\partial \varphi} \bullet (X - \varphi)}{\mu}}$$

$$\frac{\partial^2 A(\varphi)}{\partial \varphi_\alpha \partial \varphi_\beta} = \left[\frac{\partial^2 C(K)}{\partial X_\alpha \partial X_\beta} \right]^{-1}$$

$A(\varphi)$ contains all the moment information. It is like the Free Energy in statistical physics.

$\frac{\partial A_0(X)}{\partial X_\alpha} = 0$ gives the `bare' orbit.

$\frac{\partial A(\varphi)}{\partial \varphi_\alpha} = 0$ gives the complete expected state variable <X> including all corrections due to fluctuations in measurements and model error.

$$e^{-\frac{A(\varphi)}{\mu}} = \int dX \; e^{-\frac{A_0(X)+\frac{\partial A(\varphi)}{\partial \varphi}\bullet(X-\varphi)}{\mu}}$$

$$A(\varphi) = \sum_{l=0} \mu^l A_l(\varphi)$$

$$e^{-\sum_{l=1} \mu^{l-1} A_l(\varphi)} = \sqrt{\frac{(2\pi\mu)^D}{\det(A_0^{''}(\varphi))}} \int dU e^{-U^2} e^{-\sum_{r=3} 2^r \mu^{r-1} (\frac{U}{\sqrt{A_0^{''}(\varphi)}})^r + \sqrt{2} \sum_{l=1} \mu^{l-1/2} \frac{\partial A_l(\varphi)}{\partial \varphi} \frac{U}{\sqrt{A_0^{''}(\varphi)}}}$$

$$A(\varphi) = A_0(\varphi) - \frac{\mu}{2} tr(\log A_0^{''}(\varphi)) + O(\mu^2)$$

$\frac{\partial A(\varphi)}{\partial \varphi} = 0$ is the generalized 4Dvar

Using the integral representation of the Data Assimilation questions, one can ask how many degrees of freedom are actually required to capture the physics---when do we stop improving the resolution ?

By a consideration of the continuum limit in space and time, one can integrate out the high frequency and high wavenumber components and provide a “universal” parametrization of the answer, then evaluate the remaining structure now containing all the Physics on the desired scales.

Summary

Data assimilation is a unidirectional communications system: data = transmitter, model = receiver. Synchronize the model and the data to achieve communication of information from data to model.

We gave an exact integral representation of the solution of the master equation for the conditional probability distribution in state space. It is an integral along a path in the space of states and parameters of the receiver = model.

Evaluation of high dimensional path integral using Monte Carlo methods. Works well in “low” dimensions. Eminently parallelizable, so numerical possibilities for LARGE models are available.