Entropy-Based Ensemble Prediction Schemes

Sangil Kim
Program in Applied Mathematics
University of Arizona

and

Gregory Eyink
Applied Mathematics & Statistics
The Johns Hopkins University

Supported by NSF Grant # DMS-0113649.

The Filtering Problem

A *state vector* $\mathbf{x}_t \in \mathbb{R}^p$ evolves via

$$\mathbf{x}_{t+1} = \mathbf{F}_t(\mathbf{x}_t, \boldsymbol{\eta}_t), \quad t = 0, 1, ..., T \qquad (1)$$

$\mathbf{F}_t : \mathbb{R}^p \times \mathbb{R}^r \longrightarrow \mathbb{R}^p$ and $\boldsymbol{\eta}_t \in \mathbb{R}^r$ is a random noise vector. Initial conditions $\mathbf{x}_0$ are chosen randomly from $P_0(\mathbf{x})$.

Measurements are taken

$$\mathbf{y}_t = \mathbf{h}_t(\mathbf{x}_t) + \boldsymbol{\epsilon}_t, \qquad (2)$$

for a subset of t, where $\mathbf{h}_t : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ and $\boldsymbol{\epsilon}_t \in \mathbb{R}^q$ are random observation errors. Generally,

$$1 \lll q \lll p.$$

We assume that $\boldsymbol{\epsilon}_t$ is an $N(\mathbf{0}, \mathbf{R}_t)$ random q-vector, i.e. normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_t$, and that $\mathbf{h}_t(\mathbf{x}) = \boldsymbol{\mathcal{H}}_t\mathbf{x} + \mathbf{d}_t$ is affine, for a q-vector $\mathbf{d}_t$ and $q \times p$ matrix $\boldsymbol{\mathcal{H}}_t$.

Problems in Geophysical Estimation

(1) Dynamics are nonlinear and statistics may be highly non-Gaussian

(2) States of very low a priori probability before measurements can become very probable afterward

(3) State spaces of the dynamics are often very high-dimensional and only small ensembles of solutions may be generated.

An example can help to illustrate some of these difficulties....

2D Thermohaline Convection

McWilliams & Thual (1991) considered, for $0 \leq z \leq d, -\ell < y < \ell$, the equations

$$\partial_t \nabla^2 \psi + J(\psi, \nabla^2 \psi) = g(\alpha_T \partial_y T - \alpha_S \partial_y S) + \nu \nabla^4 \psi$$

$$\partial_t T + J(\psi, T) = \kappa_T \nabla^2 T$$

$$\partial_t S + J(\psi, S) = \kappa_S \nabla^2 S$$

with free-slip b.c. for stream function

$$\psi = 0, \quad \partial_n^2 \psi = 0$$

and boundary conditions

$$T(y, d) = \Delta T \cdot \theta(y), \quad \partial_z S(y, d) = \Delta S \cdot F(y)/d$$

$$\partial_z T(y, 0) = 0, \quad \partial_z S(y, 0) = 0$$

$$\partial_y T(\pm\ell, z) = 0, \quad \partial_y S(\pm\ell, z) = 0$$

for temperature and salinity.

Surface Forcing

McWilliams & Thual took

$$\theta(y) = F(y) = \cos y$$

Instead, we'll consider

$$\theta(y) = \cos y$$

but

$$F(y,t) = \bar{F}(y) + \widetilde{F}(y,t)$$

$\bar{F}(y)$ is the *systematic salinity flux*, specified later, and $\widetilde{F}(y,t)$ is *random salinity flux*, taken to be zero-mean Gaussian white-noise with covariance

$$\langle \widetilde{F}(y,t)\widetilde{F}(y',t')\rangle = \Sigma_0^2 \cdot \delta(y-y')\delta(t-t')$$

Small Aspect-Ratio Limit

Cessi & Young (1992) considered

$$\epsilon \equiv \frac{\pi d}{\ell} \ll 1.$$

Nondimensionalize as

$$(y, z) = d\left(\frac{\widehat{y}}{\epsilon}, \widehat{z}\right), \quad t = \frac{d^2}{\kappa_T}\widehat{t}, \quad \psi = \frac{\kappa_T}{\epsilon}\widehat{\psi}$$

$$T = \frac{\nu\kappa_T}{g\alpha_T d^3\epsilon^2}\widehat{T}, \quad S = \frac{\nu\kappa_T}{g\alpha_S d^3\epsilon^2}\widehat{S},$$

In domain $0 \leq z \leq 1, -\pi < y < \pi$,

$$P^{-1}[\partial_t\zeta + J(\psi, \zeta)] = \partial_y T - \partial_y S + (\partial_z^2 + \epsilon^2\partial_y^2)\zeta$$

$$\partial_t T + J(\psi, T) = (\partial_z^2 + \epsilon^2\partial_y^2)T$$

$$L^{-1}[\partial_t S + J(\psi, S)] = (\partial_z^2 + \epsilon^2\partial_y^2)S$$

with zonal vorticity

$$\zeta = (\partial_z^2 + \epsilon^2\partial_y^2)\psi$$

and Prandtl and Lewis numbers

$$P = \nu/\kappa_T, \quad L = \kappa_S/\kappa_T.$$

The surface b.c. for T, S are now

$$T(y,1) = a\theta(y), \quad \partial_z S(y,1) = bF(y)$$

with thermal and saline Rayleigh numbers

$$a = \frac{g\alpha_T \Delta T d^3 \epsilon^2}{\nu\kappa_T} \quad b = \frac{g\alpha_S \Delta S d^3 \epsilon^2}{\nu\kappa_T}.$$

Another dimensionless group also appears

$$c = \frac{g\alpha_S \Sigma_0 (\epsilon d)^{5/2}}{\nu\kappa_T^{1/2}},$$

for the magnitude of the stochastic flux term.

For a nontrivial limit, one must take

$$a = \epsilon a_1, \quad b = \epsilon^3 b_3, \quad c = \epsilon^2 c_2$$

and expand

$$(\psi, T, S) = \epsilon(\psi_1, T_1, S_1) + \epsilon^2(\psi_2, T_2, S_2) + \cdots$$

At third-order, one obtains a solvability condition for salinity $\sigma(y,\tau) = a_1^{-1} S_1$, $\tau = \epsilon^2 Lt$:

$$\partial_\tau \sigma - \mu^2 \partial_y [\partial_y \sigma (\partial_y \sigma - \partial_y \theta)^2] = rF + \partial_y^2 \sigma - \gamma^2 \partial_y^4 \sigma.$$

Amplitude Equation

With *meridional thermal and salinity gradients*

$$\eta(y) \equiv \partial_y \theta(y), \quad \chi(y,\tau) \equiv \partial_y \sigma(y,\tau),$$

the solvability equation becomes

$$\boxed{\begin{aligned}\partial_\tau \chi = \partial_y^2[\mu^2 \chi(\chi-\eta)^2 - r\bar{f}(y) + \chi - \gamma^2 \partial_y^2 \chi] \\ + \partial_y \tilde{f}(y,\tau)\end{aligned} \quad (1)}$$

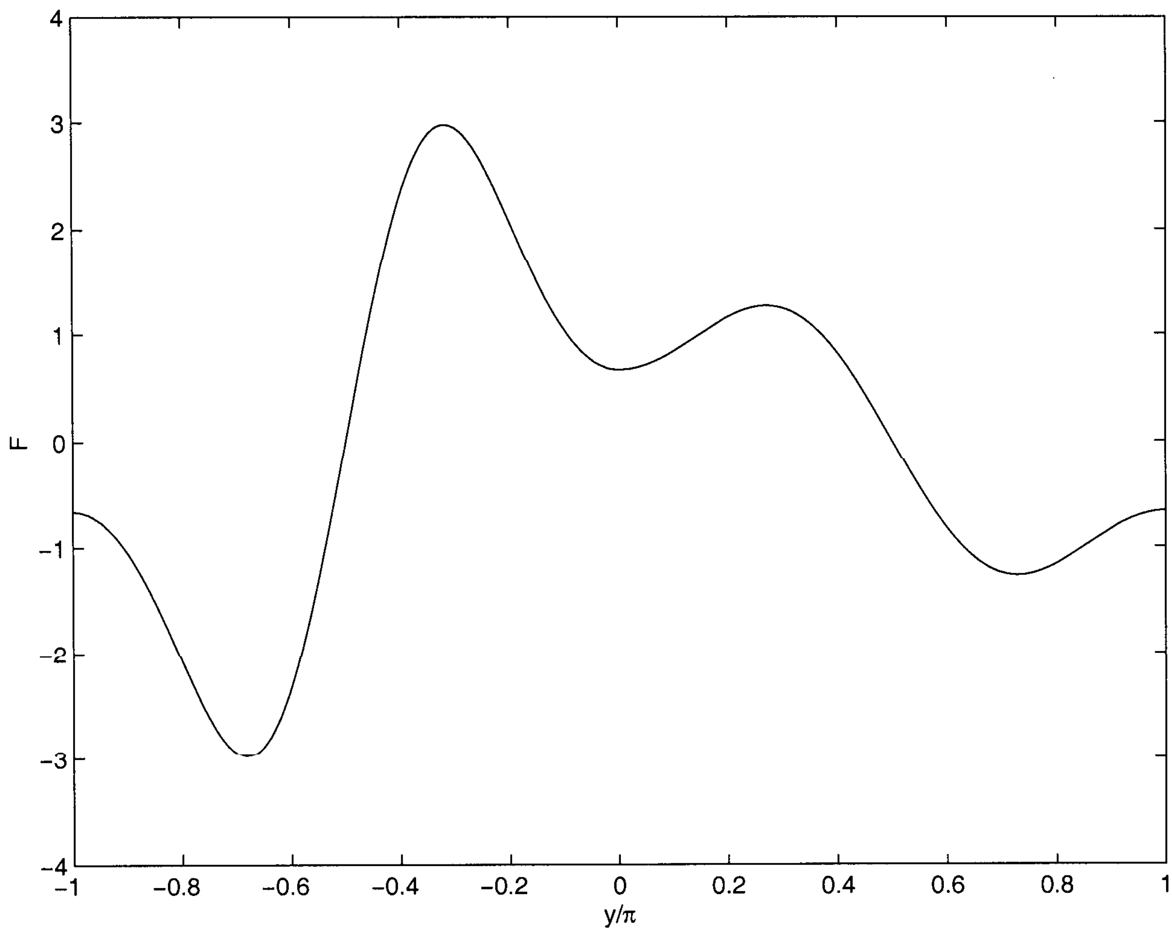
with

$$\bar{f}(y) = -\int_{-\pi}^{y} \bar{F}(\bar{y})d\bar{y}$$

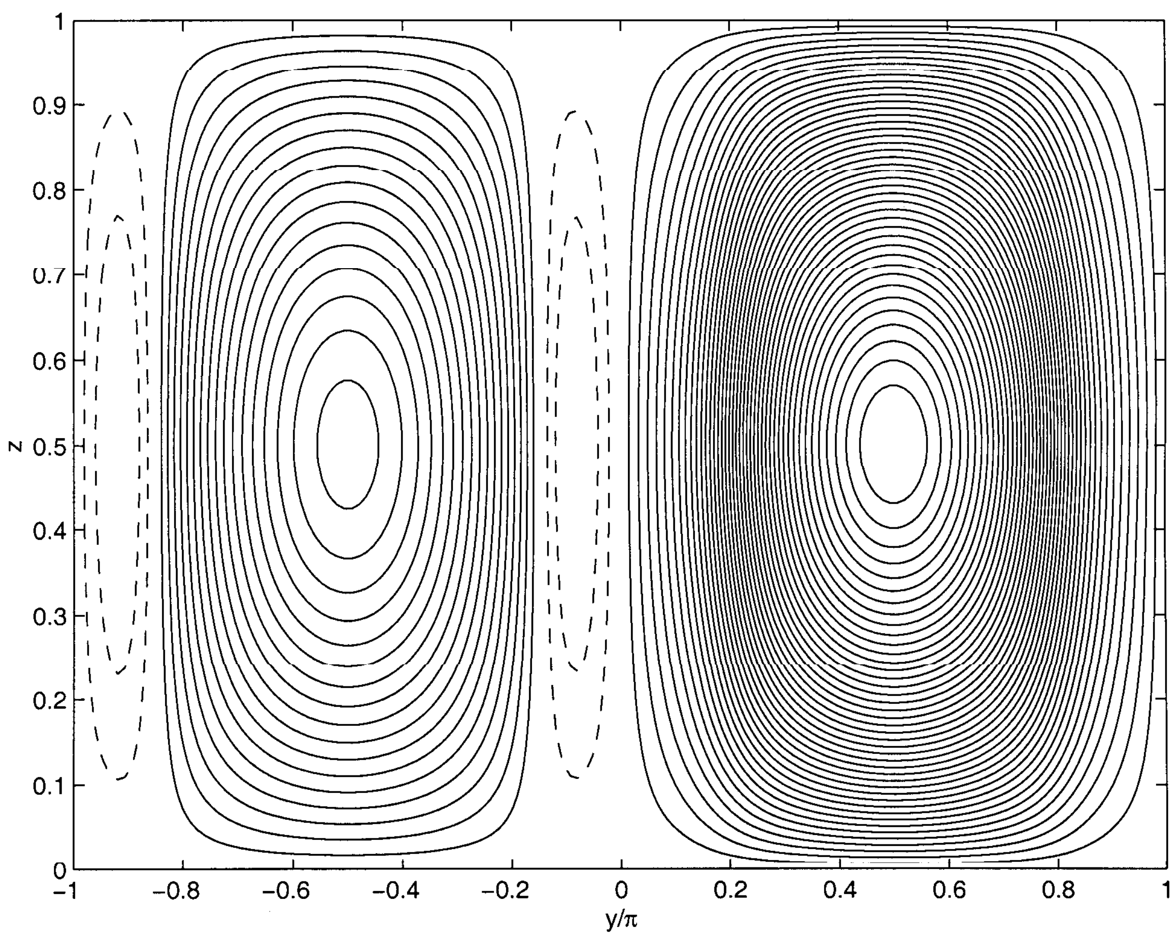
and

$$\langle \tilde{f}(y,\tau)\tilde{f}(y',\tau')\rangle = \sigma_0^2 \cdot \delta(y-y')\delta(\tau-\tau')$$

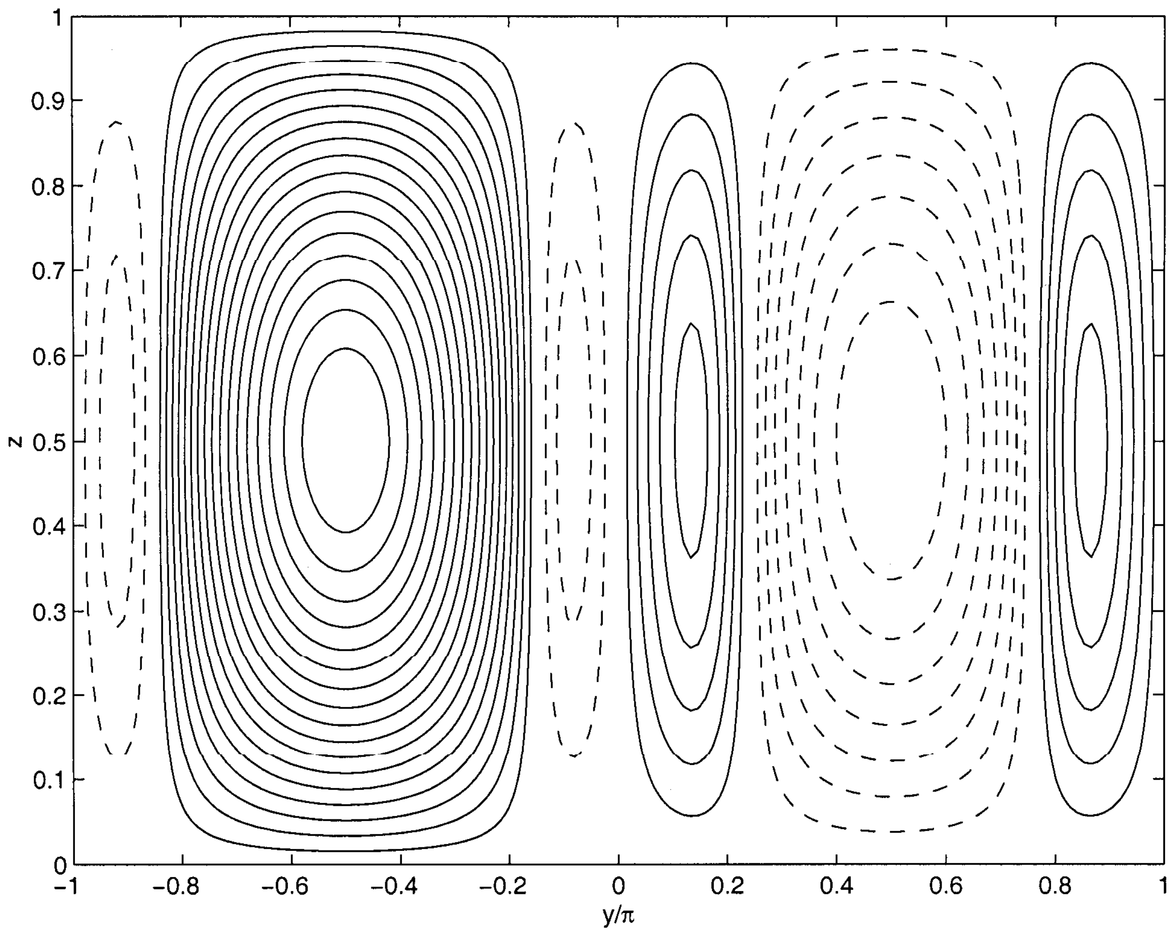
with $\sigma_0 = L^{1/2}c_2/a_1$.



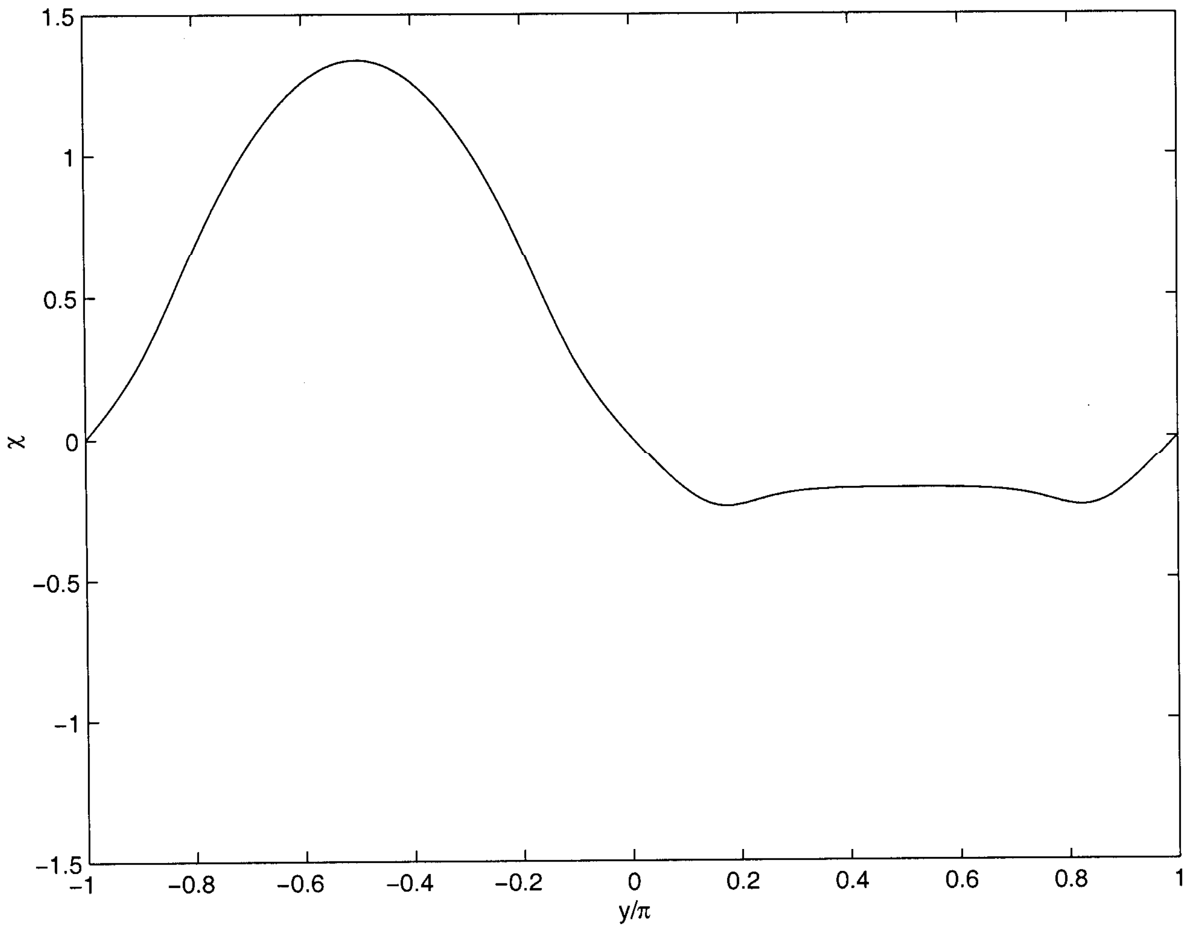
4
3
2
1
0
-1
-2
-3
-4
F
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
y/π

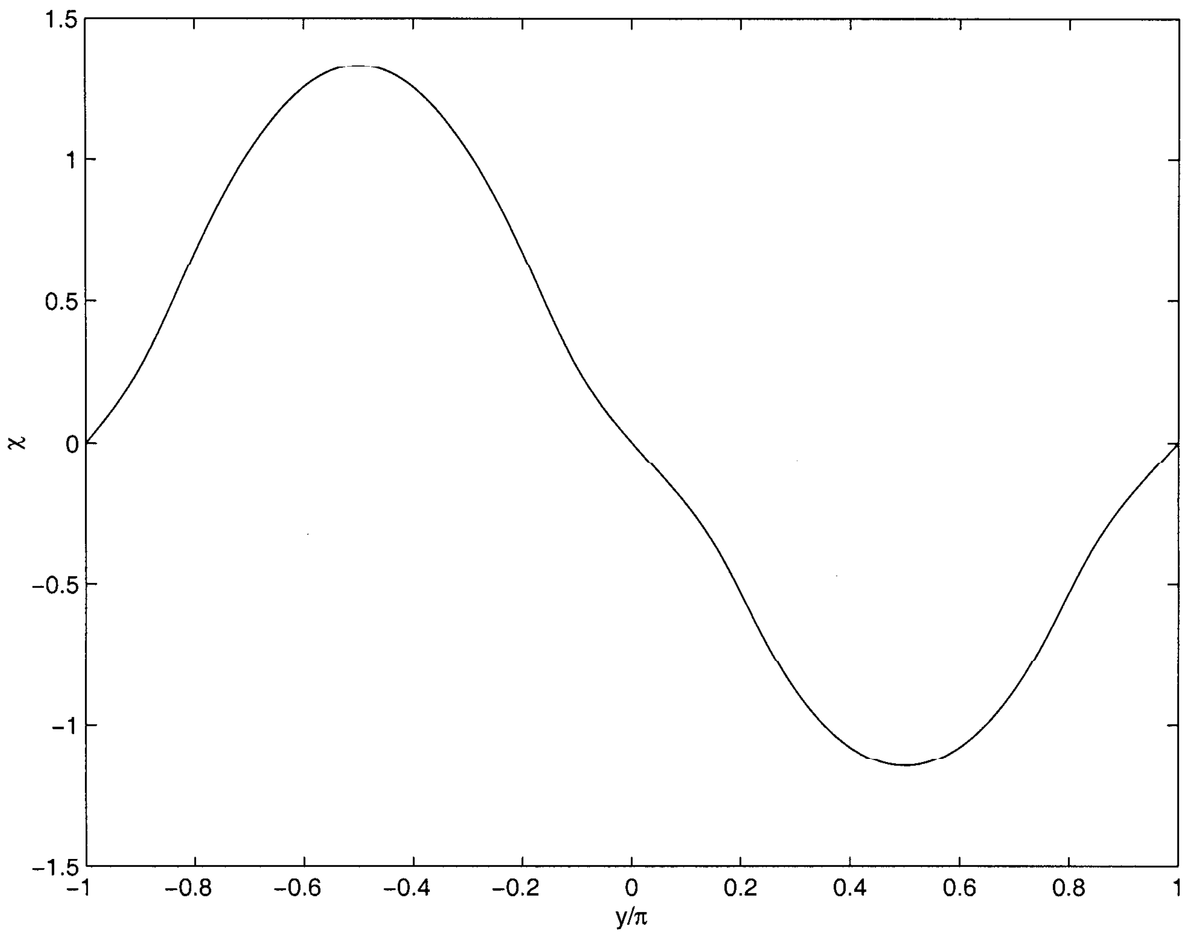


1
0.9
0.8
0.7
0.6
z 0.5
0.4
0.3
0.2
0.1
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
y/π



z
y/π


1.5
1
0.5
0
−0.5
−1
−1.5
χ
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
y/π



1.5
1
0.5
0
−0.5
−1
−1.5
χ
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
y/π

Order Parameter

The *salinity field* is

$$\sigma(y,\tau) = \sigma(0) + \int_0^y d\bar{y}\ \chi(\bar{y},\tau),$$

where the value of the equatorial salinity $\sigma(0)$ may be freely defined, e.g. $\sigma(0) = 0$.

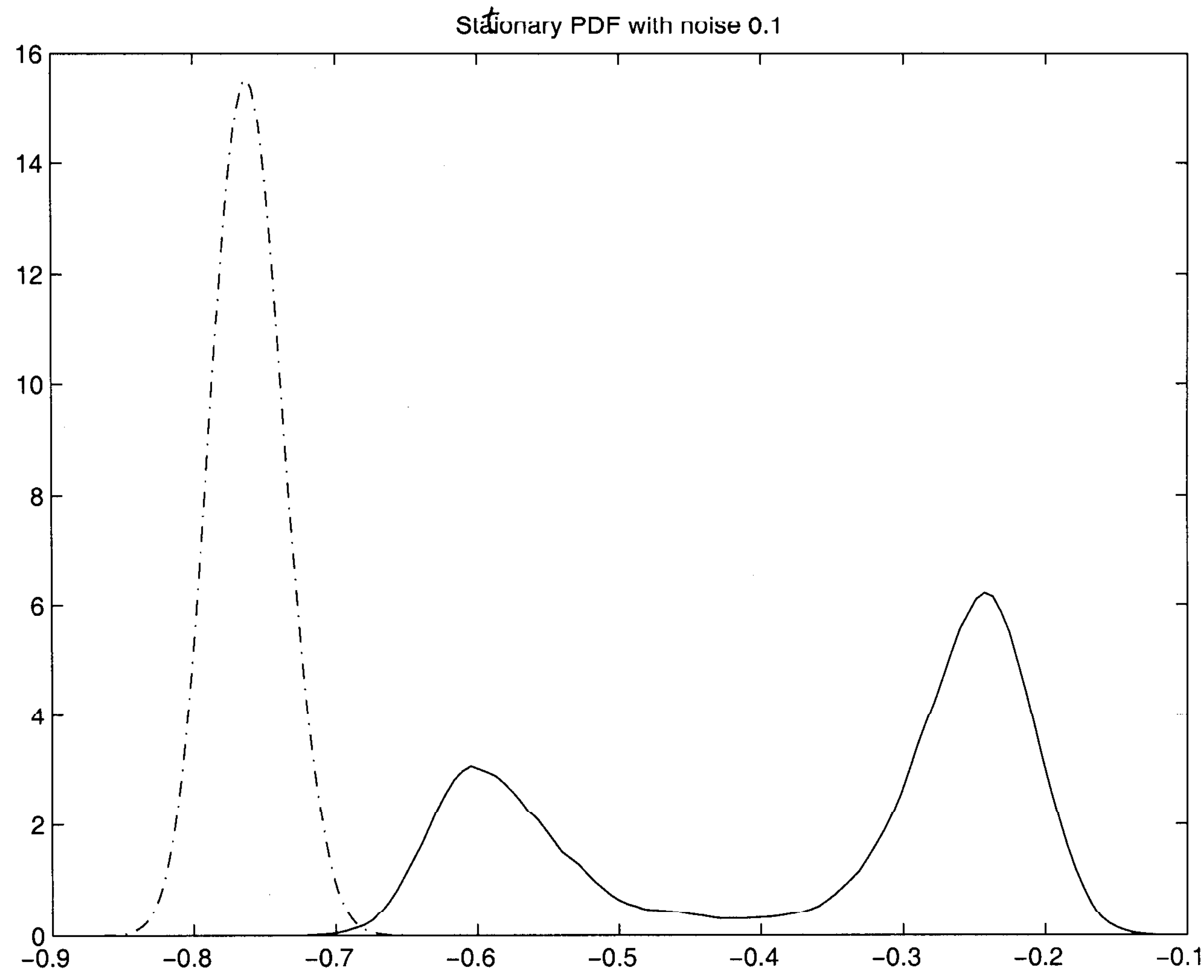
The salinity of the north polar water

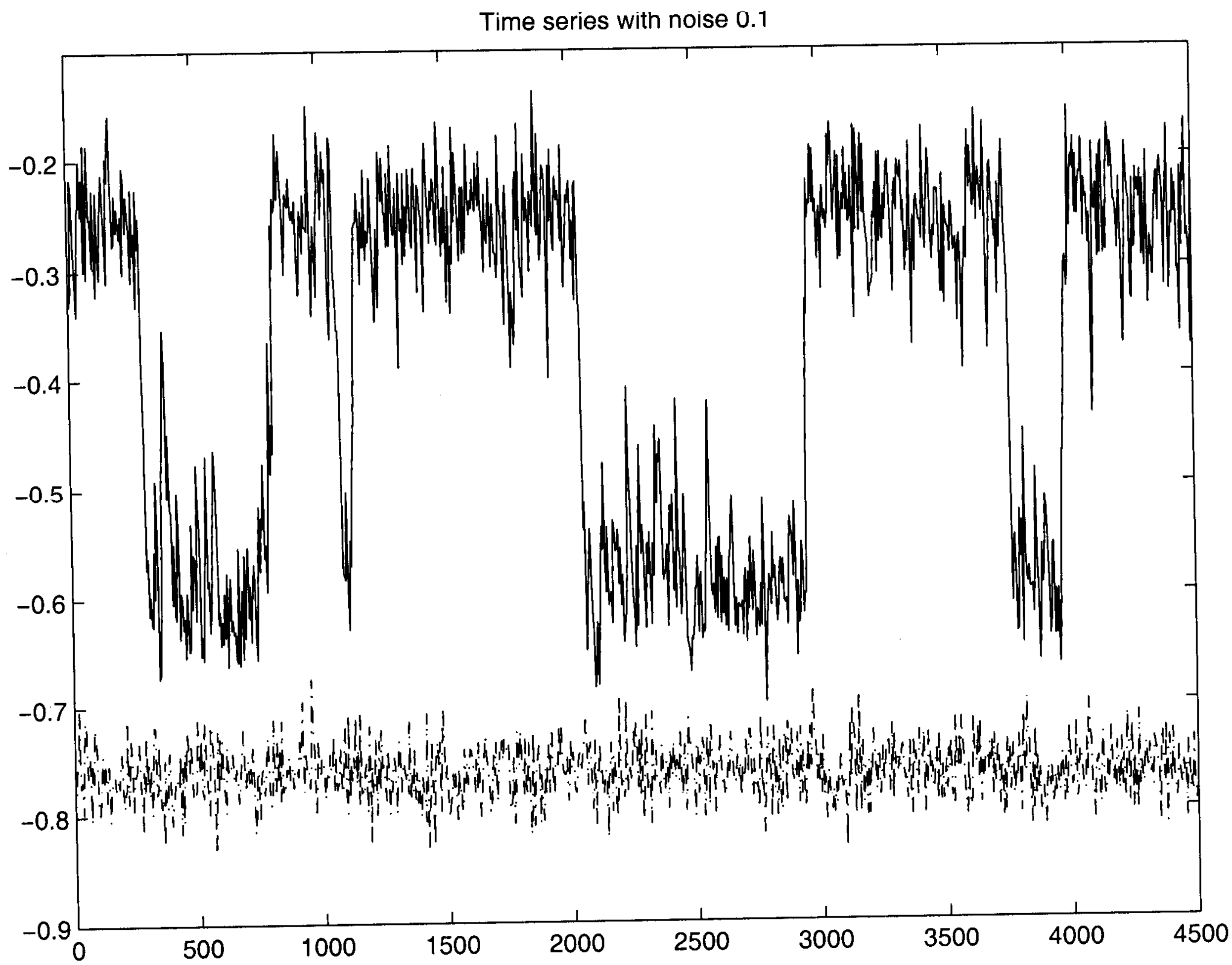
$$\sigma_N(\tau) = \sigma(\pi,\tau)$$

acts as an "order parameter" to distinguish in which of the stable equilibria the system resides at time τ.

We also consider salinity of the south polar water

$$\sigma_S(\tau) = \sigma(-\pi,\tau)$$


Stationary PDF with noise 0.1
16
14
12
10
8
6
4
2
0
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1



Time series with noise 0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
0
500
1000
1500
2000
2500
3000
3500
4000
4500

Model Measurements

At each observation time τ, we shall measure salinity field $\sigma(y,\tau)$ at various latitudes y_k,

$$\boxed{\gamma_k = \sigma(y_k,\tau) + \epsilon_k \quad (2)}$$

where ϵ_k is an $N(0,R_k)$ random measurement error for $k = 1,...,q$.

In previous notations, this corresponds to the affine measurement function $\mathbf{h}[\chi,\tau]$ given by

$$h_k[\chi;\tau] = \int_0^{y_k} d\bar{y}\ \chi(\bar{y},\tau), \quad k = 1,...,q.$$

and error covariance matrix

$$\mathbf{R}(\tau) = \begin{bmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & R_q \end{bmatrix}$$

In our example we take $q = 2$, with

$$y_1 = \pi/2 \quad \text{or } 45^\circ \text{ N}$$

$$y_2 = \pi \quad \text{or } 90^\circ \text{ N}$$

and $R_1 = R_2 = 10^{-2}$, corresponding to 10% accuracy in the measurements.

Remark: We have also performed experiments with measurements on many other quantities, such as:

Temperature change:

$$\Delta\theta(y, z; \tau) = -\eta(y)[\chi(y, \tau) - \eta(y)]U(z)$$

Meridional flow velocity:

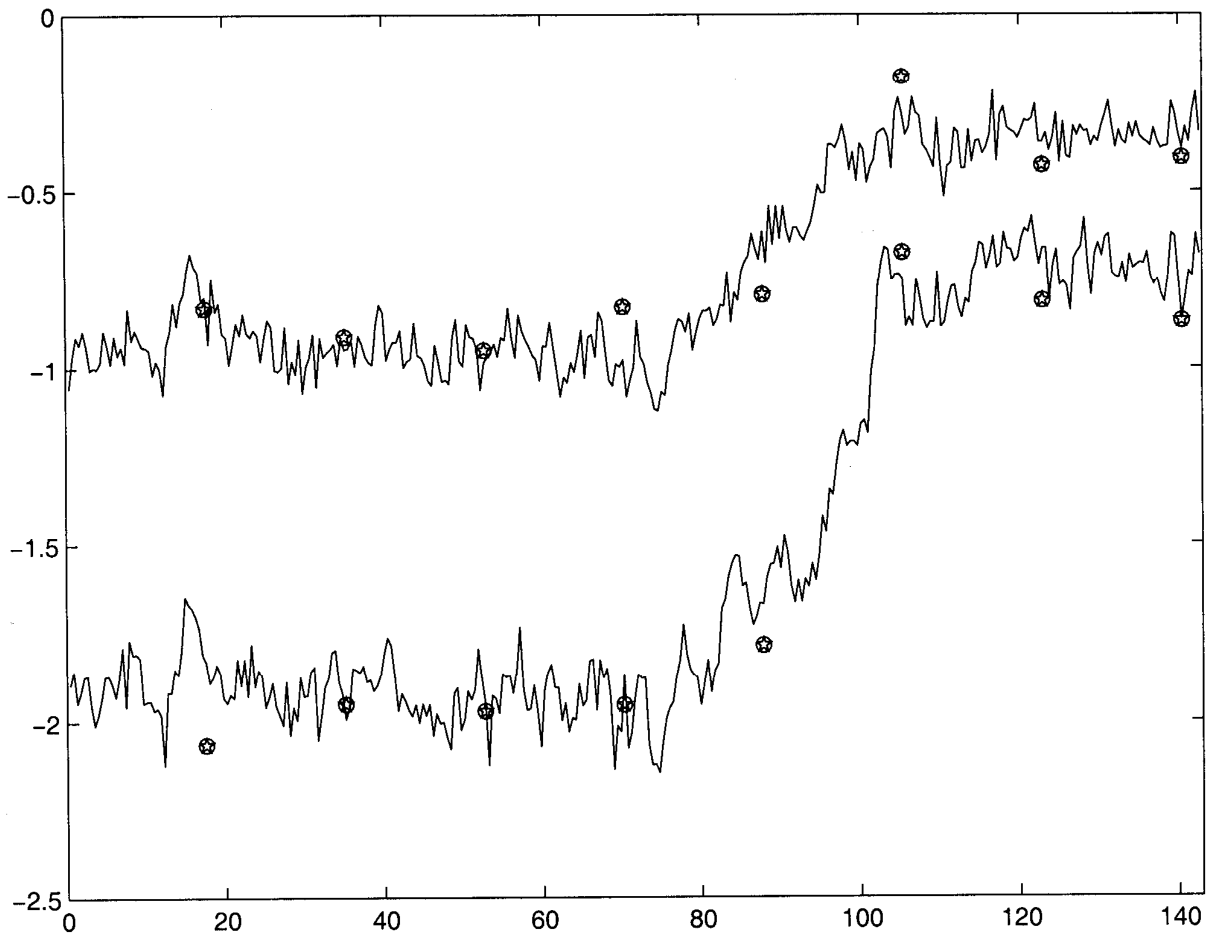
Oops!

$$w(y, z; \tau) = \partial_y[\chi(y, \tau) - \eta(y)]\, W(z)$$

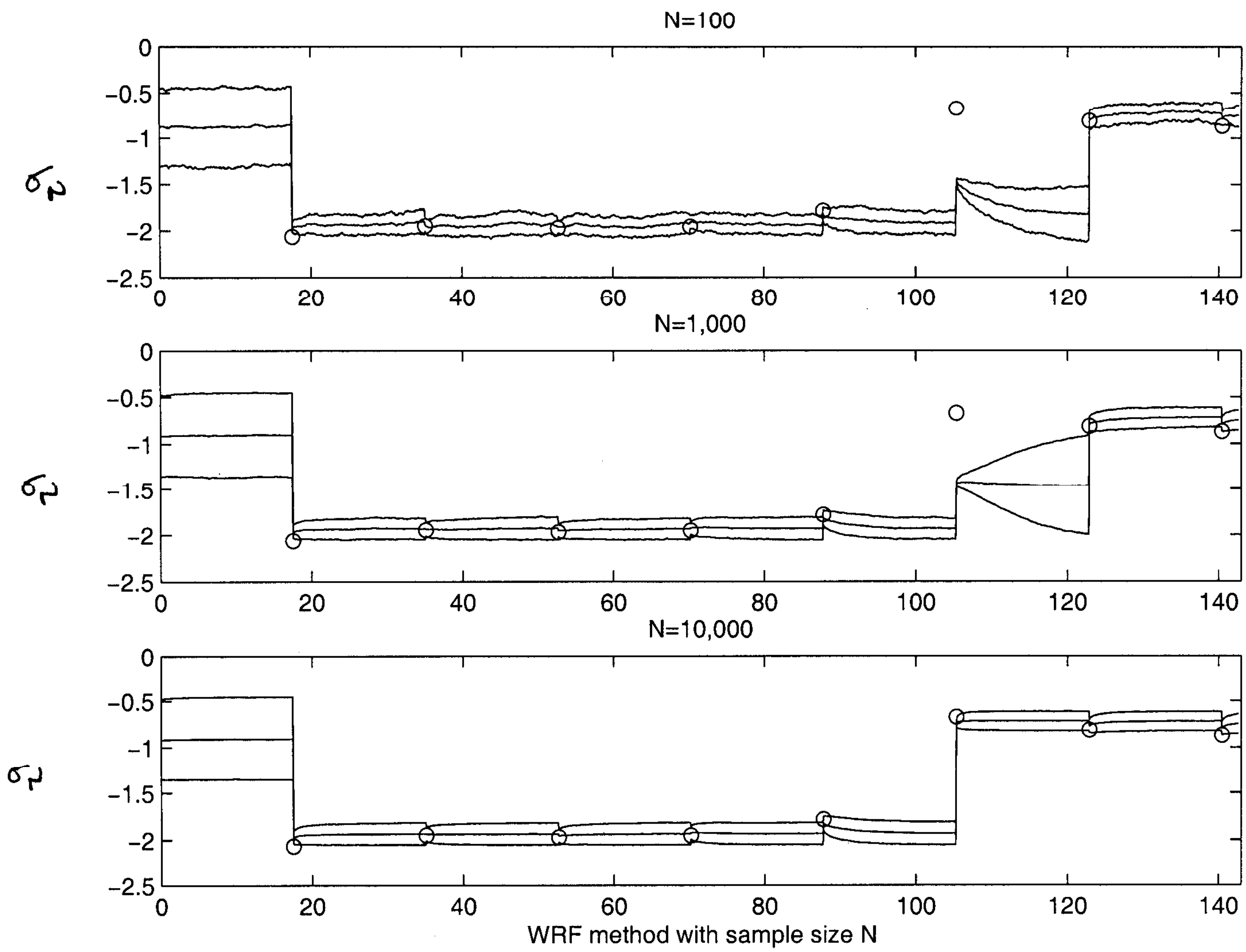
Vertical flow velocity:

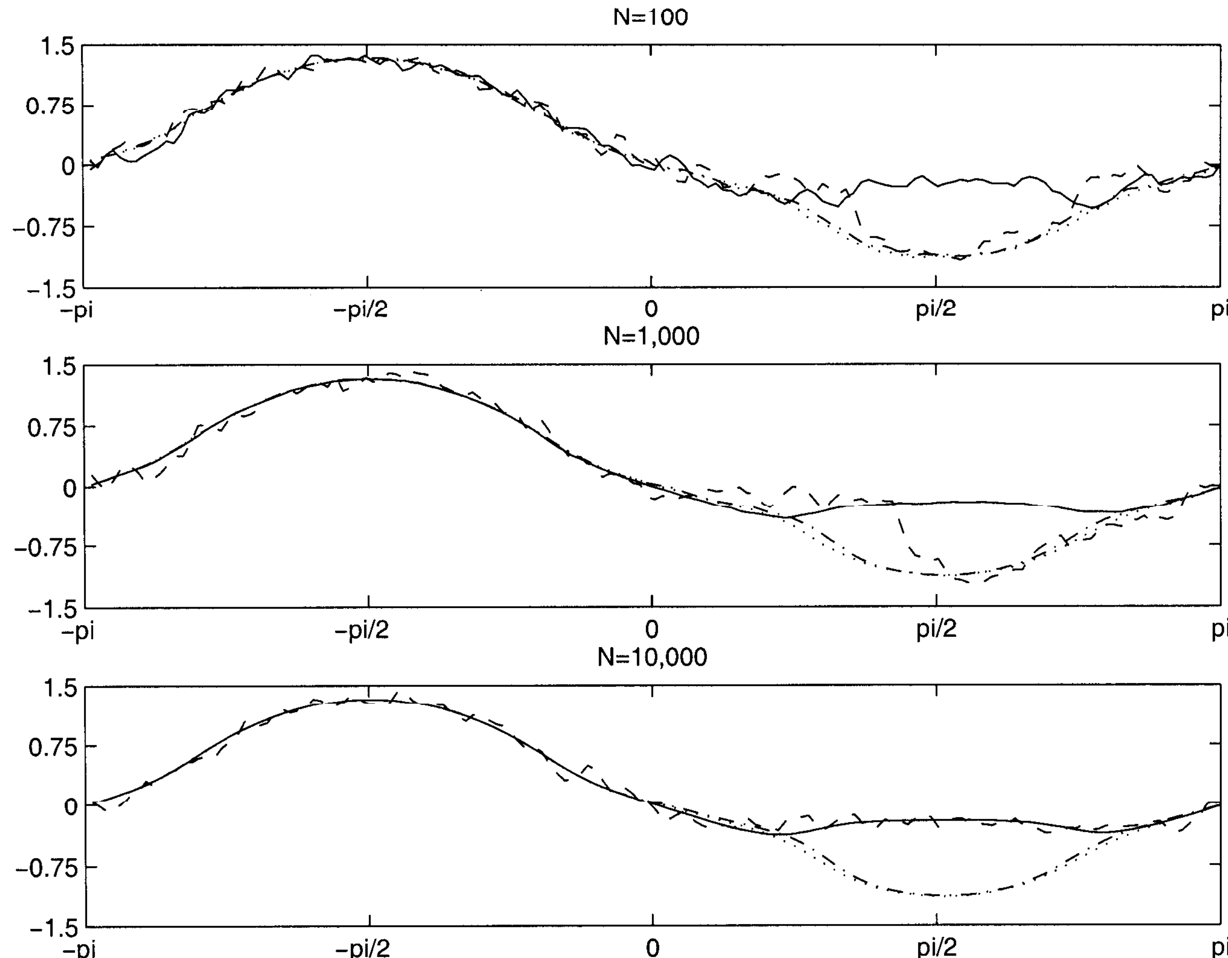
$$v(y, z; \tau) = -[\chi(y, \tau) - \eta(y)]W'(z)$$

All of these gave similar results.

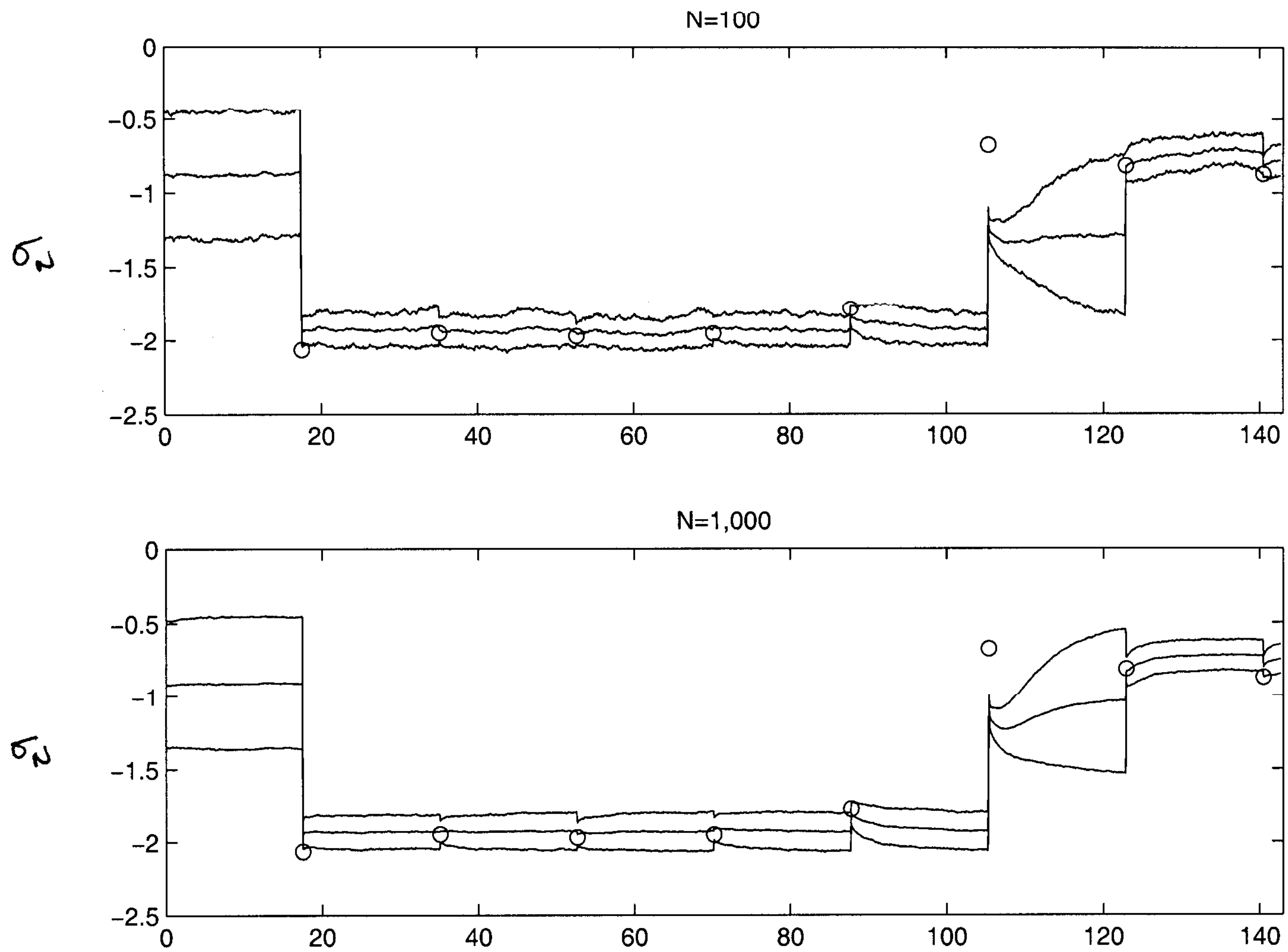


True trajectory and Measurements

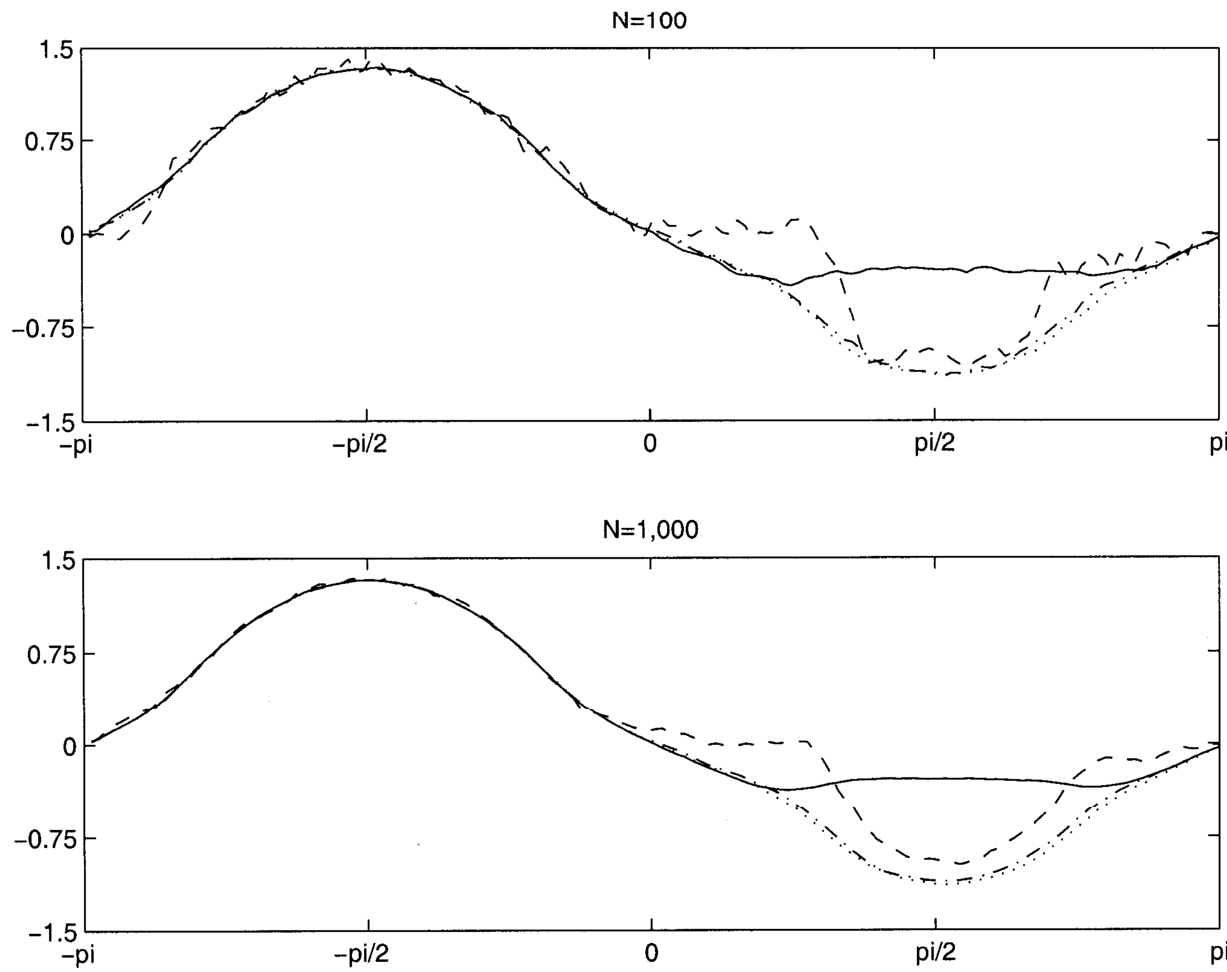

N=100
N=1,000
N=10,000
WRF method with sample size N



Mean profiles of salinity gradient



EnKF with sample size N.



Mean profiles of salinity-gradient

Outline of Maximum-Entropy Filter

Samples $\mathbf{x}_t^{(n)}$, $n = 1, ..., N$ evolve between measurements under the dynamics (1). At measurement times (2), there are three steps:

(i) *Matching:* A parametric model $P(\mathbf{x}; \boldsymbol{\lambda}_{t-})$ is determined by matching to some moments of the ensemble $\mathbf{x}_{t-}^{(n)}$, $n = 1, ..., N$.

(ii) *Updating:* Bayes theorem is now applied to update $P(\mathbf{x}; \boldsymbol{\lambda}_{t-})$ to $P(\mathbf{x}; \boldsymbol{\lambda}_{t+})$.

(iii) *Resampling:* A new N-sample ensemble $\mathbf{x}_{t+}^{(n)}$, $n = 1, ..., N$ is created, by sampling from the model posterior $P(\mathbf{x}; \boldsymbol{\lambda}_{t+})$.

The ensemble $\mathbf{x}_t^{(n)}$, $n = 1, ..., N$ represents the filter distribution $P(\mathbf{x}, t)$.

H-Theorem for Relative Entropy

Let $Q(\mathbf{x},t)$ be the *prior distribution* in the absence of any measurements. For example, if $P_0(\mathbf{x}) = P_*(\mathbf{x})$, the invariant measure of the dynamics (1), then $Q(\mathbf{x},t) = P_*(\mathbf{x})$.

For any nondegenerate Markov process the *relative entropy* or *Kullback-Leibler distance*,

$$H(P(t)|Q(t)) = \int d\mathbf{x}\ P(\mathbf{x},t) \ln\left(\frac{P(\mathbf{x},t)}{Q(\mathbf{x},t)}\right)$$

is non-increasing in time between measurements and vanishes only when $P(t) = Q(t)$.

At long times between measurements $P(\mathbf{x},t)$ "loses information" and converges back toward its prior $Q(\mathbf{x},t)$.

Maximum-Entropy Distributions

The moments of the measured variable,

$$\boldsymbol{\eta}_{t^-} = \langle \mathbf{h}_t \rangle_{t^-}, \ \mathbf{H}_{t^-} = \langle \mathbf{h}_t \mathbf{h}_t^\top \rangle_{t^-},$$

represent the *measurement forecast* at the time t, both the mean $\boldsymbol{\eta}_{t^-}$ and the covariance matrix $\mathbf{C}_{t^-}^H = \mathbf{H}_{t^-} - \boldsymbol{\eta}_{t^-}\boldsymbol{\eta}_{t^-}^\top$.

We take as our model of $P(\mathbf{x}, t^-)$ the *maximum-entropy* (minimum-information) distribution consistent with the measurement forecast. It belongs to an exponential family:

$$P(\mathbf{x}, t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) =$$

$$\frac{\exp[\boldsymbol{\lambda}\cdot\mathbf{h}_t(\mathbf{x}) + \frac{1}{2}\boldsymbol{\Lambda}:\mathbf{h}_t(\mathbf{x})\mathbf{h}_t^\top(\mathbf{x})]}{Z_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})} Q(\mathbf{x}, t)$$

with q-vector $\boldsymbol{\lambda}$ and $q \times q$ symmetric matrix $\boldsymbol{\Lambda}$ as Lagrange multipliers and denominator $Z_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ a normalization factor.

Matching Algorithm

Define convex *cumulant-generating function*

$$F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \log Z_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$$

$$= \log \left[\int d\mathbf{x}\; e^{\boldsymbol{\lambda}\cdot\mathbf{h}_t(\mathbf{x}) + \frac{1}{2}\boldsymbol{\Lambda}:\mathbf{h}_t(\mathbf{x})\mathbf{h}_t^{\top}(\mathbf{x})} Q(\mathbf{x}, t) \right].$$

The moments $(\boldsymbol{\eta}, \mathbf{H})$ are obtained as:

$$\eta_i = \frac{\partial F_t}{\partial \lambda_i}, \quad H_{ij} = \frac{\partial F_t}{\partial \Lambda_{ij}}.$$

The parameters $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ corresponding to given $(\boldsymbol{\eta}, \mathbf{H})$ are determined as optimizers:

$$H_t(\boldsymbol{\eta}, \mathbf{H}) = \sup_{\boldsymbol{\lambda}, \boldsymbol{\Lambda}} \{\boldsymbol{\eta}\cdot\boldsymbol{\lambda} + \frac{1}{2}\mathbf{H}:\boldsymbol{\Lambda} - F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\}$$

which gives the relative entropy for the model density. This involves the minimization of a convex function of $\frac{q(q+3)}{2}$ variables $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. The computational cost is reduced when $q \ll p$

Mixture Models for Priors

We use a Gaussian *mixture model*

$$Q_M(\mathbf{x},t) = \sum_{m=1}^{M} w_m(t) N(\mathbf{x}; \boldsymbol{\mu}_m(t), \mathbf{C}_m(t)).$$

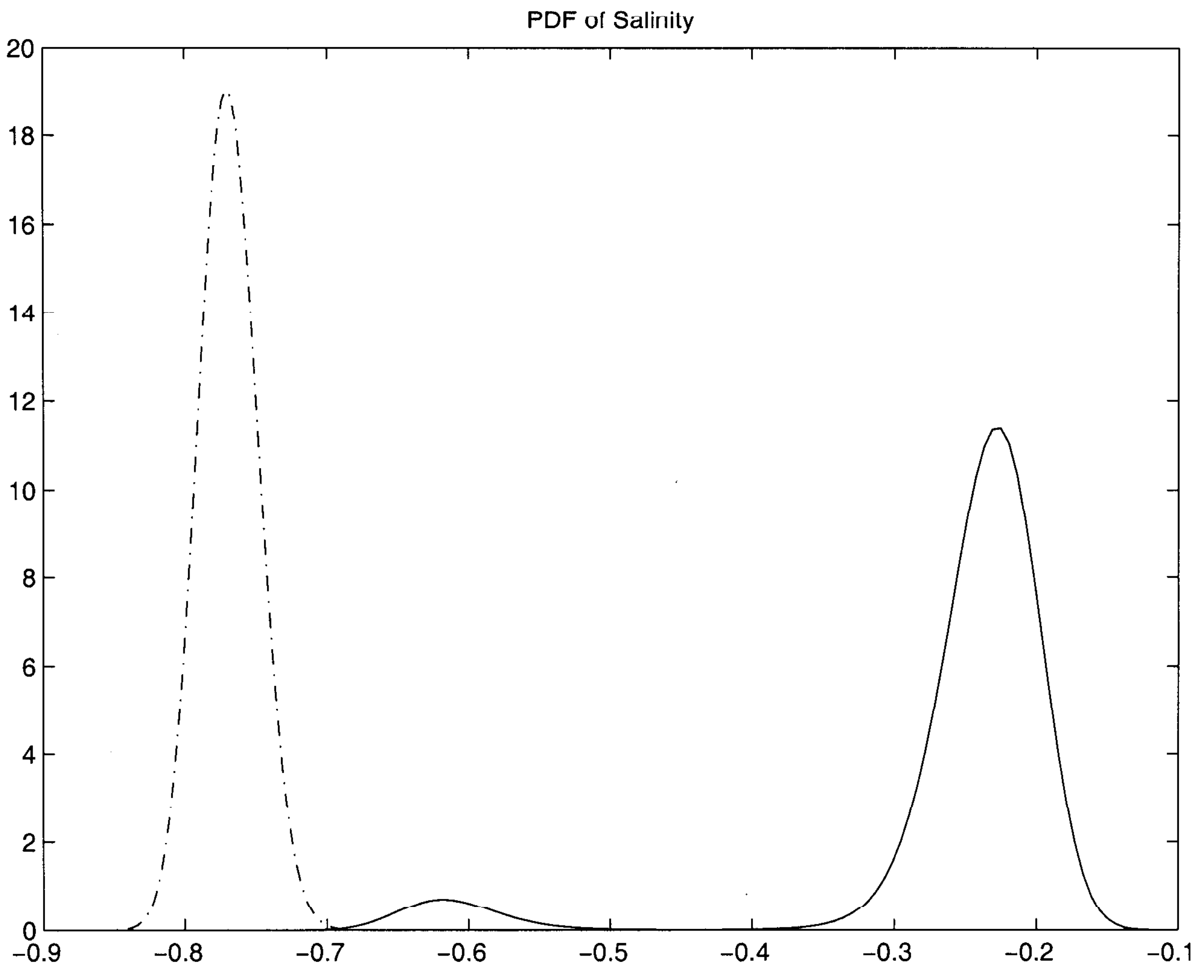
The weights of the components satisfy

$$\sum_{m=1}^{M} w_m(t) = 1,$$

and $N(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$.

If there is one Gaussian component ($M = 1$) if we match all the moments $\langle \mathbf{x} \rangle$, $\langle \mathbf{x}\mathbf{x}^\top \rangle$ and if $\mathbf{h}_t(\mathbf{x}) = \boldsymbol{\mathcal{H}}_t \mathbf{x} + \mathbf{d}_t$ is affine, then our method is equivalent to Ensemble Kalman Filter.

To construct $w_m(t), \boldsymbol{\mu}_m(t), \mathbf{C}_m(t)$ we will use conditional sampling of a solution of (1).


PDF of Salinity
20
18
16
14
12
10
8
6
4
2
0
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1

Mixture Model for Cessi-Young

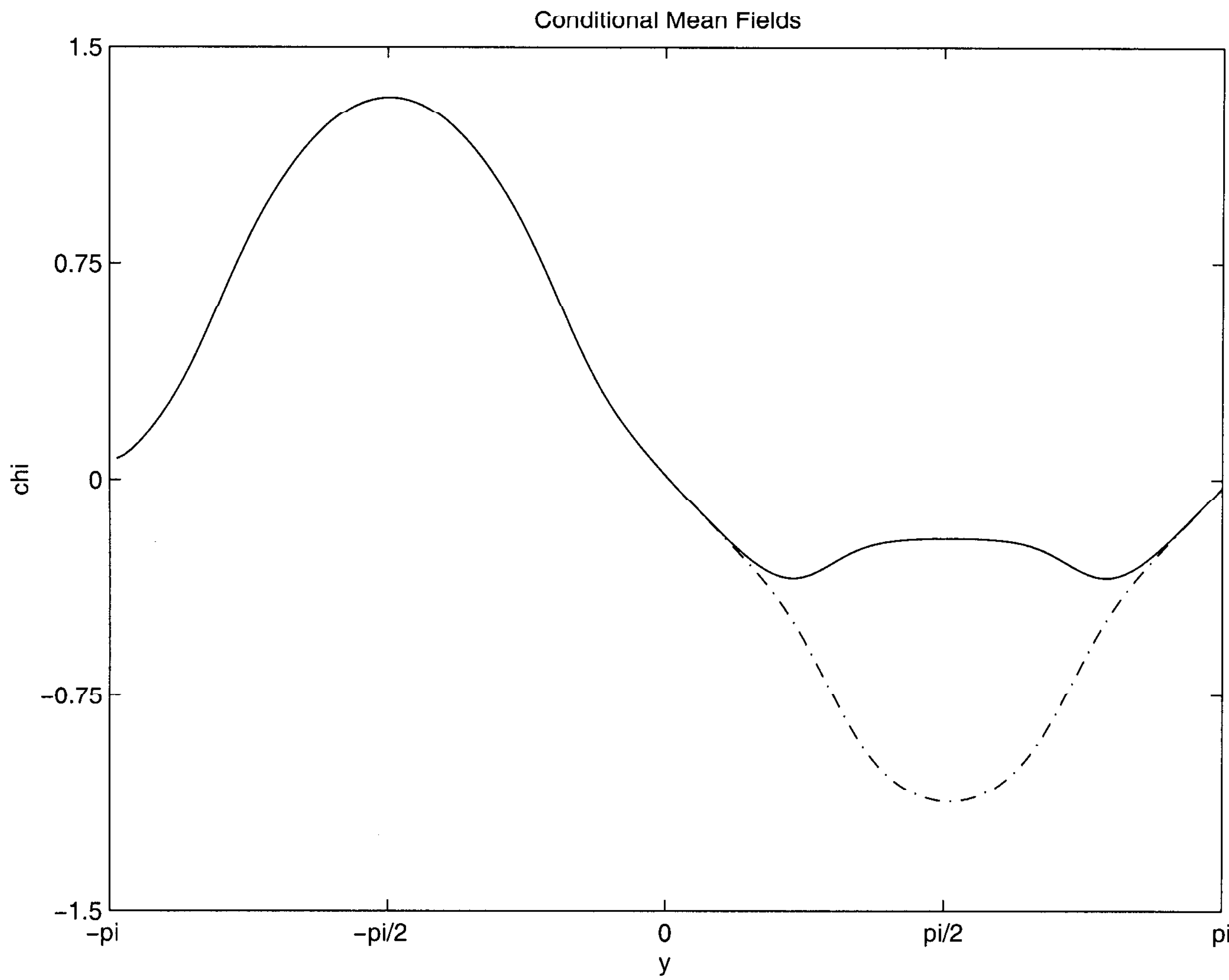
With $\sigma_0 = 0.08$ we perform conditional sampling on the two events

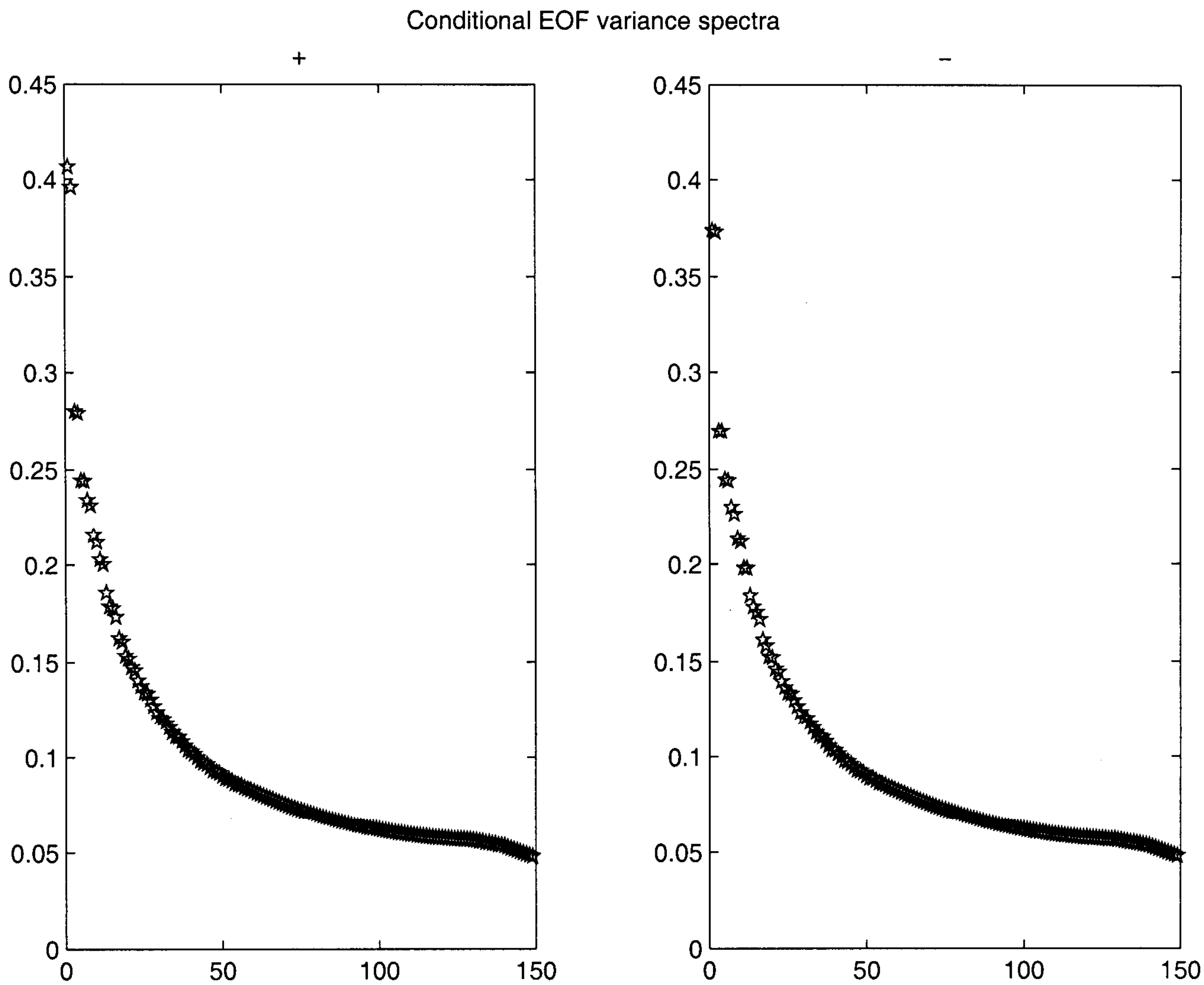
$$E_- = \{\sigma_N < -0.45\} \ \text{weak circulation}$$

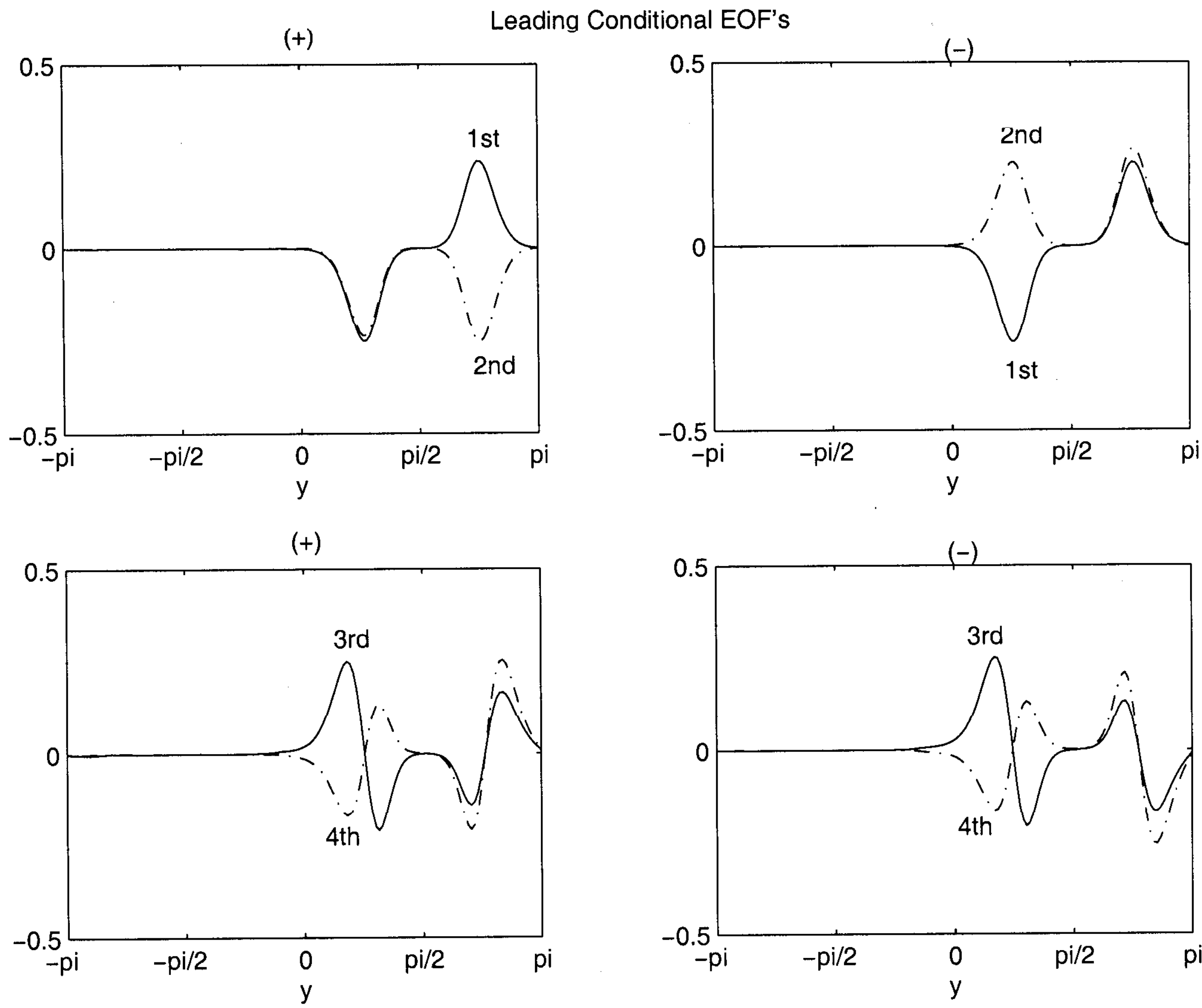
$$E_+ = \{\sigma_N > -0.45\} \ \text{strong circulation}$$

From this we obtain:

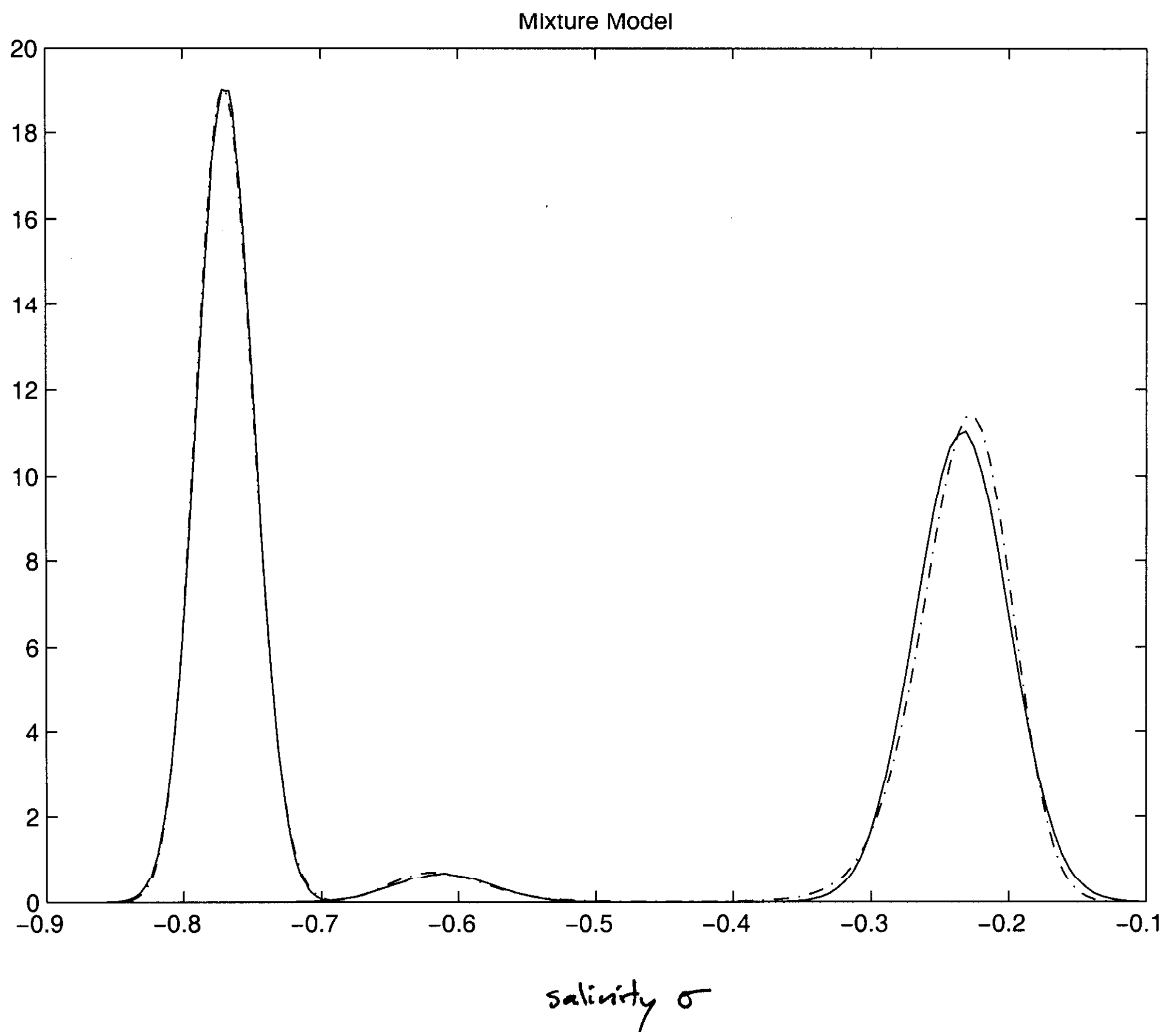
- Weights: $w_\pm = P(E_\pm)$
- Conditional mean profiles $\bar{\chi}_\pm(y_k)$, $k = 1, ..., 150$
- Conditional covariance matrices $\mathbf{C}_\pm(y_k, y_l)$, $k, l = 1, ..., 150$
- Conditional EOF variance spectra $\gamma_\pm^{(a)}$, $a = 1, ..., 150$
- Conditional EOF's $\hat{\mathbf{e}}_\pm^{(a)}, a = 1, ..., 150$


Conditional Mean Fields
1.5
0.75
0
-0.75
-1.5
chi
-pi
-pi/2
0
pi/2
pi
y


Conditional EOF variance spectra
+
−
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
50
100
150



Leading Conditional EOF's
(+)
0.5
1st
0
2nd
−0.5
−pi
−pi/2
0
pi/2
pi
y
(−)
0.5
2nd
0
1st
−0.5
−pi
−pi/2
0
pi/2
pi
y
(+)
0.5
3rd
0
4th
−0.5
−pi
−pi/2
0
pi/2
pi
y
(−)
0.5
3rd
0
4th
−0.5
−pi
−pi/2
0
pi/2
pi
y


Mixture Model
20
18
16
14
12
10
8
6
4
2
0
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
salinity σ

Generalized Representer Algorithm

Approximating $Q \approx Q_M$, the model distribution with parameters $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is also a mixture:

$$P_M(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) =$$

$$\sum_{m=1}^{M} w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) N(\mathbf{x}; \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(\boldsymbol{\Lambda})).$$

To determine $w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(\boldsymbol{\Lambda})$:

*Define q-vectors $\boldsymbol{\mu}_m^H = \boldsymbol{\mathcal{H}}\boldsymbol{\mu}_m + \mathbf{d}$ and $q \times q$ matrices $\mathbf{C}_m^H = \boldsymbol{\mathcal{H}}\mathbf{C}_m\boldsymbol{\mathcal{H}}^\top$, $\boldsymbol{\Gamma}_m^H = [\mathbf{C}_m^H]^{-1}$

*Determine the Cholesky decomposition of $\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}$, $m = 1, ..., M$. These matrices must be positive-definite for the model density to be statistically realizable with the given $\boldsymbol{\Lambda}$.

* Using the Cholesky factorizations, solve the linear equation

$$(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}) \cdot \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda}$$

and calculate the inverse $[\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}]^{-1}$ and determinant $\mathrm{Det}\,(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})$, $m = 1, ..., M$.

Finally, for $m = 1, ..., M$, set

$$Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sqrt{\frac{\mathrm{Det}\,\boldsymbol{\Gamma}_m^H}{\mathrm{Det}\,(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})}} \times$$

$$\exp\left[-\frac{1}{2}(\boldsymbol{\mu}_m^H)^\top \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \frac{1}{2}(\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda})^\top \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\right].$$

$$w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = w_m \frac{Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})}{Z(\boldsymbol{\lambda}, \boldsymbol{\Lambda})}$$

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Gamma}_m^H \cdot [\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) - \boldsymbol{\mu}_m^H]$$

$$\mathbf{C}_m(\boldsymbol{\Lambda}) = \mathbf{C}_m + \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Gamma}_m^H [\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}]^{-1} \boldsymbol{\Lambda} \boldsymbol{\mathcal{H}} \mathbf{C}_m,$$

and

$$Z(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sum_{m=1}^{M} w_m Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$$

$$F(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \ln Z(\boldsymbol{\lambda}, \boldsymbol{\Lambda}).$$

$$\frac{\partial F}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$$

$$\frac{\partial F}{\partial \Lambda_{ij}} = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) \{ [(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1}]_{ij}$$

$$+ [\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})]_i [\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})]_j \}$$

We use these to determine

$$(\boldsymbol{\lambda}_{t^-}, \boldsymbol{\Lambda}_{t^-}) =$$

$$\text{argsup}_{\boldsymbol{\lambda}, \boldsymbol{\Lambda}} \{ \boldsymbol{\eta}_{t^-} \cdot \boldsymbol{\lambda} + \frac{1}{2} \mathbf{H}_{t^-} : \boldsymbol{\Lambda} - F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) \}$$

by conjugate-gradient minimization with a feasible Armijo line-search. We can monitor if trials $(\boldsymbol{\lambda}_k, \boldsymbol{\Lambda}_k)$ in the CG iteration remain in the domain of $F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ by existence of the Cholesky factorizations of $\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda},\ m = 1, ..., M.$

Updating the Model Distribution

Bayes theorem is now applied, which, for normal error statistics,

$$\mathbf{y}_t = \mathbf{h}_t(\mathbf{x}_t) + \boldsymbol{\epsilon}_t$$

$$\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{R}_t)$$

yields another maximum-entropy distribution with parameters $(\boldsymbol{\lambda}_{t+}, \boldsymbol{\Lambda}_{t+})$ given by

$$\boldsymbol{\lambda}_{t+} = \boldsymbol{\lambda}_{t-} + \mathbf{R}_t^{-1}\mathbf{y}_t,$$

$$\boldsymbol{\Lambda}_{t+} = \boldsymbol{\Lambda}_{t-} - \mathbf{R}_t^{-1}.$$

Updating the model distribution is trivial!

Resampling the Model Distribution

With $(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = (\boldsymbol{\lambda}_{t+}, \boldsymbol{\Lambda}_{t+})$, resample

$$P_M(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) N(\mathbf{x}; \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(\boldsymbol{\Lambda})).$$

by repeating the following steps for $n = 1, ..., N$:

(1) Choose a component m_n with probability $w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}),\ m = 1, ..., M$

(2) Sample an element $\mathbf{x}_n$ from the distribution $N(\boldsymbol{\mu}_{m_n}(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_{m_n}(\boldsymbol{\Lambda}))$ using its Karhunen-Loève representation:

$$\mathbf{x}_n = \boldsymbol{\mu}_{m_n}(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) + \sum_{a=1}^{p} \xi_n^{(a)} \sqrt{\gamma_{m_n}^{(a)}(\boldsymbol{\Lambda})}\ \hat{\mathbf{e}}_{m_n}^{(a)}(\boldsymbol{\Lambda}).$$

Here $\gamma_m^{(a)}(\boldsymbol{\Lambda}), \hat{\mathbf{e}}_m^{(a)}(\boldsymbol{\Lambda})$ are the eigenvalues and eigenvectors of $\mathbf{C}_m(\boldsymbol{\Lambda})$ and $\xi_n^{(a)}$ are i.i.d. normal random variables, $a = 1, ..., p$, $n = 1, ..., N$

Calculating $\gamma_m^{(a)}(\mathbf{\Lambda}), \hat{\mathbf{e}}_m^{(a)}(\mathbf{\Lambda})$ *for every new value of* $\mathbf{\Lambda}$ *is expensive!*

To avoid this, sample $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda}), \mathbf{C}_m(\mathbf{\Lambda}))$ by the Metropolis-Hastings algorithm with the Gaussian $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda}), \mathbf{C}_m)$ as the proposal distribution. Thus, proposed updates have the form

$$\mathbf{x}' = \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda}) + \sum_{a=1}^{p} \xi_a \sqrt{\gamma_m^a}\, \hat{\mathbf{e}}_m^a,$$

where $\gamma_m^a, \hat{\mathbf{e}}_m^a$ are the eigenvalues and eigenvectors of $\mathbf{C}_m$. (Note that $\mathbf{C}_m$ does not depend on $\mathbf{\Lambda}$!) These are the *conditional EOF's*.

These updates are accepted with probability $\min\{1, e^{-\Delta E}\}$ to replace a current state vector $\mathbf{x}$, where $\Delta E = E(\mathbf{x}') - E(\mathbf{x})$ and

$$E(\mathbf{x}) =$$

$$-\frac{1}{2}[\mathbf{h}(\mathbf{x}) - \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \mathbf{\Lambda})]^\top \mathbf{\Lambda} [\mathbf{h}(\mathbf{x}) - \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \mathbf{\Lambda})]$$

Costs of the Algorithm

Matching: Calculation of F_t and its gradients at one value of $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ requires $O(Mq^3)$ multiplications. The total cost of the minimization by conjugate-gradient is $O(n_{CG}Mq^3)$, where n_{CG} is the number of CG iterations.

Resampling: To calculate EOF's of $\mathbf{C}_m$, $m = 1, ..., M$ at the outset is a single-time cost of $O(Mp^3)$. If a number n_T of trials is made in each Metropolis step, then resampling requires Npn_T random numbers and $O(Np^2n_T)$ multiplications at each measurement time.

To simplify, truncate K-L expansion to a maximum number of EOF's $p_{\text{max}} \ll p$. Finding the p_{max} leading eigenvalues and eigenvectors of $\mathbf{C}_m$, $m = 1, ..., M$ requires $O(Mp^2p_{\text{max}})$ operations, e.g. by iterative Arnoldi methods. Likewise, Metropolis sampling from the truncated K-L expansion uses $Np_{\text{max}}n_T$ random numbers and $O(Npp_{\text{max}}n_T)$ multiplications. These are smaller by a factor of $p_{\text{max}}/p \ll 1$.

Mean-Field Filter

Matching: Minimize $H(P_t|Q_t)$ subject to the single constraint $\langle \mathbf{h}_t \rangle_{t-} = \boldsymbol{\eta}_{t-}$. This gives

$$P(\mathbf{x}, t; \boldsymbol{\lambda}) = \frac{1}{Z_t(\boldsymbol{\lambda})} \exp[\boldsymbol{\lambda} \cdot \mathbf{h}_t(\mathbf{x})] \cdot Q(\mathbf{x}, t)$$

with q-vector $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{t-}$ yielding the supremum

$$H_t(\boldsymbol{\eta}) = \sup_{\boldsymbol{\lambda}} \{\boldsymbol{\eta} \cdot \boldsymbol{\lambda} - F_t(\boldsymbol{\lambda})\}$$

for $\boldsymbol{\eta} = \boldsymbol{\eta}_{t-}$. Here $F_t(\boldsymbol{\lambda}) = \log Z_t(\boldsymbol{\lambda})$.

Updating:

$$\boldsymbol{\eta}_{t+} = \operatorname*{arginf}_{\boldsymbol{\eta}} \{H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t-}) + \frac{1}{2}[\boldsymbol{\eta} - \mathbf{y}_t]^\top \mathbf{R}_t^{-1} [\boldsymbol{\eta} - \mathbf{y}_t]\Big\}$$

where

$$H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t-}) = H_t(\boldsymbol{\eta}) - H_t(\boldsymbol{\eta}_{t-}) - (\boldsymbol{\eta} - \boldsymbol{\eta}_{t-}) \cdot \boldsymbol{\lambda}_{t-}.$$

Resampling: Essentially the same as before.

Interpretation of Mean-Field Update

Suppose samples $\mathbf{x}_{t-}^{(n)}$, $n = 1, ..., N$ are drawn independently from the distribution $P(\mathbf{x}, t; \boldsymbol{\lambda}_{t-})$.

Also take an i.i.d. set $\{\boldsymbol{\epsilon}_t^{(n)},\ n = 1, ..., N\}$ of $N(\mathbf{0}, \mathbf{R}_t)$ random variables and define the ensemble of measured values

$$\mathbf{y}_t^{(n)} = \mathbf{h}_t(\mathbf{x}_{t-}^{(n)}) + \boldsymbol{\epsilon}_t^{(n)},\ n = 1, ..., N.$$

Then:

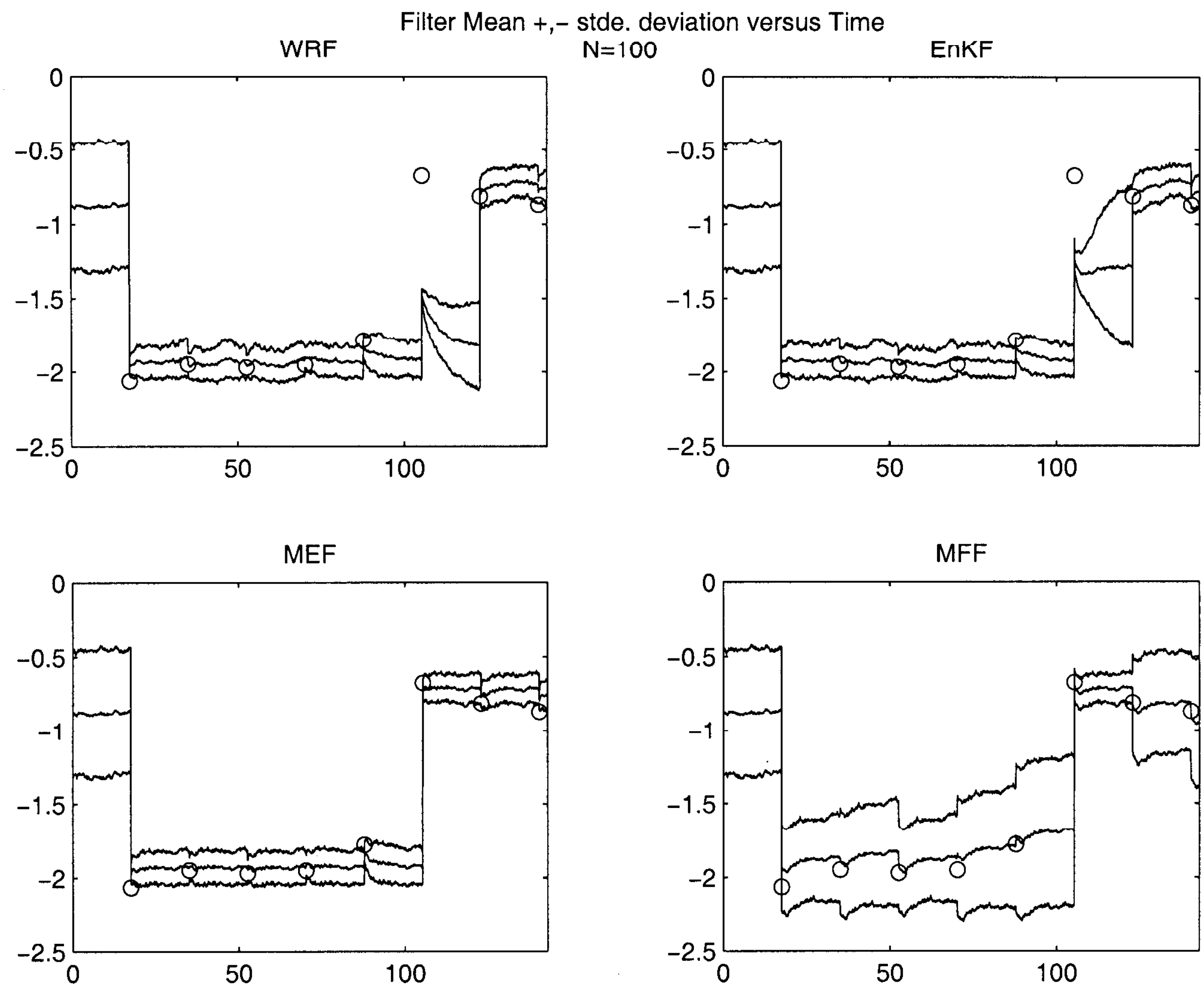
$\boldsymbol{\eta}_{t+}$ is the most probable value of $\frac{1}{N}\sum_{n=1}^{N} \mathbf{h}_t(\mathbf{x}_{t-}^{(n)})$ for the ensemble conditioned upon the event $\frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_t^{(n)} = \mathbf{y}$, in the limit as $N \to \infty$.

Computational Costs

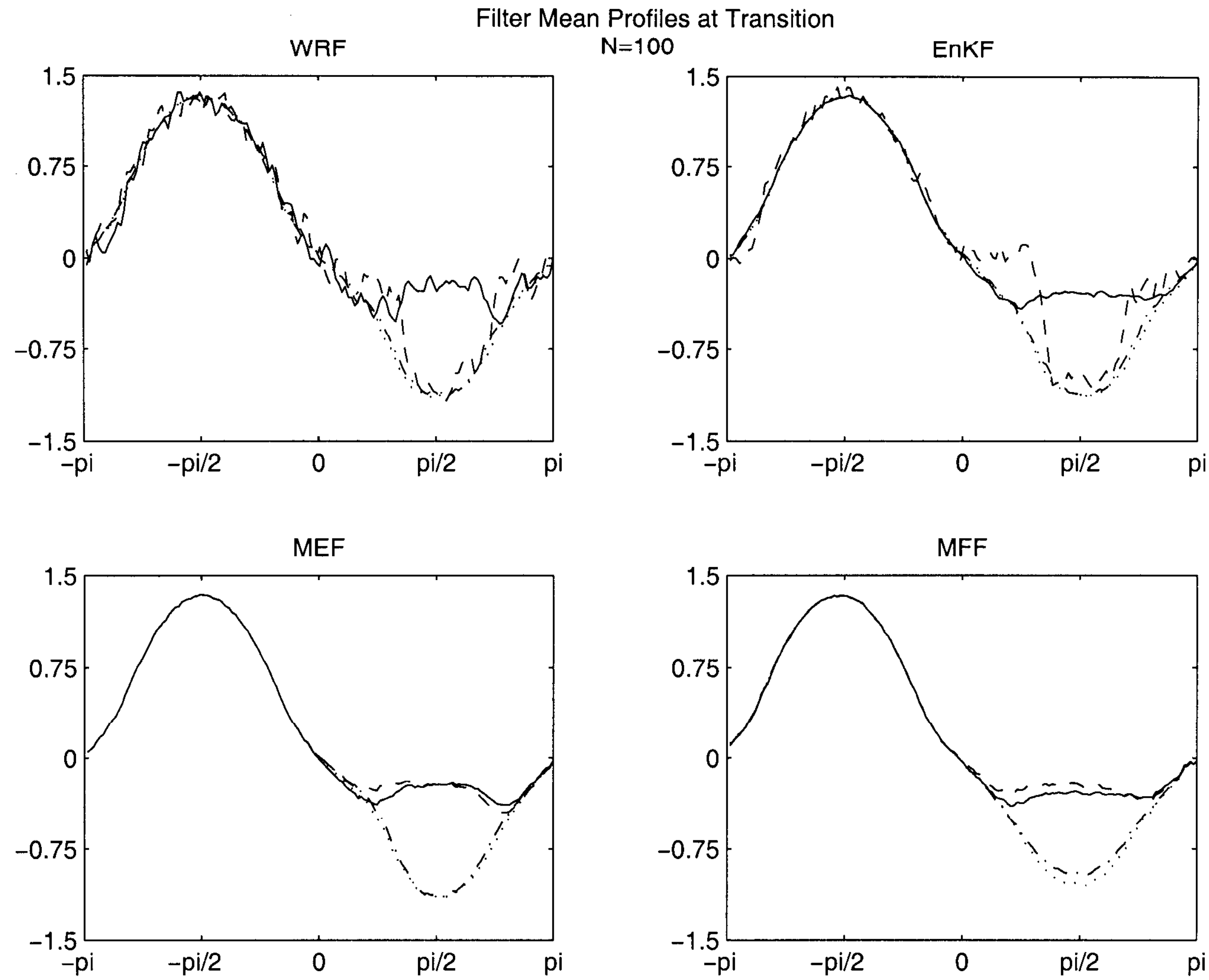
The cost to calculate $\boldsymbol{\eta}_m(\boldsymbol{\lambda}), Z_m(\boldsymbol{\lambda})$ for $m = 1, ..., M$ and $F(\boldsymbol{\lambda})$ and its derivatives is $O(Mq^2)$. Hence, the total cost of the matching step is $O(n_{CG}Mq^2)$. This is smaller by $1/q$ than for MEF and smaller by $O(n_{CG}M(q/p)^2/q)$ than the cost of the Kalman gain matrix in EnKF.

The resampling step in the mean-field MEF uses $O(Mpq)$ multiplications to calculate the means $\boldsymbol{\mu}_m(\boldsymbol{\lambda})$, $m = 1, ..., M$. As in MEF, there is a single-time expense of $O(Mp^3)$ to calculate EOF's of the component covariances $\mathbf{C}_m$, $m = 1, ..., M$. Also, Np random numbers and $O(Np^2)$ multiplications are needed to generate new samples.

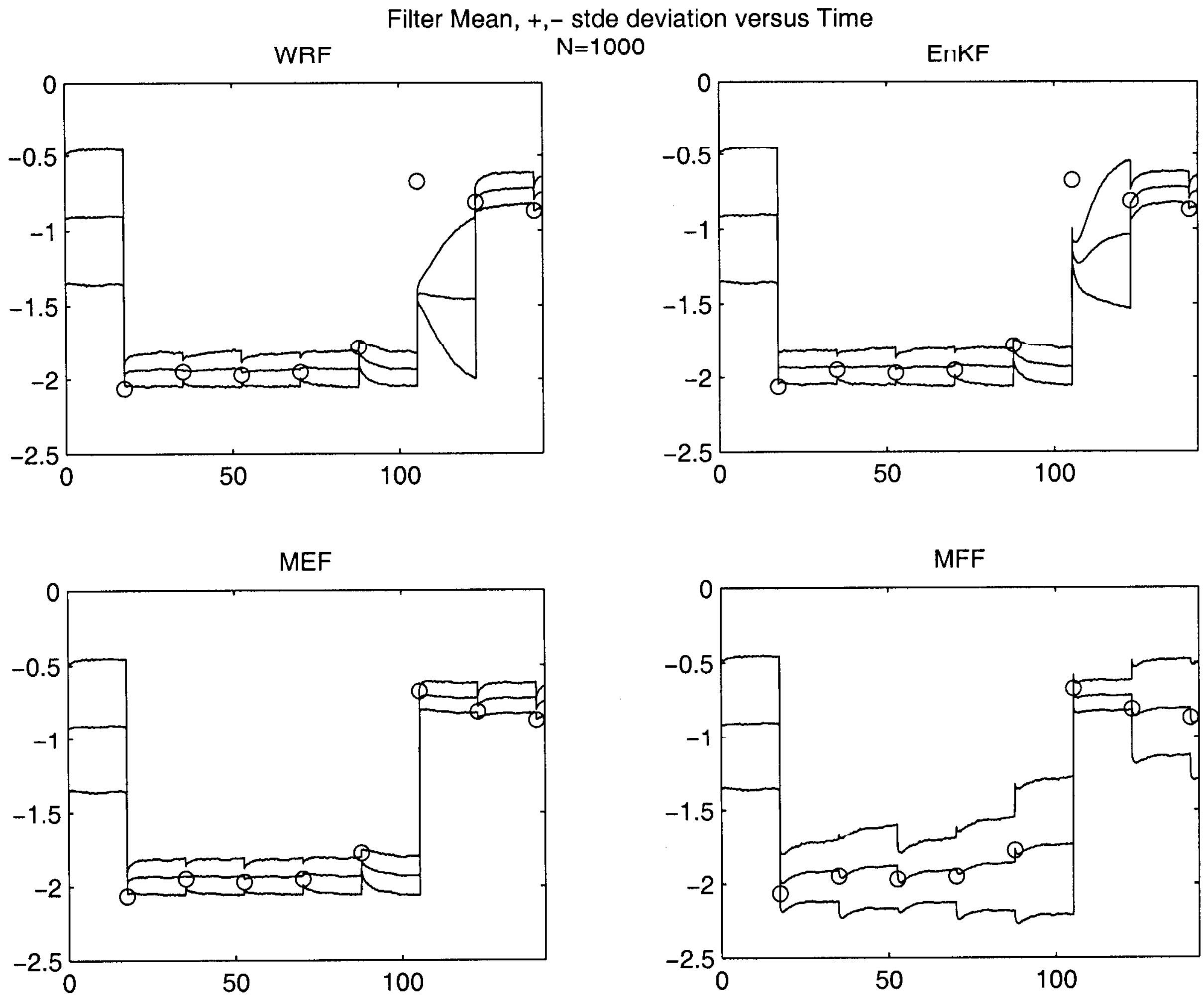
Thus, resampling in MFF is cheaper than in full MEF by a factor of $1/n_T$ and more expensive than in EnKF by a factor of p/q. However, if a truncated K-L expansion is used with only p_{max} terms, then this factor is p_{max}/q and the cost will be similar as for EnKF if $p_{\text{max}} \approx q$.

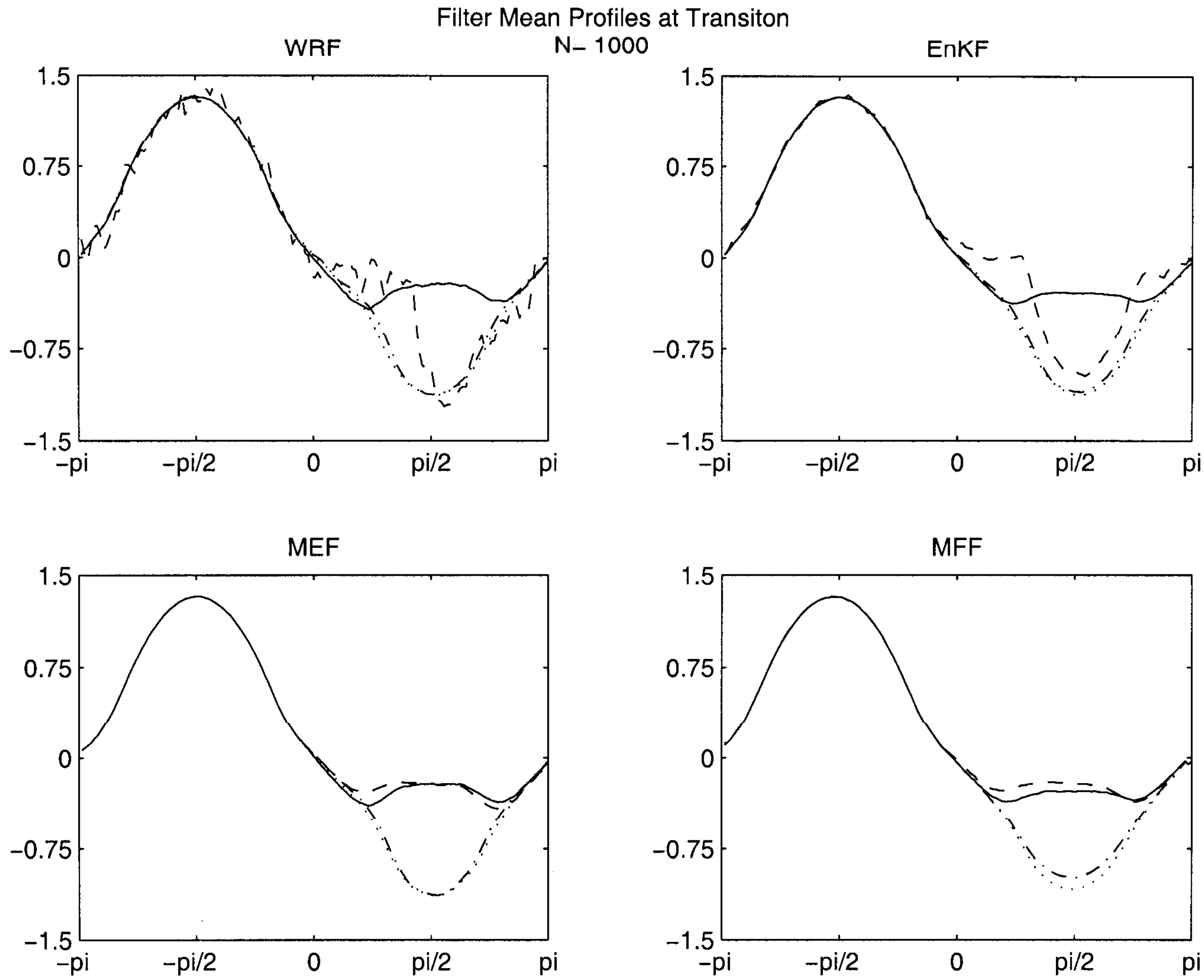


Filter Mean +,- stde. deviation versus Time
N=100
WRF
EnKF
MEF
MFF
0
-0.5
-1
-1.5
-2
-2.5
0
50
100



Filter Mean Profiles at Transition
N=100
WRF
EnKF
MEF
MFF
1.5
0.75
0
−0.75
−1.5
−pi
−pi/2
0
pi/2
pi


Filter Mean, +,- stde deviation versus Time
N=1000
WRF
EnKF
MEF
MFF
0
-0.5
-1
-1.5
-2
-2.5
0
50
100


Filter Mean Profiles at Transiton
N– 1000
WRF
EnKF
MEF
MFF
1.5
0.75
0
−0.75
−1.5
−pi
−pi/2
0
pi/2
pi

Entropy

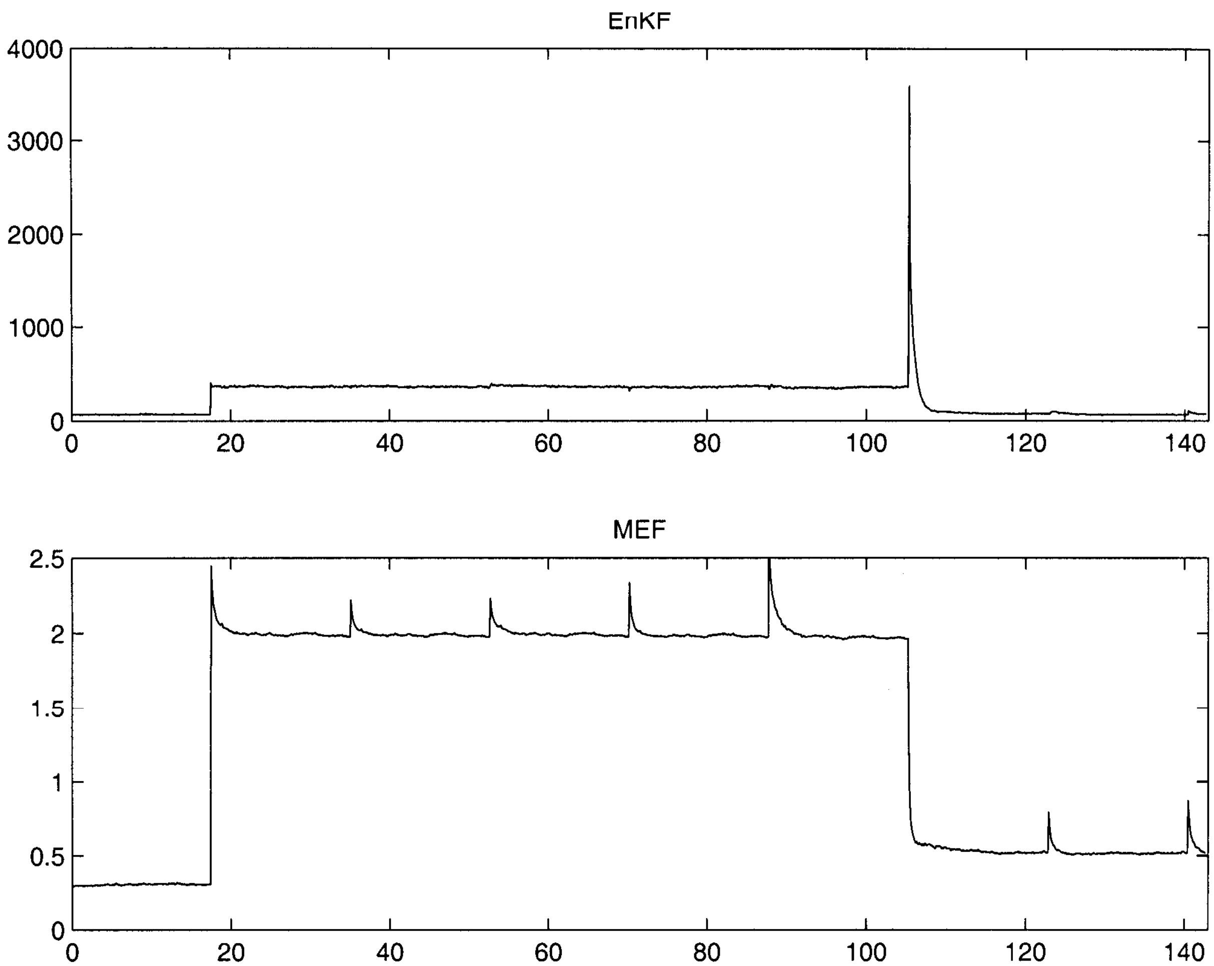
MEF and MFF yield as by-products estimates of the relative entropy $H(P_t|Q_t)$, i.e. $H_t(\boldsymbol{\eta}, \mathbf{H})$ and $H_t(\boldsymbol{\eta})$. These can be calculated at any time desired by matching to the particle ensemble moments $\boldsymbol{\eta}_t, \mathbf{H}_t$.

EnKF also gives an estimate, if one assumes a pair of normal densities $P = N(\boldsymbol{\mu}_t, \mathbf{C}_t), Q = N(\boldsymbol{\nu}_t, \mathbf{G}_t)$:

$$H(P_t|Q_t) = \frac{1}{2}(\boldsymbol{\mu}_t - \boldsymbol{\nu}_t)^\top \mathbf{G}_t^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\nu}_t)$$

$$+\frac{1}{2}\text{Tr}\,[\mathbf{C}_t\mathbf{G}_t^{-1} - \mathbf{I}] - \frac{1}{2}\ln\left(\frac{\text{Det}\,\mathbf{C}_t}{\text{Det}\,\mathbf{G}_t}\right).$$

However, it is very expensive to calculate the determinant $\text{Det}\,\mathbf{C}_t$ at each desired time t, needing $O(p^3)$ multiplications.



Entropy versus Time (N = 1,000)

Log-Likelihood & Parameter Estimation

All filtering schemes yield the log-likelihood in the *innovation form* $\Lambda_T = \sum_{t=1}^{T} \ln \mathcal{N}_t$ where the sum is over measurement times and $\mathcal{N}_t$ is the normalization in Bayes theorem.

MEF: With $\Delta F_t = F_t(\boldsymbol{\lambda}_{t+}, \boldsymbol{\Lambda}_{t+}) - F_t(\boldsymbol{\lambda}_{t-}, \boldsymbol{\Lambda}_{t-})$

$$\ln \mathcal{N}_t = \Delta F_t - \frac{1}{2}\mathbf{y}_t^\top \mathbf{R}_t^{-1}\mathbf{y}_t - \frac{1}{2}\log[(2\pi)^q \mathrm{Det}\,\mathbf{R}_t]$$

MFF: $\ln \mathcal{N}_t = -H_t^Y(\mathbf{y}_t|\boldsymbol{\eta}_{t-})$ with

$$H_t^Y(\mathbf{y}|\boldsymbol{\eta}_{t-}) = \min_{\boldsymbol{\eta}} \Big\{ H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t}\ ) + \frac{1}{2}[\boldsymbol{\eta} - \mathbf{y}]^\top \mathbf{R}_t^{-1}[\boldsymbol{\eta} - \mathbf{y}] \Big\}$$

EnKF: With $\boldsymbol{\mu}_{t-}^Y = \boldsymbol{\mu}_{t-}^{H}$, $\mathbf{C}_{t-}^Y = \mathbf{C}_{t-}^H + \mathbf{R}_t$.

$$\ln \mathcal{N}_t = -\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}_{t-}^Y)^\top (\mathbf{C}_{t-}^Y)^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{t-}^Y) - \frac{1}{2}\log[(2\pi)^q \mathrm{Det}\,(\mathbf{C}_{t-}^Y)]$$

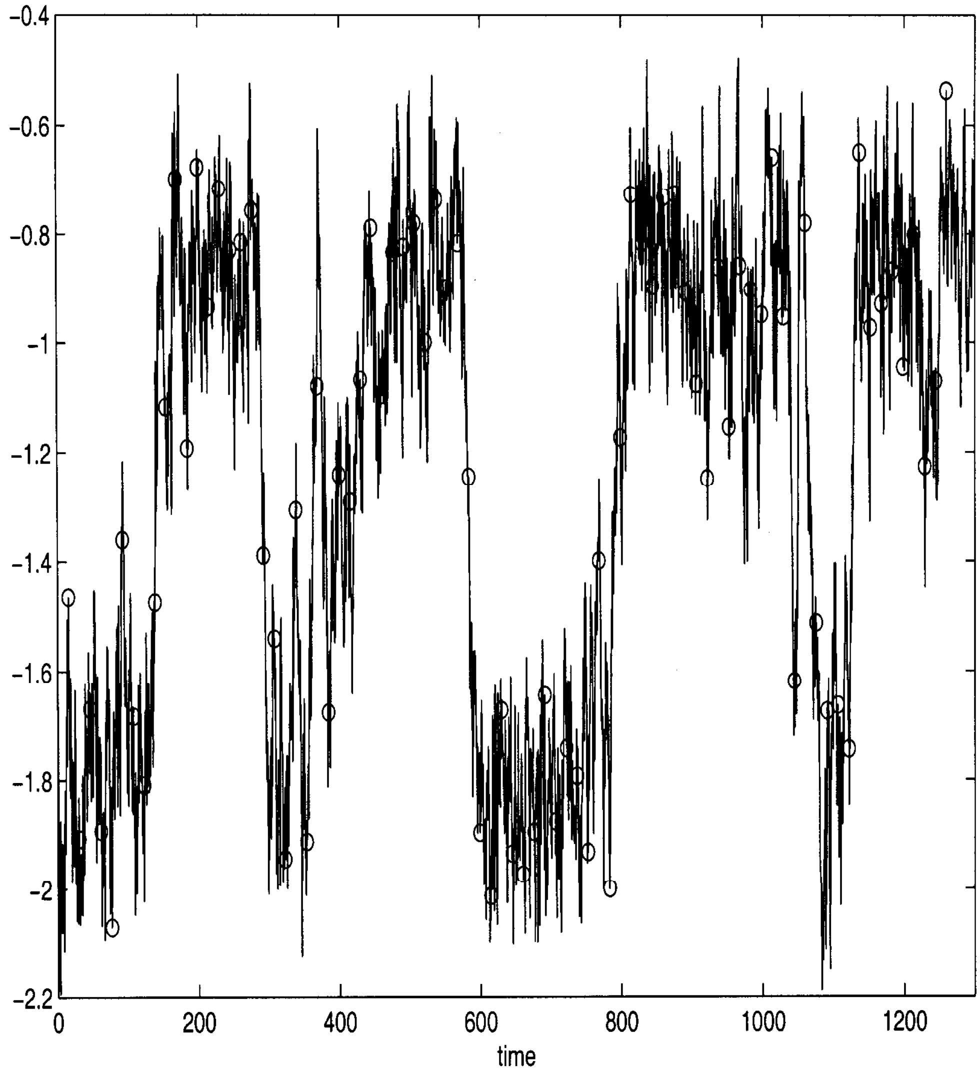


Figure 1: Evolution of salinity at the north pole with $\sigma = 0.115$.

Maximum-Likelihood Estimates

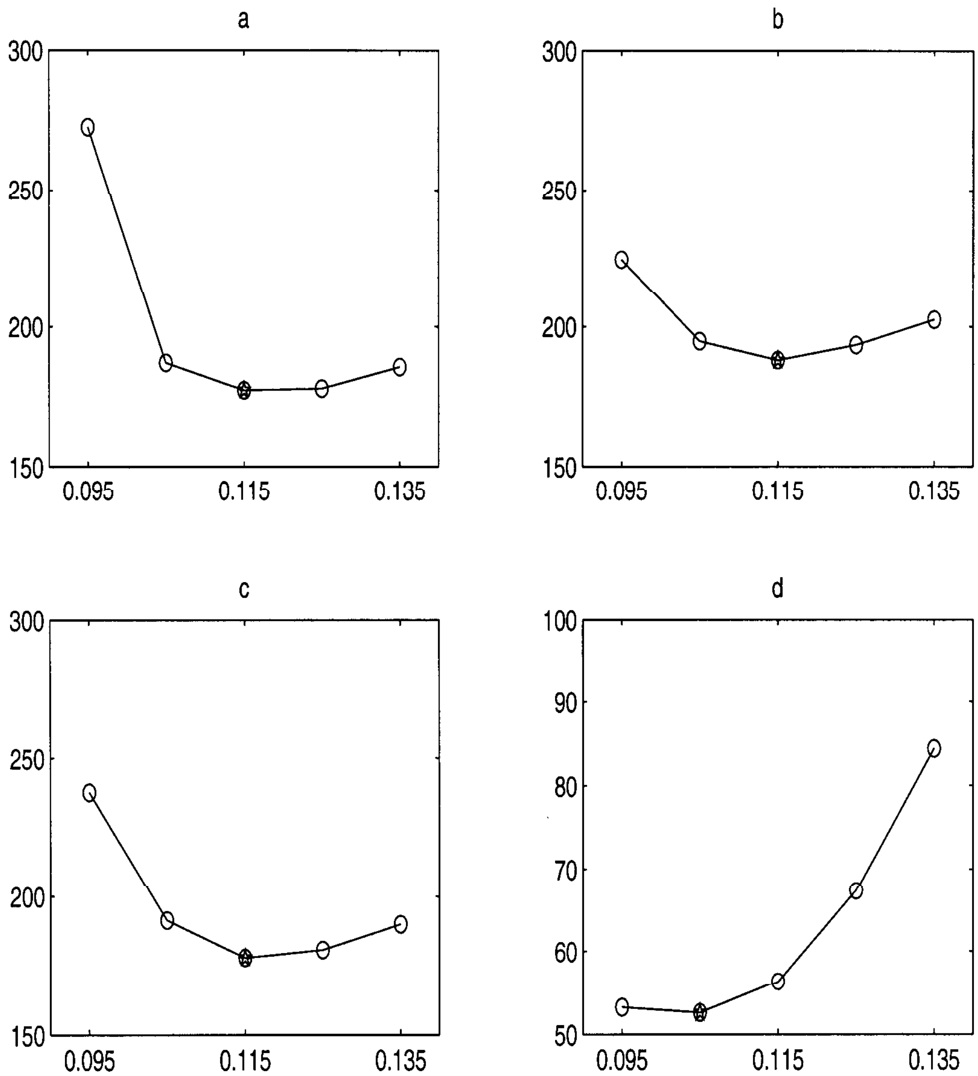


Figure 2: MLE. (a) WRF, (b) EnKF, (c) MEF, (d) MFF

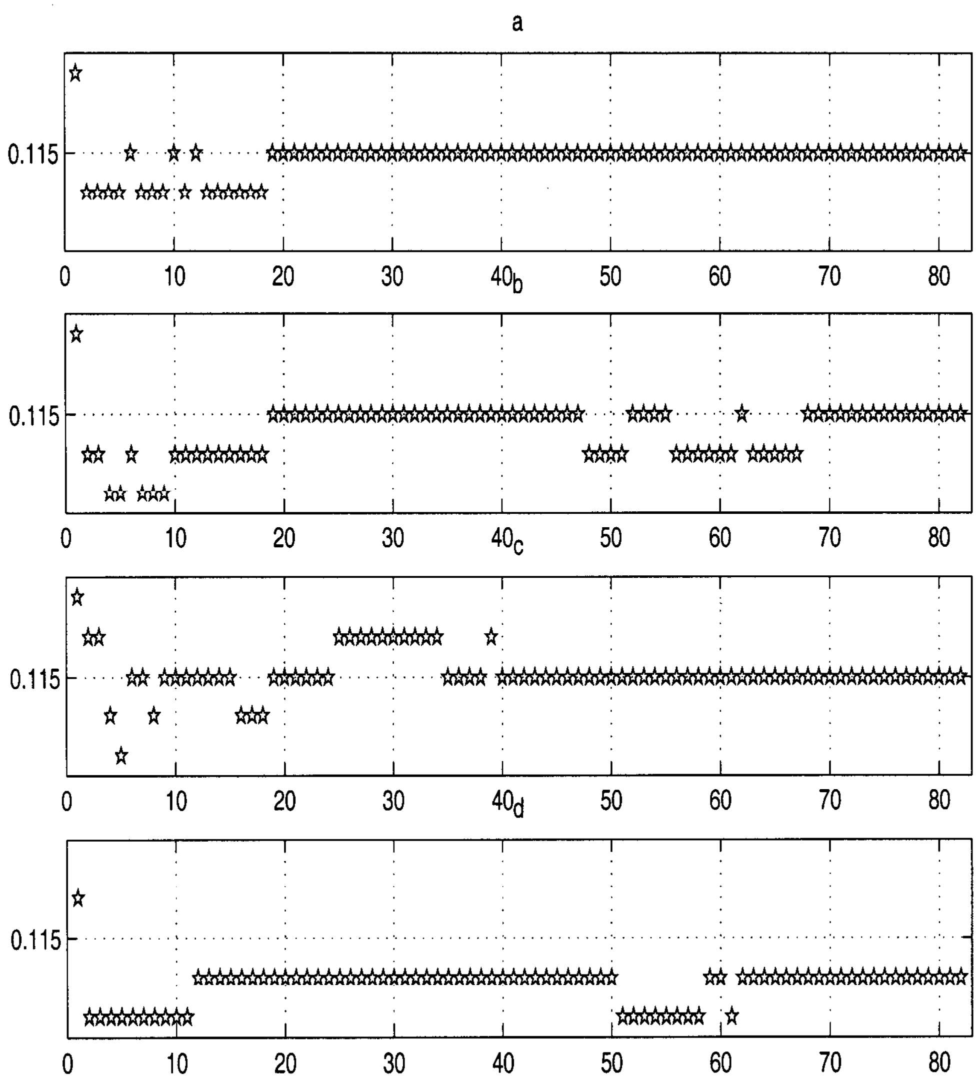


Figure 3: Change of MLE. (a) WRF (b) EnKF (c) MEF (d) MFF.

Conclusions

We have developed a maximum-entropy method for particle filtering, or *Maximum-Entropy Filter*. When prior distributions are represented by Gaussian mixture models, this method generalizes the Ensemble Kalman Filter to better handle non-normal statistics. The method gives excellent results in a test problem with highly non-Gaussian distributions with as few as $N = 100$ samples. This method is very economical when $p \gg q$ and q is not too large.

When also $q \gg 1$, then a practical alternative uses a mean-field conditioning rather than a full Bayes update, the *Mean-Field Filter*. This method is much cheaper than MEF, but, like MEF, is well-converged with as few as $N = 100$ samples. MFF gives good results for the filter means, but less good results for variances and entropy.