Reconstruction of Effective Stochastic Dynamics from Data

Daan Crommelin

Courant Institute of Mathematical Sciences
New York University

NCAR, 2 March 2006

joint work with Eric Vanden-Eijnden

- Inverse modeling / stochastic model identification
- Numerical examples
- Reconstruction of homogenized dynamics
- Application to subgrid-scale parameterizaton in Lorenz 96 model

Data-driven (re)construction of stochastic models

Why?

– Systematic derivation difficult or impossible
– Interest in modeling an observed variable
– Need for describing effective dynamics

Inverse modeling

What is the best stochastic model that fits given timeseries?

- Diffusion process / Stochastic Differential Equation (SDE)

- Continuous-time Markov chain: $\frac{d}{dt}\rho_i = \sum_j L_{ij}\rho_j, \quad \rho(t+\Delta t) = e^{\Delta t L}\rho(t)$

Diffusion reconstruction

Reconstruct drift vector b and diffusion matrix a from timeseries $\{X_t\}$

- Direct estimation (e.g. Siegert et al., Friedrich et al., Berner and Branstator, Sura)

$$\begin{cases} b(x) = \lim\limits_{\Delta t \to 0} \mathbb{E}_x \left(X_{\Delta t} - x \right) \\ a(x) = \lim\limits_{\Delta t \to 0} \mathbb{E}_x \left(X_{\Delta t} - x \right) \otimes \left(X_{\Delta t} - x \right) \end{cases}$$

$\Delta t \to 0$ limit problematic (availability of data, non-Markov effects at short timescales); no flexibility

- Maximum Likelihood estimation (e.g. Heyde, Sørensen, Aït-Sahalia)

Problematic if data has no underlying diffusion or if sampling interval too large.

New approach:

Process characterised by generator L
(backward Fokker-Planck operator or Markov chain generator matrix)

Spectrum of L: $\{\psi_k, \phi_k, \lambda_k\}$ (ψ_1 = invariant distribution)

Construct L such that its spectrum resembles a given reference spectrum as closely as possible

Given a reference spectrum $\{\psi_k^{\mathsf{r}}, \phi_k^{\mathsf{r}}, \lambda_k^{\mathsf{r}}\}$, construct optimal generator L by minimizing the object function

$$E = \sum_{k=1}^{K} \left(\alpha_k \| L^* \psi_k^{\mathsf{r}} - \lambda_k^{\mathsf{r}} \psi_k^{\mathsf{r}} \|^2 + \beta_k \| L \phi_k^{\mathsf{r}} - \lambda_k^{\mathsf{r}} \phi_k^{\mathsf{r}} \|^2 + \gamma_k \left| \langle \psi_k^{\mathsf{r}}, L \phi_k^{\mathsf{r}} \rangle - \lambda_k^{\mathsf{r}} \right|^2 \right)$$

- Continuous-time Markov chain: $E = E(L)$
  Minimize E under variation of matrix elements L_{ij}

- Diffusion process: $L = L(a, b) \longrightarrow E = E(a, b)$
  Minimize E under variation of a, b

- Weights $\alpha_k, \beta_k, \gamma_k$ allow to put emphasis on (e.g.) leading eigenmodes

- E is **quadratic** in L or a, b
  $\rightarrow$ quadratic programming problem; unique minimum

- Constraints: $L_{i,j\neq i} \geq 0$, or a positive semi-definite $\rightarrow$ convex domain

- Special processes / parametric estimation:
  - E.g., birth-death Markov chain $\rightarrow$ L tri-diagonal
  - Expand a, b on a basis: $a(x) = \sum_m a_m f_m(x),\ b(x) = \sum_n b_n g_n(x)$,
    minimize E under variation of the a_m, b_n

  (flexibility; dimension reduction)

The reference spectrum $\{\psi_k^{\mathsf{r}}, \phi_k^{\mathsf{r}}, \lambda_k^{\mathsf{r}}\}$

How to obtain? Construct Markov chain from timeseries; sampling interval h. Spectrum of stochastic matrix P: $\{\psi_k^{\mathsf{r}}, \phi_k^{\mathsf{r}}, \Lambda_k^{\mathsf{r}}\}$

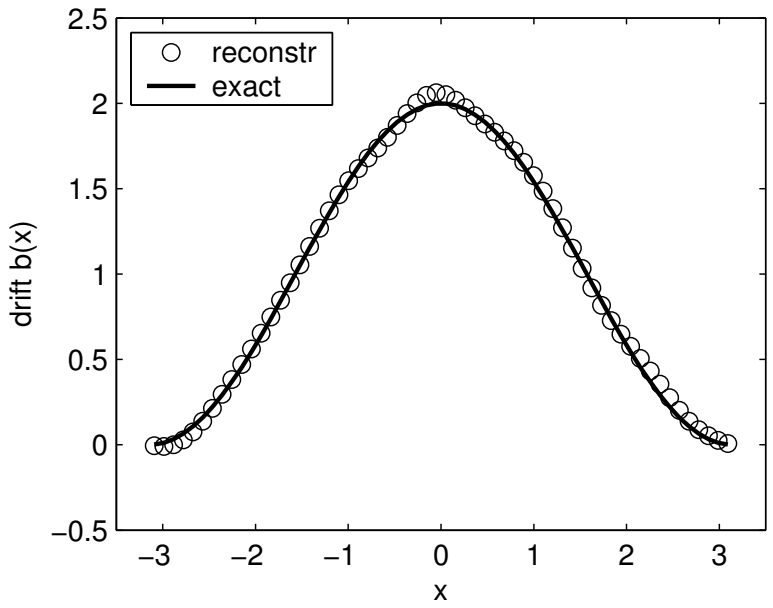
$\{\psi_k^{\mathsf{r}}, \phi_k^{\mathsf{r}}\}$: (discrete approximations of) reference eigenvectors
$\{\lambda_k^{\mathsf{r}} = h^{-1} \log \Lambda_k^{\mathsf{r}}\}$: reference eigenvalues

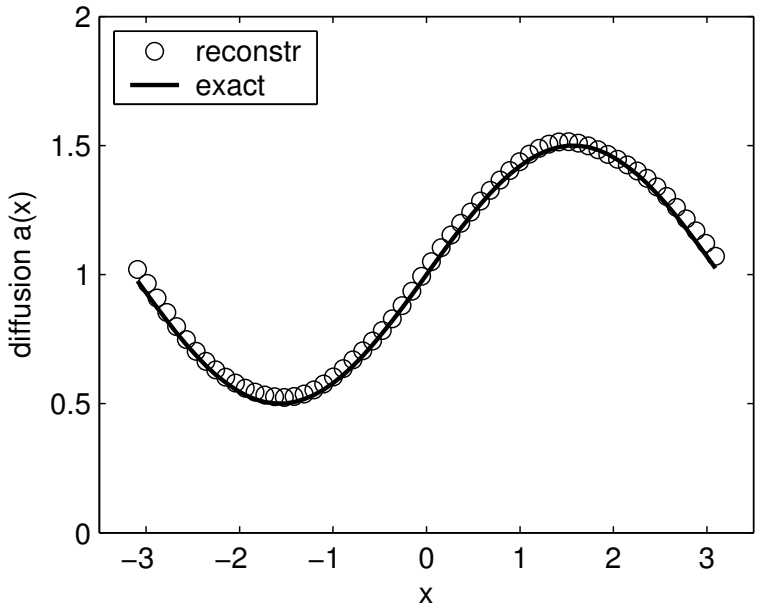
What h to choose? Too short: non-Markov effects
Too long: sampling error

How many needed? In practice: only leading modes
E.g., 1-dim SDE: $k = 1, 2$

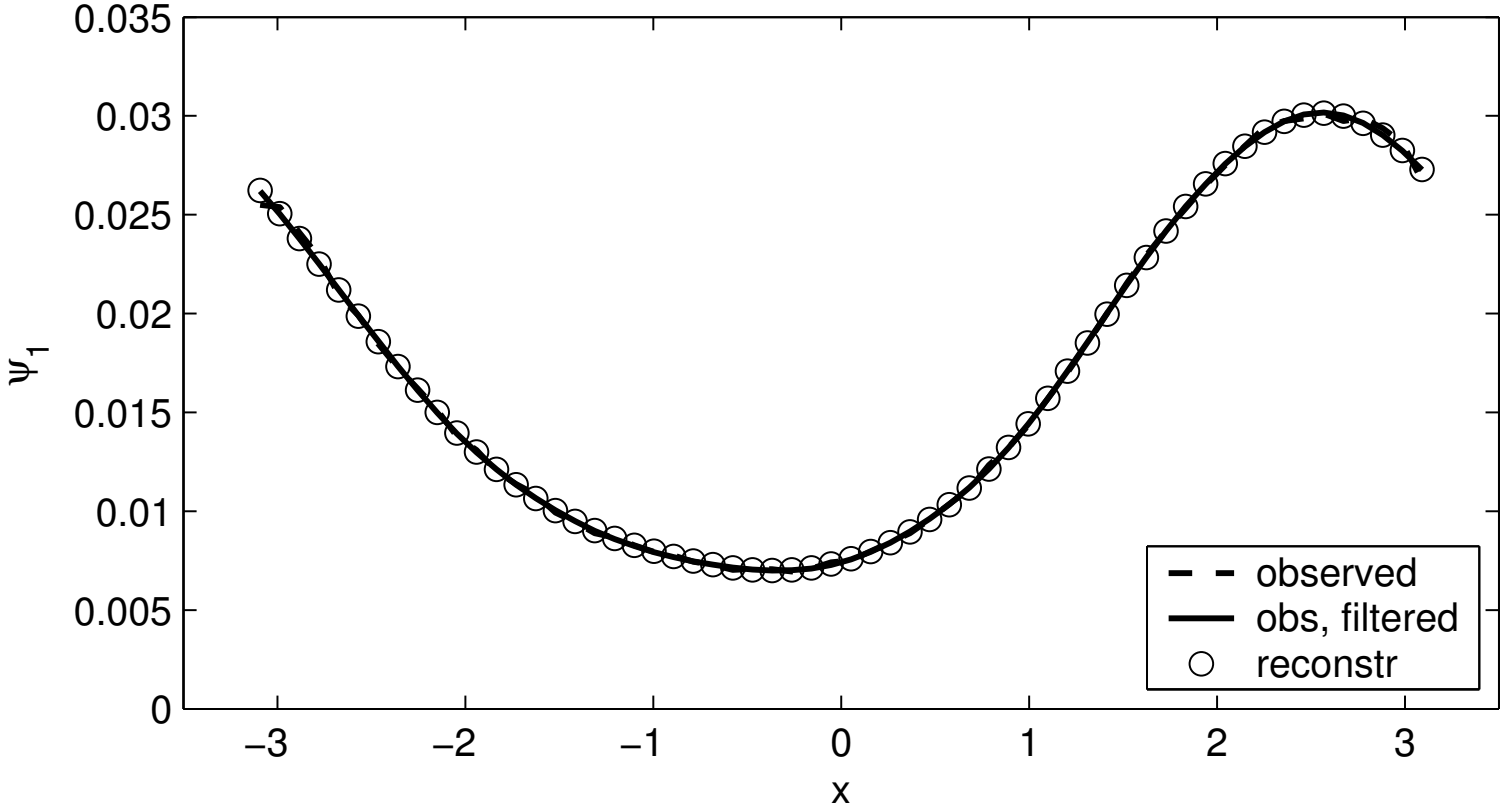
Example: 1-d SDE

Drift $b(x) = 1 + \cos x$, diffusion $a(x) = 1 + \frac{1}{2}\sin x$,
periodic domain $x \in [-\pi, \pi]$.

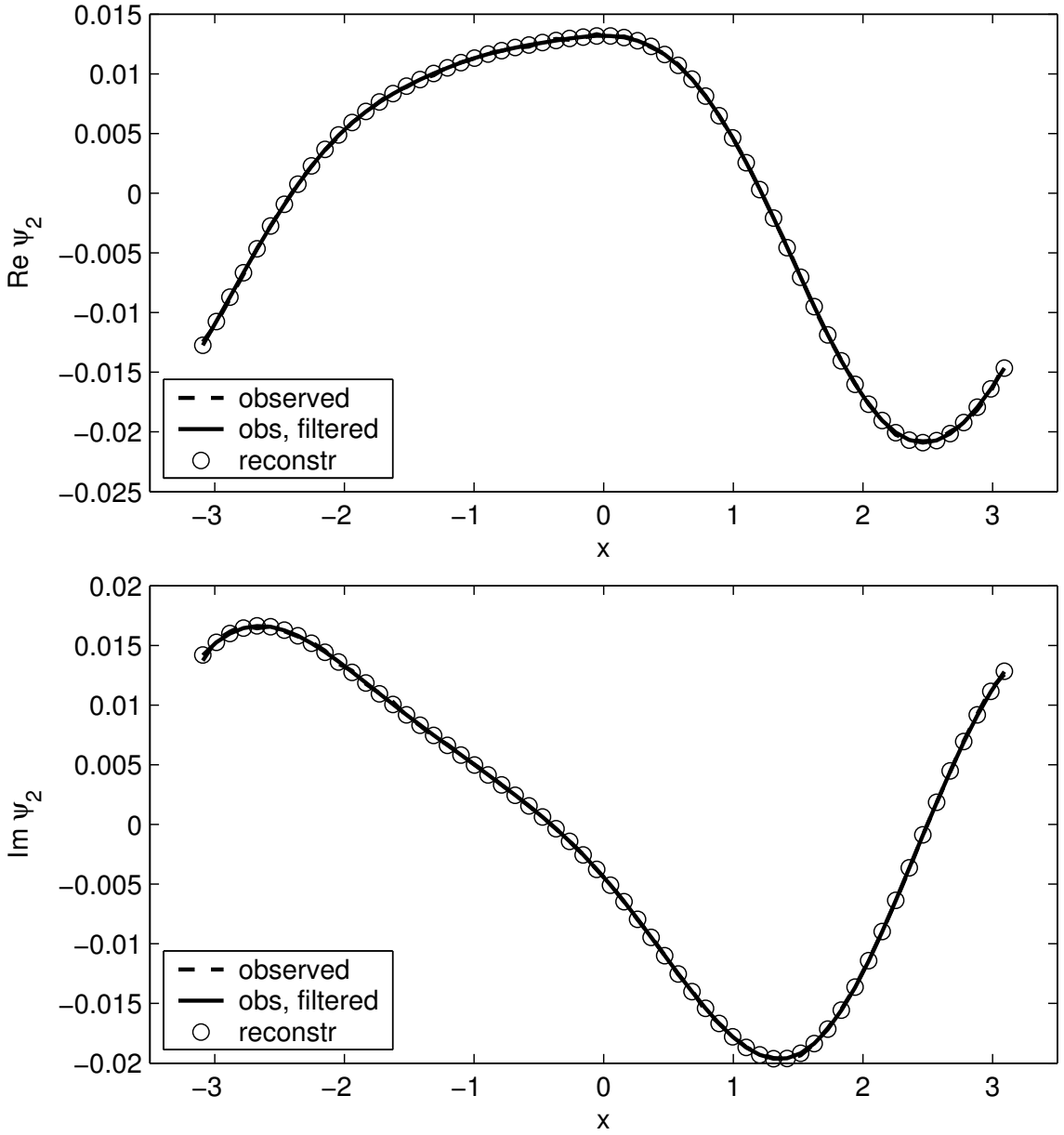




Invariant distribution ψ_1



ψ_2 (Re, Im)


Re ψ2
0.015
0.01
0.005
0
−0.005
−0.01
−0.015
−0.02
−0.025
−3
−2
−1
0
1
2
3
x
observed
obs, filtered
reconstr
Im ψ2
0.02
0.015
0.01
0.005
0
−0.005
−0.01
−0.015
−0.02
−3
−2
−1
0
1
2
3
x
observed
obs, filtered
reconstr

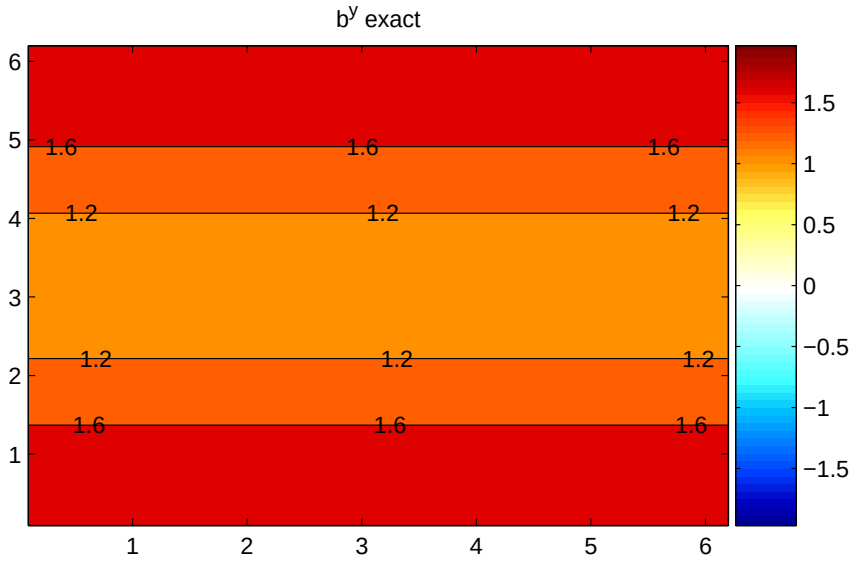
Example: 2-d SDE

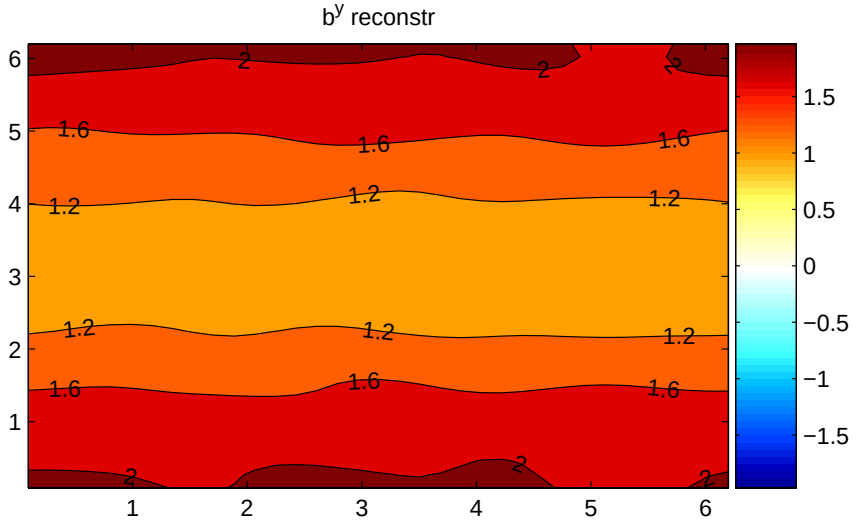
Drift: $b^x = 1, \quad b^y = \frac{3}{2} + \frac{1}{2}\cos y$

Diffusion: $a^{xx} = \left(\frac{3}{4} + \frac{1}{4}\sin y\right)^2, \quad a^{yy} = 1$

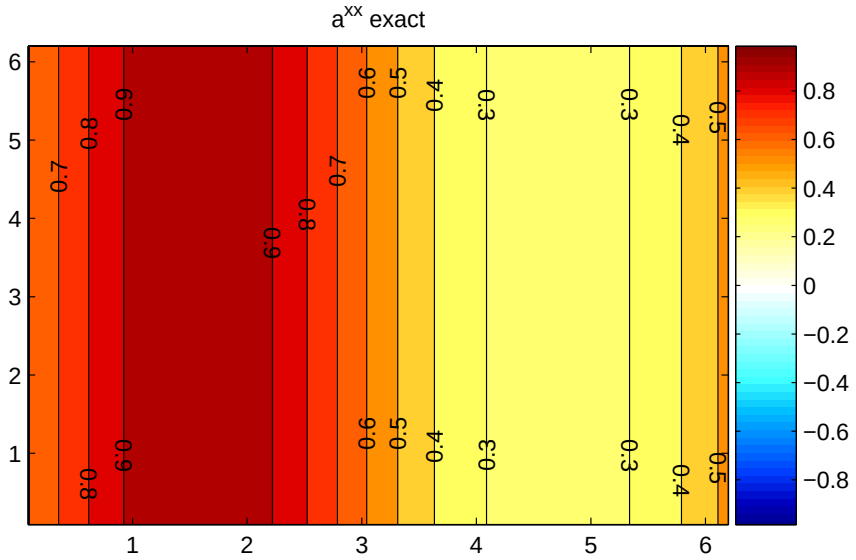
Periodic domain $(x, y) \in [-\pi, \pi] \times [-\pi, \pi]$.

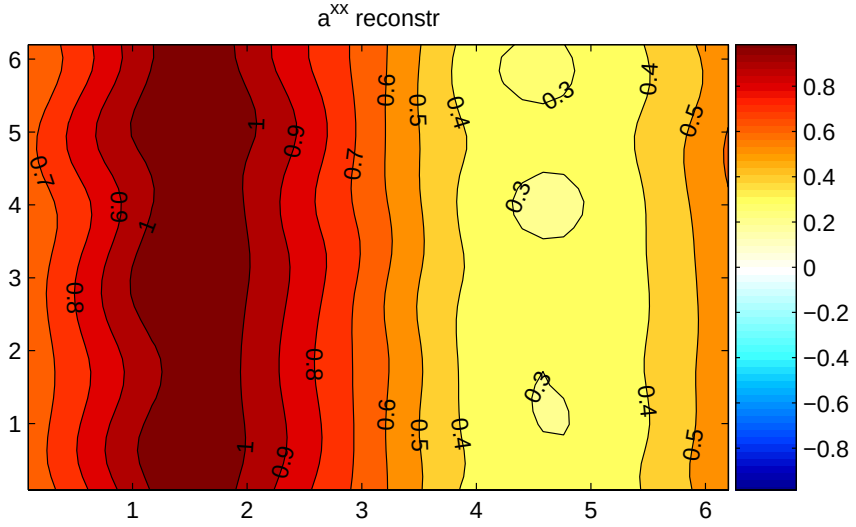
$$b^y = \frac{3}{2} + \frac{1}{2}\cos y$$



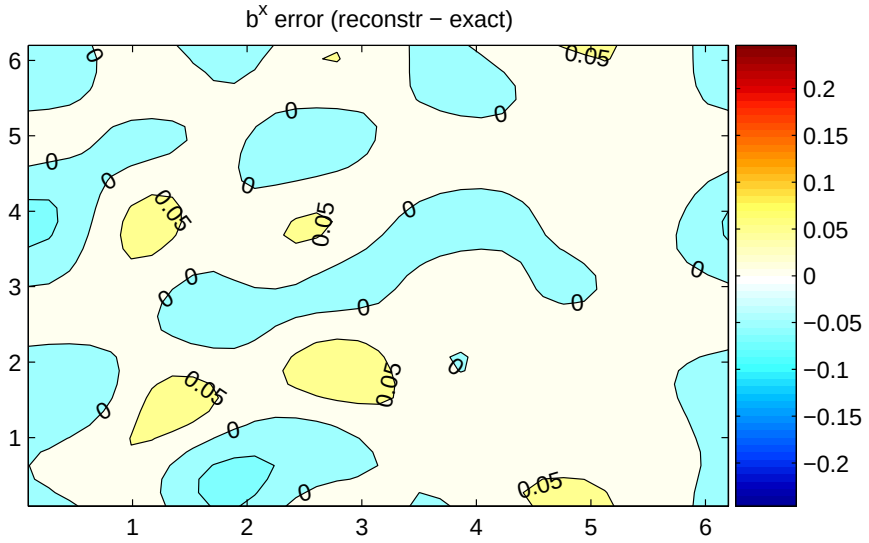


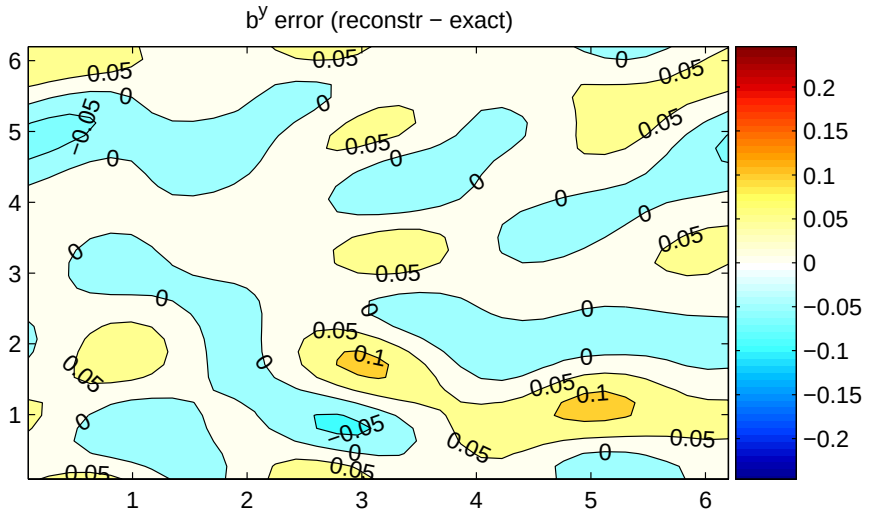
$$a^{xx} = \left(\frac{3}{4} + \frac{1}{4}\sin y\right)^2$$



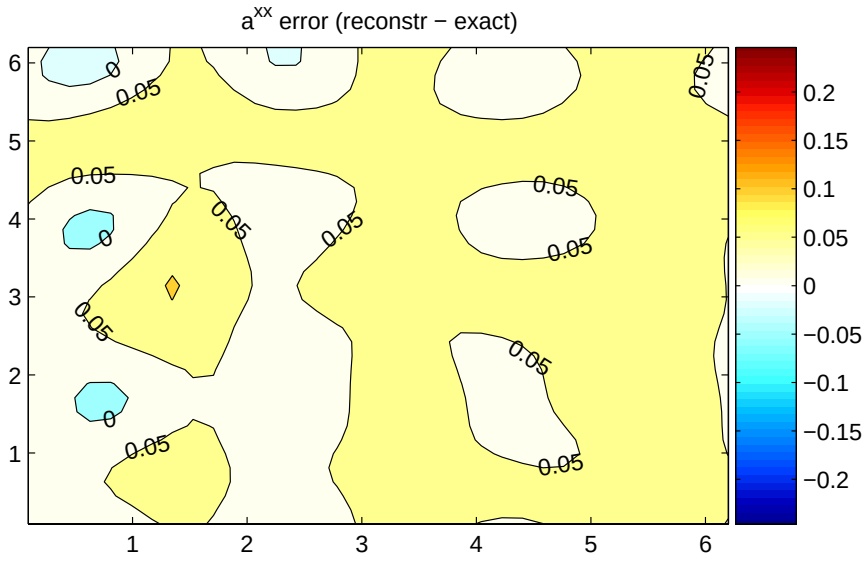


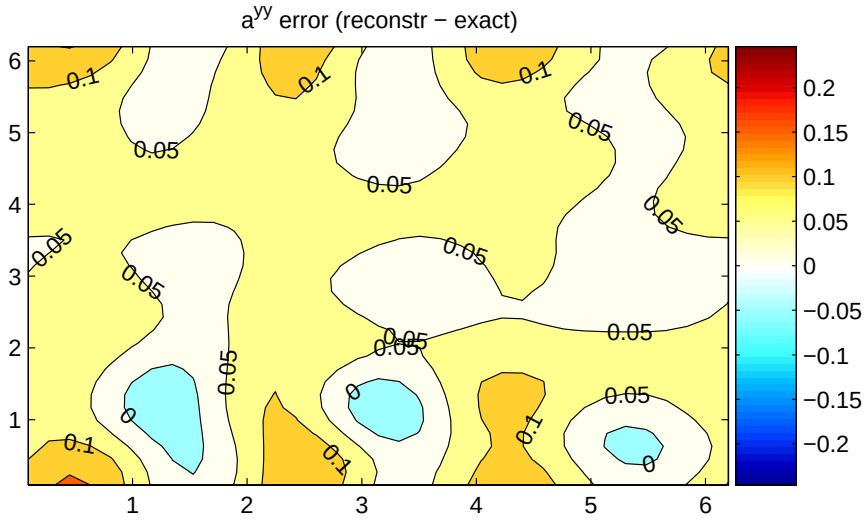
errors: drift





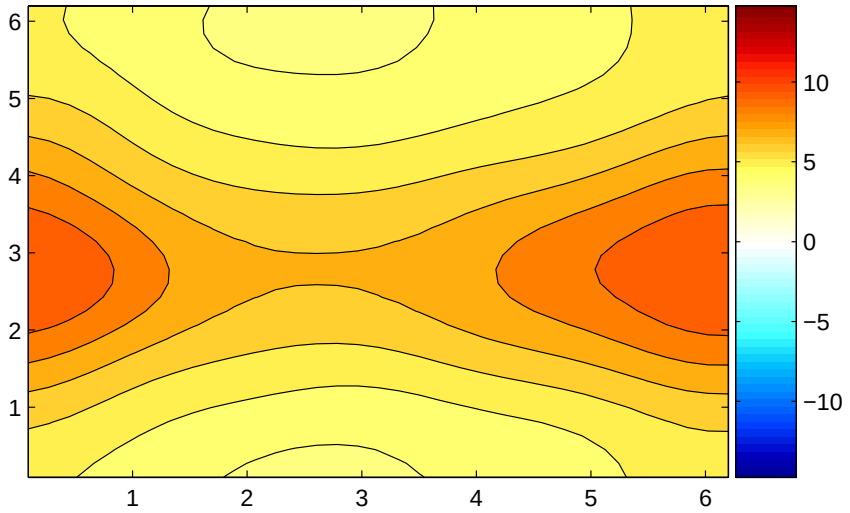
errors: diffusion



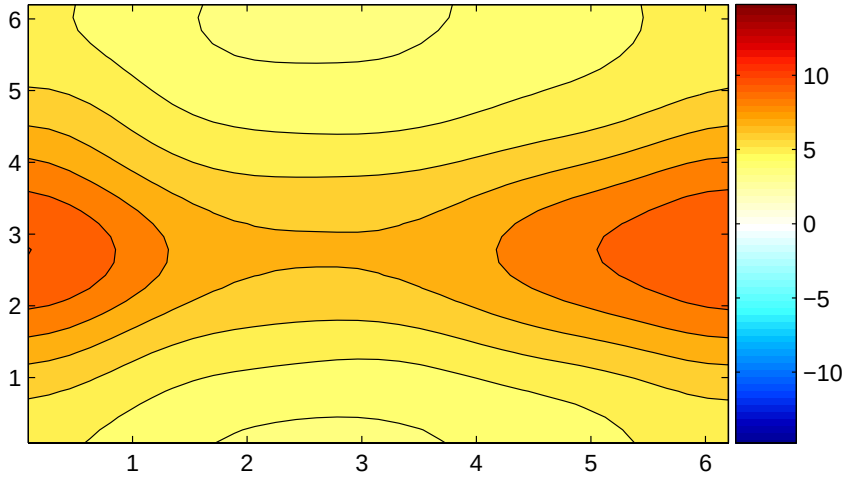


invariant distribution

ψ_1 (x 10), reference



ψ_1 (x 10), reconstructed

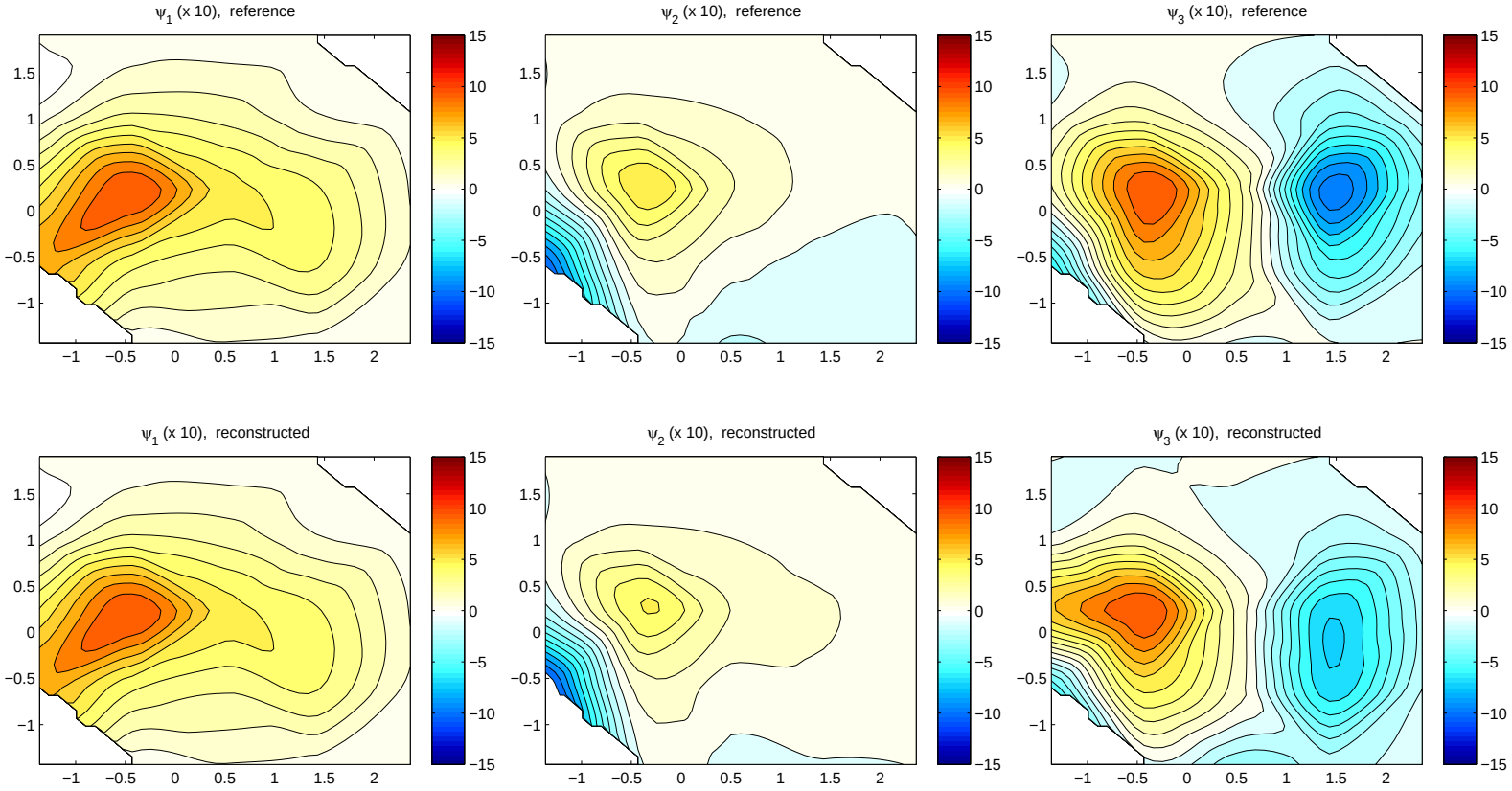


Example: atmospheric timeseries, 2-d

Leading Principle Component timeseries, generated by T21 barotropic model of flow over topography (NH).

Effective model: 23-state continuous-time Markov chain

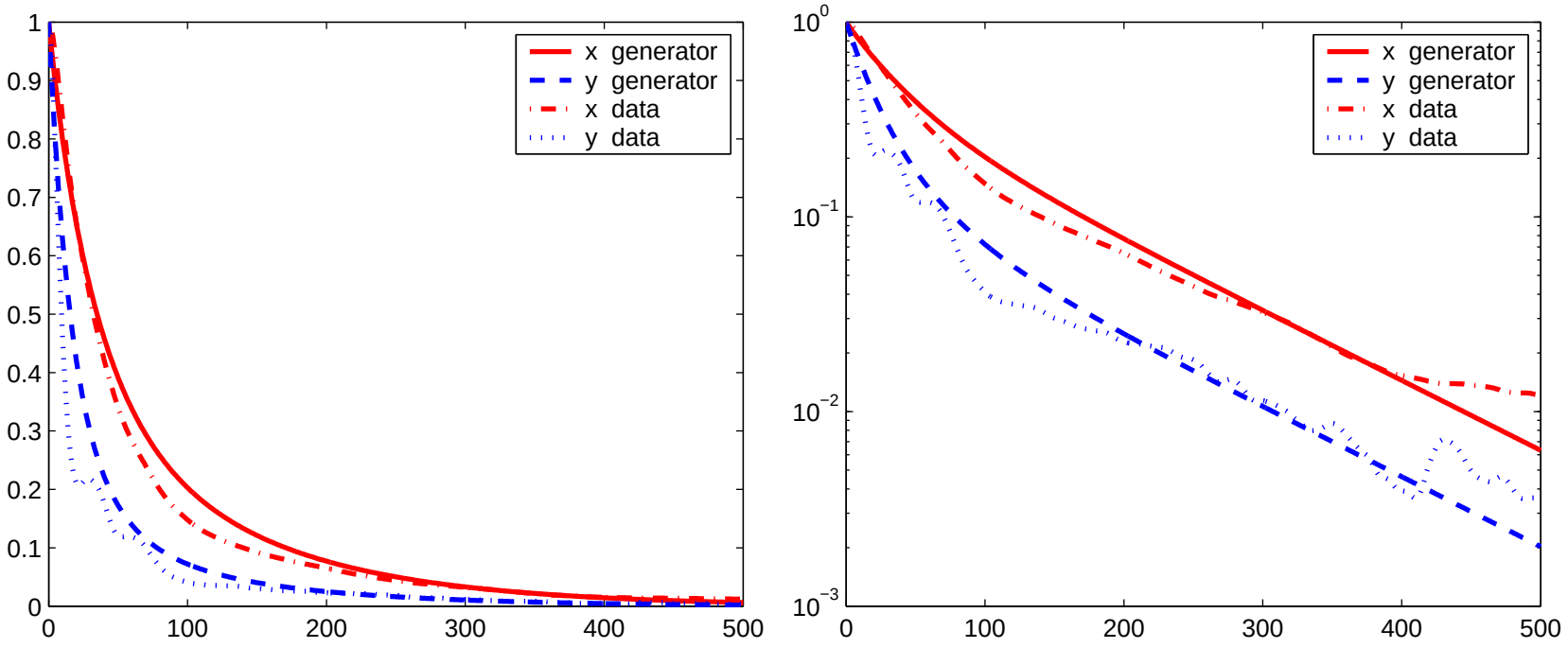
Leading eigenvectors ψ_k:



k=1 k=2 k=3

Autocorrelation functions determined by eigenspectrum:

$$\mathbb{E}(X_{t+\tau}X_t) = \sum_k \mathrm{e}^{\lambda_k t} \int dx\, x\, \psi_k(x) \int dy\, y\, \phi_k(y)\mu(y)$$



Example: a system with two timescales

$$\begin{cases} \dot{x} = b_x(x,y) + \sigma_x(x,y)\dot{W}_1, \\ \dot{y} = \dfrac{1}{\varepsilon} b_y(x,y) + \dfrac{1}{\sqrt{\varepsilon}} \sigma_y(x,y)\dot{W}_2, \qquad \varepsilon \ll 1 \end{cases}$$

Limiting diffusion: as $\varepsilon \to 0$, slow variable x can be approximated by process $\dot{x} = \bar{b}(x) + \bar{\sigma}(x)\dot{W}_1$

with $\bar{b}(x) = \int b_x(x,y) d\mu_x(y), \quad \bar{a}(x) = \int \sigma_x(x,y)\sigma_x^T(x,y) d\mu_x(y)$

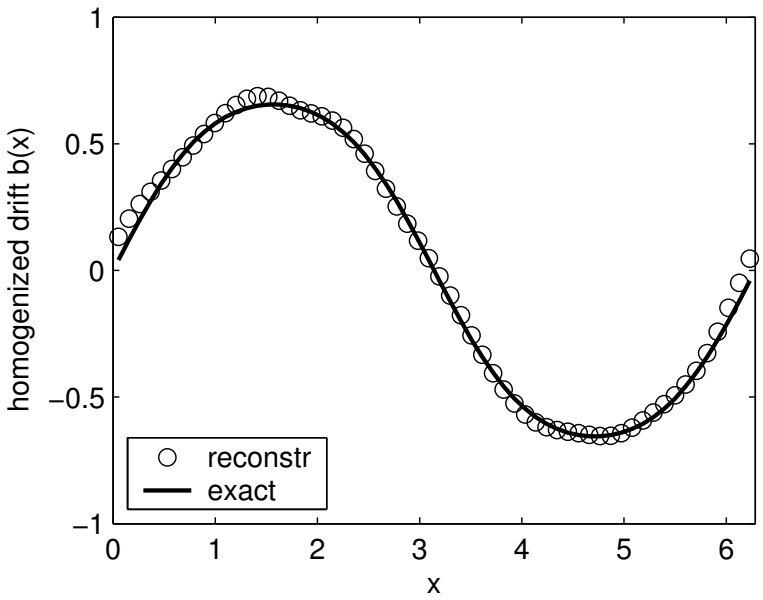
(Khasminskii, Kurtz, Papanicolaou, ...)

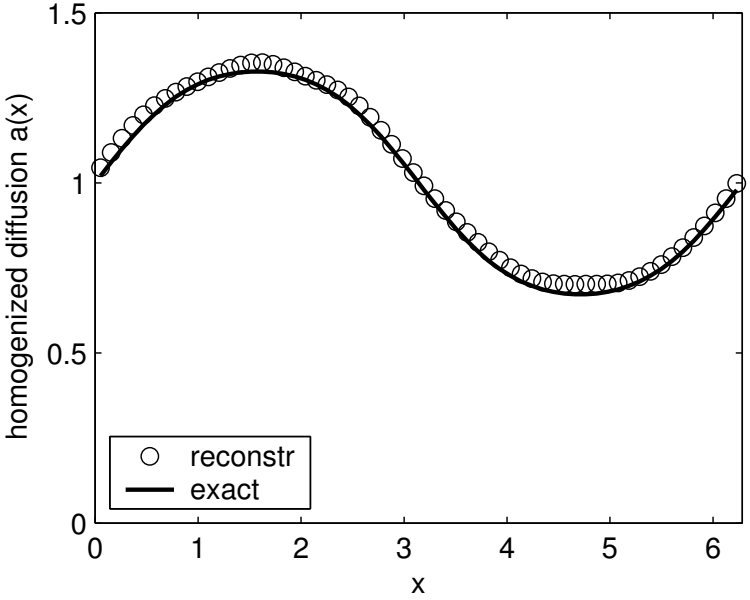
Take $\begin{cases} \dot{x} = \sin y + \sqrt{1 + \frac{1}{2}\sin y}\,\dot{W}_1 \\ \dot{y} = \frac{1}{\varepsilon}(y - \sin x) + \frac{1}{\sqrt{\varepsilon}}\,\dot{W}_2 \end{cases}$ ($\varepsilon = 10^{-3}$, x periodic on $[0, 2\pi]$)

$\rightarrow \bar{b}$ and $\bar{a}$ can be explicitly calculated
(y: OU-process, mean $\sin x$, var 1/2 $\rightarrow \mu_x(y)$ known)

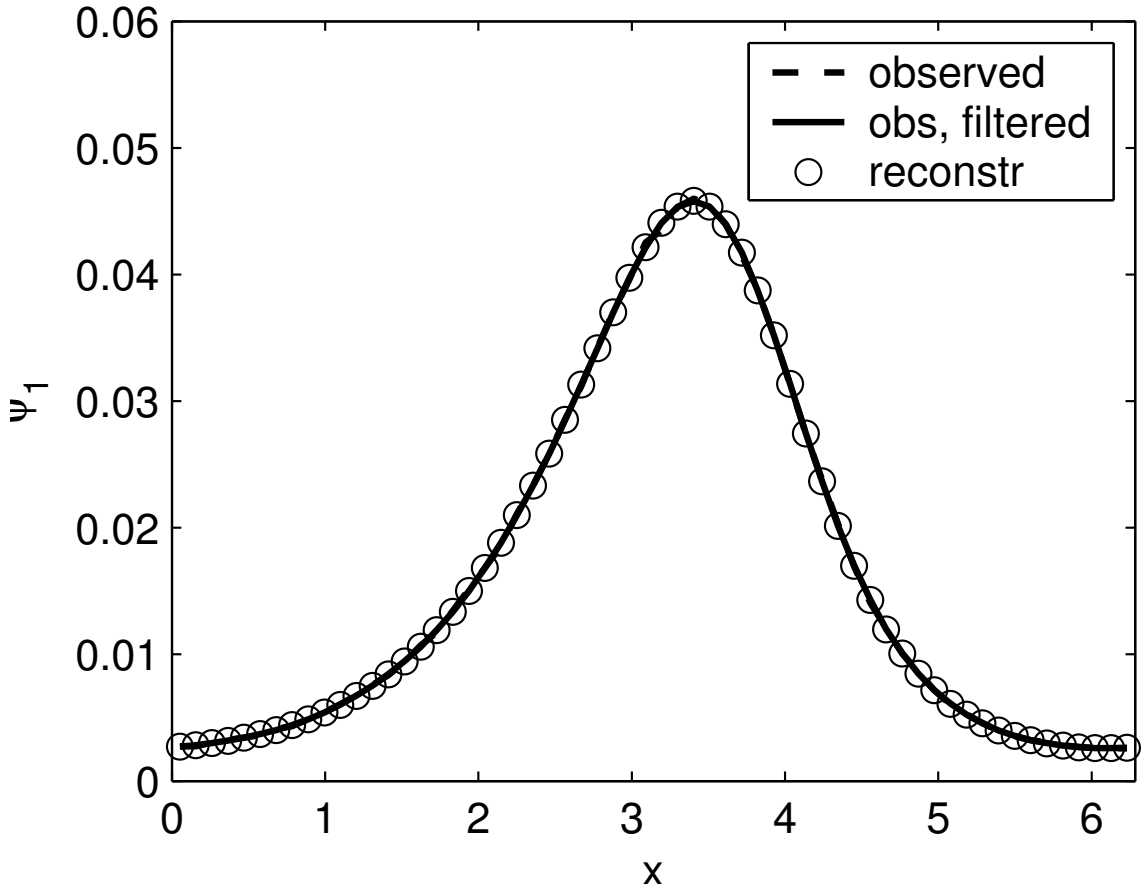
Reconstruction from data: use timeseries for x only, $h = 0.1 \gg \varepsilon$

Homogenised (effective) drift and diffusion
well captured by reconstruction:





Invariant distribution ψ_1 for x:



Stochastic parameterization for L96 model

2-layer L96 model (Lorenz 1996):

$$\begin{cases} \dot{X}_k = X_{k-1}(X_{k+1} - X_{k-2}) - X_k + F_x + B_k \\ \dot{Y}_{j,k} = \frac{1}{\varepsilon}\Big(Y_{j+1,k}(Y_{j-1,k} - Y_{j+2,k}) - Y_{j,k} + h_y X_k\Big) \end{cases}$$

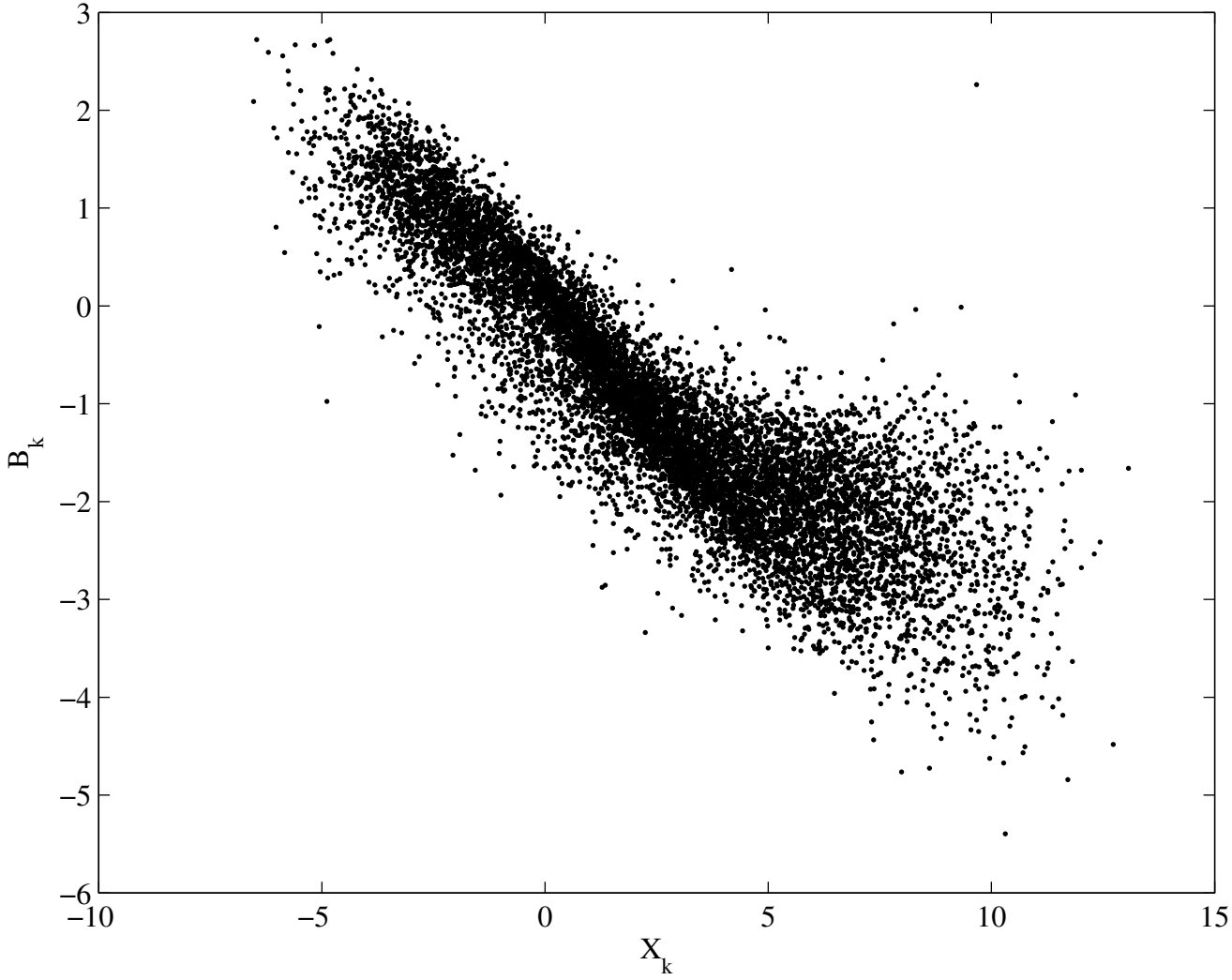
with $B_k = \dfrac{h_x}{J}\sum_{j=1}^{J} Y_{j,k}$ $\qquad (k = 1, \ldots, K;\ j = 1, \ldots, J)$

Parameter settings: $\varepsilon = 0.1$

$K = 36, J = 10$

$F_x = 10, h_x = -1, h_y = 1$

B_k versus X_k



Find model for the X_k alone

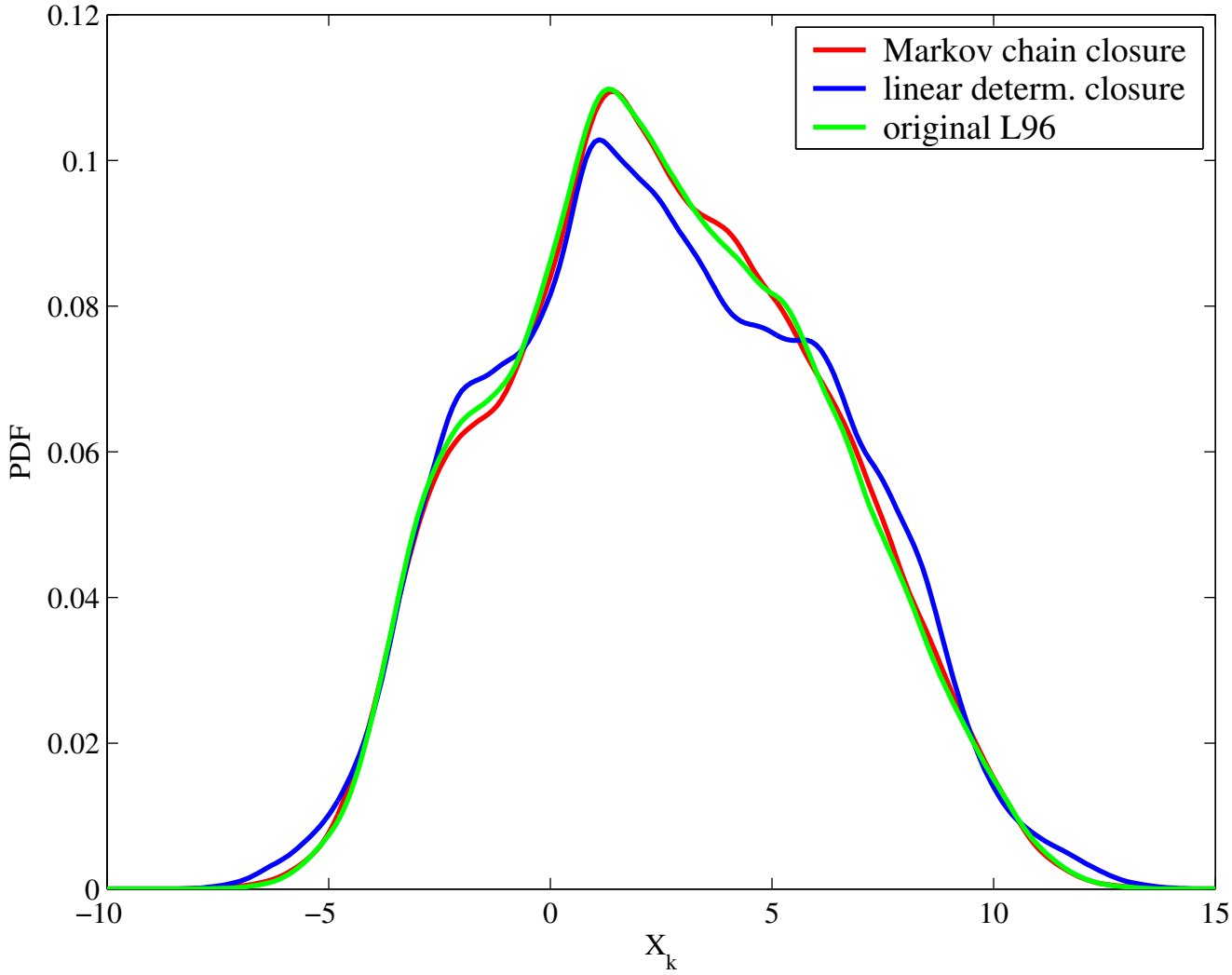
Some parameterizations / closures:

- Truncation: $B_k = 0$

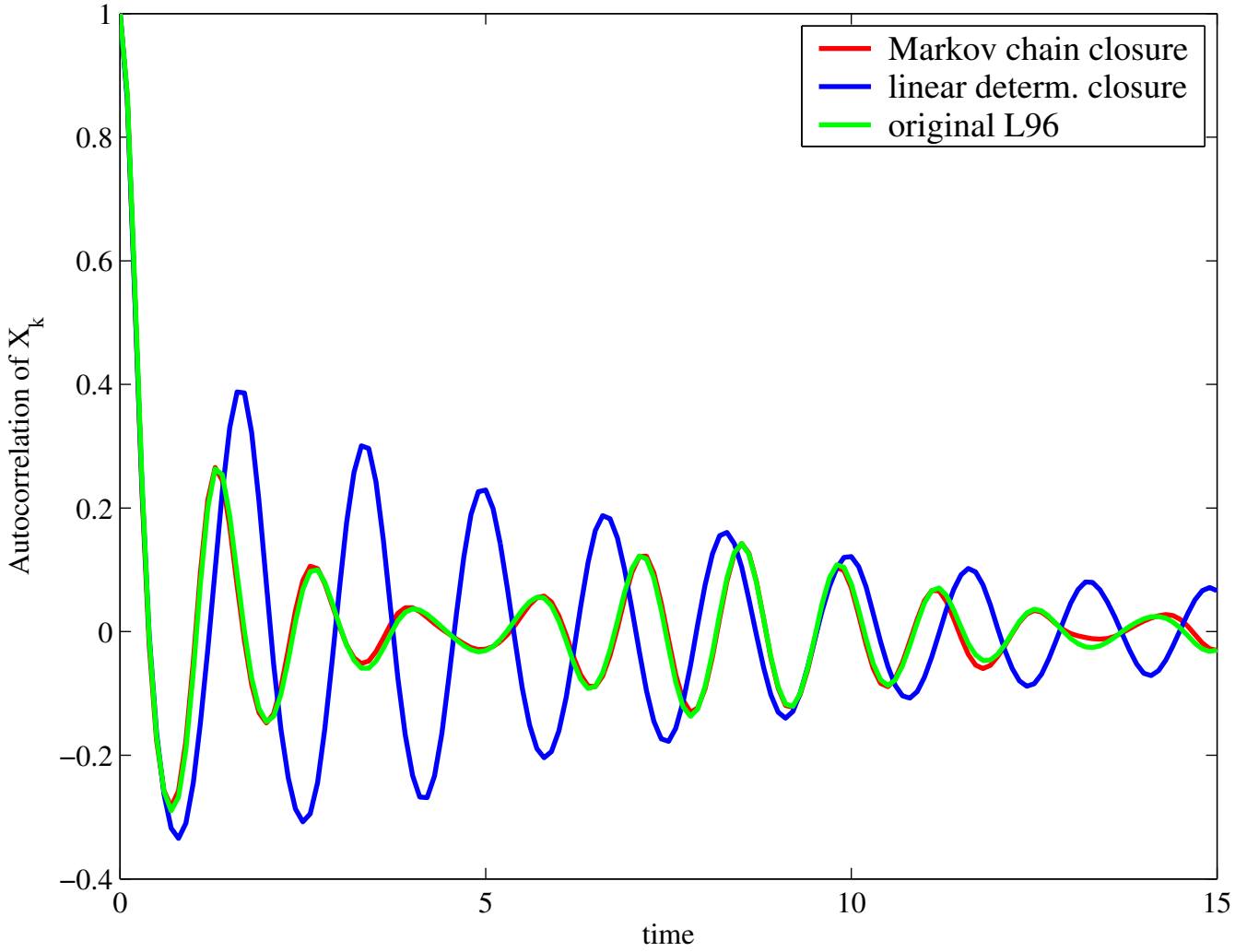
- Deterministic: fit curve through X_k, B_k scatterplot

- Stochastic: construct continuous-time Markov chains for B_k on different intervals of X_k ($|X_k - 1| < 0.5$, $|X_k - 2| < 0.5$ etcetera). Dynamics of the B_k: MCMC; switch between chains if X_k changes.

Probability density functions



Autocorrelation functions



Conclusion/outlook

- Reconstruction by matching eigenmodes works well
- Avoids non-Markov effects at short timescales
- Focus on effective dynamics on long timescales

- Different numerical implementations, e.g. parametrical
- Efficiency using data
- Higher dimensions
- Parameterizations/subgrid-scale modeling
- Effective dynamics of LFV modes (NAO etc.)