Data analysis in the low noise regime

Jonathan Weare

University of Chicago

February 19, 2013

Example: the Kuroshio current

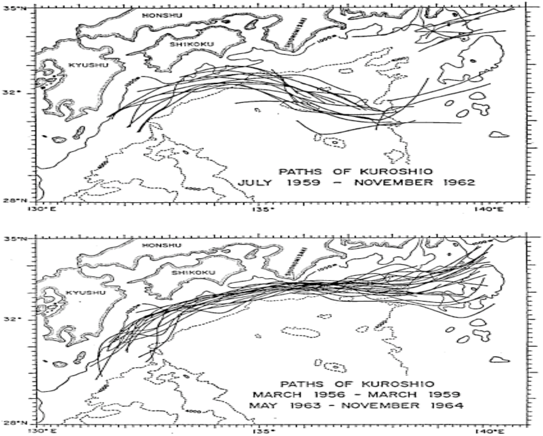


Figure : **Top:** Mean flow paths in the large meander state. **Bottom:** Mean flow paths in the small meander state

In the case of the meander transition of the Kuroshio we might like to know:

- How does this event occur, i.e. what rearrangements have to happen to trigger the event?
- How does the frequency or severity of the event depend on various environmental parameters?
- Can we predict the event from data in real-time?

Similar questions are relevant for **chemical reactions** or the **failure of a reliable electronic device** or **dramatic swings in a stock price**.

Answering any of these questions requires not only that we can simulate the underlying system but that we can simulate the rare event itself (many times).

When numerical stability and accuracy require you to simulate on a timescale orders of magnitude smaller than timescale of the event of interest (5–10 years in the case of the Kuroshio) this is a problem.

Data assimilation

What happens to the data assimilation problem when the underlying system undergoes rare dramatic fluctuations like the Kuroshio's meander transitions?

We'll model the underlying system by a small noise stochastic differential equation

$$dX^{\epsilon}(t) = b(X^{\epsilon}(t))\,dt + \sqrt{\epsilon}\,\sigma(X^{\epsilon}(t))\,dW(t)$$

and, at a sequence of discrete times $t_1, t_2, t_3, \ldots,$ record observations of the form (for example)

$$Y^{\epsilon}(t_n) = X^{\epsilon}(t_n) + \sqrt{\epsilon}\,\xi_n,$$

where the ξ_n are (for example) i.i.d. standard Gaussian.

We would like to generate sample trajectories of the system given a particular sequence of observations.

The parameter ϵ appears in both the dynamics and the observation model. We trust the dynamics and the observations roughly equally. What do we do when they disagree?

In other words what happens when the observation at time t_{i+1} lies in the tail of the distribution of our prediction of the state at time t_{i+1}?

By taking $\epsilon \to 0$ we can identify the most important consequences of this point of view and address them.

The two step recursive filtering procedure:

1. Starting from an ensemble of copies of $X^{\epsilon}(t_i)$ evolve each copy forward to the next observation time (t_{i+1}) and compute the contribution to the weight from the next observation

$$W_j^{\epsilon}(t_{i+1}) = \exp\left(-\frac{1}{\epsilon}\|y(t_{i+1}) - X_j^{\epsilon}(t_{i+1})\|^2\right)$$

2. Resample the copies of $X^{\epsilon}(t_{i+1})$ according to the weights, i.e. duplicate copies with large weights and eliminate copies with low weight.
3. Repeat for next observation.

Various approximations are needed to make this scheme practical on large problems (e.g. the ensemble Kalman filter).

The two step recursive filtering procedure:

1. Starting from an ensemble of copies of $X^{\epsilon}(t_i)$ evolve each copy forward to the next observation time (t_{i+1}) and compute the contribution to the weight from the next observation

$$W_j^{\epsilon}(t_{i+1}) = \exp\left(-\frac{1}{\epsilon}\|y(t_{i+1}) - X_j^{\epsilon}(t_{i+1})\|^2\right)$$

2. Resample the copies of $X^{\epsilon}(t_{i+1})$ according to the weights, i.e. duplicate copies with large weights and eliminate copies with low weight.
3. Repeat for next observation.

Various approximations are needed to make this scheme practical on large problems (e.g. the ensemble Kalman filter).

The two step recursive filtering procedure:

1. Starting from an ensemble of copies of $X^\epsilon(t_i)$ evolve each copy forward to the next observation time (t_{i+1}) and compute the contribution to the weight from the next observation

$$W_j^\epsilon(t_{i+1}) = \exp\left(-\frac{1}{\epsilon}\|y(t_{i+1}) - X_j^\epsilon(t_{i+1})\|^2\right)$$

2. Resample the copies of $X^\epsilon(t_{i+1})$ according to the weights, i.e. duplicate copies with large weights and eliminate copies with low weight.
3. Repeat for next observation.

Various approximations are needed to make this scheme practical on large problems (e.g. the ensemble Kalman filter).

The two step recursive filtering procedure:

1. Starting from an ensemble of copies of $X^{\epsilon}(t_i)$ evolve each copy forward to the next observation time (t_{i+1}) and compute the contribution to the weight from the next observation

$$W_j^{\epsilon}(t_{i+1}) = \exp\left(-\frac{1}{\epsilon}\|y(t_{i+1}) - X_j^{\epsilon}(t_{i+1})\|^2\right)$$

2. Resample the copies of $X^{\epsilon}(t_{i+1})$ according to the weights, i.e. duplicate copies with large weights and eliminate copies with low weight.
3. Repeat for next observation.

Various approximations are needed to make this scheme practical on large problems (e.g. the ensemble Kalman filter).

Now imagine that you apply this scheme to the Kuroshio:

Suppose at observation time t_i all of your copies of the system are in states consistent with the large meander.

Suppose that between t_i and t_{i+1} the true state of the Kuroshio transitions from the large meander to the small meander. The observation generated at time t_{i+1} will probably be consistent with the small meander.

However, at best only a very small portion of your copies will make the transition to the small meander between times t_i and t_{i+1}.

The few copies that make the transition will have relatively huge weights and you have no or very little resolution in the region that is suddenly important.

One can measure of the statistical quality of the weighted ensemble generated by our filter by:

$$R = \frac{\mathbf{E}\left[(W_j^\epsilon(t_{i+1}))^2\right]}{\mathbf{E}\left[W_j^\epsilon(t_{i+1})\right]^2}$$

Roughly an ensemble of N weighted samples has the quality of N/R independent unweighted samples from the target distribution.

More generally this is a very severe (stronger than total variation) measure of the difference between forecast distribution and posterior.

$R \geq 1$ and we'd like it to be as close to 1 as possible.

Considering only one observation window and treating the observation as fixed, note the the weights are of the form

$$W_j^\epsilon(t_{i+1}) = e^{-\frac{1}{\epsilon} g(X^\epsilon(t_{i+1}))}$$

The **Laplace Principle** for X^ϵ gives us constants γ_1 and γ_2 such that

$$\mathbf{E}\left[e^{-\frac{1}{\epsilon} g(X^\epsilon)}\right] = e^{\frac{-\gamma_1+o(1)}{\epsilon}} \quad \text{and} \quad \mathbf{E}\left[e^{-\frac{1}{\epsilon} 2g(X^\epsilon)}\right] = e^{\frac{-\gamma_2+o(1)}{\epsilon}}$$

Therefore

$$R = \exp\left(\frac{\gamma_2 - 2\gamma_1 + o(1)}{\epsilon}\right)$$

Since $\gamma_2 \leq 2\gamma_1$ this is very bad news. We'll need exponentially many samples.

More precisely

$$R = \exp\left(\frac{\gamma_2 - 2\gamma_1 + o(1)}{\epsilon}\right)$$

where, for a fixed position x of the system at time t_i,

$$\gamma_1 = \inf_{\substack{\varphi, \\ \varphi(t_i)=x}} \left\{ \int_{t_i}^{t_{i+1}} \frac{1}{2} \|\sigma^{-1}(\dot{\varphi} - b)\|^2 \, ds + g\left(\varphi(t_{i+1})\right) \right\}, \qquad (1)$$

and

$$\gamma_2 = \inf_{\substack{\varphi, \\ \varphi(t_i)=x}} \left\{ \int_{t_i}^{t_{i+1}} \frac{1}{2} \|\sigma^{-1}(\dot{\varphi} - b)\|^2 \, ds + 2g\left(\varphi(t_{i+1})\right) \right\}, \qquad (2)$$

One possibility: Between observations we can try to "**pull**" each copy toward the region where the likelihood is relatively large.

Instead of sampling the solution, X^ϵ, of

$$dX^\epsilon(t) = b(X^\epsilon(t))\,dt + \sqrt{\epsilon}\,\sigma(X^\epsilon(t))\,dW(t)$$

sample the solution, $\hat{X}^\epsilon$, of

$$d\hat{X}^\epsilon(t) = \Big(b(\hat{X}^\epsilon(t)) + \sigma(\hat{X}^\epsilon(t))v(t,\hat{X}^\epsilon(t))\Big)\,dt + \sqrt{\epsilon}\,\sigma(\hat{X}^\epsilon(t))\,dW(t).$$

Where v will be chosen to direct our samples to regions where the likelihood is larger.

An old idea... and a disaster if not done very carefully.

One possibility: Between observations we can try to “**pull**” each copy toward the region where the likelihood is relatively large.

Instead of sampling the solution, X^{ϵ}, of

$$dX^{\epsilon}(t) = b(X^{\epsilon}(t))\,dt + \sqrt{\epsilon}\,\sigma(X^{\epsilon}(t))\,dW(t)$$

sample the solution, $\hat{X}^{\epsilon}$, of

$$d\hat{X}^{\epsilon}(t) = \Big(b(\hat{X}^{\epsilon}(t)) + \sigma(\hat{X}^{\epsilon}(t))v(t,\hat{X}^{\epsilon}(t))\Big)\,dt + \sqrt{\epsilon}\,\sigma(\hat{X}^{\epsilon}(t))\,dW(t).$$

Where v will be chosen to direct our samples to regions where the likelihood is larger.

An old idea... and a disaster if not done very carefully.

To correct for the bias we'll have to multiply our weights by

$$Z^{\epsilon}(t_{i+1}) = \exp\left(-\frac{1}{\sqrt{\epsilon}}\int_{t_i}^{t_{i+1}} v(t,\hat{X}_j^{\epsilon}(t))\,dW_j(t) - \frac{1}{2\epsilon}\int_{t_i}^{t_{i+1}} v(t,\hat{X}_j^{\epsilon}(t))^2\,dt\right).$$

at each observation.

So now we also have to worry about the behavior of these new weights.

The relative variation of the weights is measured by

$$R = \frac{\mathbf{E}\left[\left(W_j^{\epsilon}(t_{i+1})Z^{\epsilon}(t_{i+1})\right)^2\right]}{\mathbf{E}\left[W_j^{\epsilon}(t_{i+1})\right]^2}$$

Unless we are very careful this number will grow exponentially as $\epsilon \to 0$.

Large deviations tells us that, for small ϵ, a transition like the Kuroshio's occurs along a predictable pathway. That path is the minimizer of the cost functional

$$\inf_{\substack{\varphi \in \mathcal{AC}([t_i,t_{i+1}]),\\ \varphi(t_i)=x}} \left\{ \int_{t_i}^{t_{i+1}} \frac{1}{2} \|\sigma^{-1}(\dot{\varphi} - b)\|^2 \, ds \right\}.$$

where the minimization is restricted to paths undergoing the transition.

It also tells us how a particular observation $y(t_{i+1})$ is generated in the low noise regime:

$$\inf_{\substack{\varphi \in \mathcal{AC}([t_i,t_{i+1}]),\\ \varphi(t_i)=x}} \left\{ \int_{t_i}^{t_{i+1}} \frac{1}{2} \|\sigma^{-1}(\dot{\varphi} - b)\|^2 \, ds + g(\varphi(t_{i+1})) \right\}.$$

Based on this it's tempting to choose

$$v(t,x) = \sigma^{-1}(x)(\dot{\hat{\varphi}}(t) - b(x))$$

and

$$d\hat{X}^{\epsilon}(t) = \dot{\hat{\varphi}}(t)\,dt + \sqrt{\epsilon}\,\sigma(\hat{X}^{\epsilon}(t))\,dW(t).$$

where $\hat{\varphi}$ minimizes

$$\inf_{\substack{\varphi\in\mathcal{AC}([t_i,t_{i+1}]),\\ \varphi(t_i)=x}} \left\{ \int_{t_i}^{t_{i+1}} \frac{1}{2}\|\sigma^{-1}(\dot{\varphi} - b)\|^2\,ds + g(\varphi(t_{i+1})) \right\}.$$

This sometimes works but on some problems it can be a disastrous choice.

But a related choice works extremely well.

Based on this it's tempting to choose

$$v(t,x) = \sigma^{-1}(x)(\dot{\hat{\varphi}}(t) - b(x))$$

and

$$d\hat{X}^{\epsilon}(t) = \dot{\hat{\varphi}}(t)\,dt + \sqrt{\epsilon}\,\sigma(\hat{X}^{\epsilon}(t))\,dW(t).$$

where $\hat{\varphi}$ minimizes

$$\inf_{\substack{\varphi\in\mathcal{AC}([t_i,t_{i+1}]),\\ \varphi(t_i)=x}} \left\{ \int_{t_i}^{t_{i+1}} \frac{1}{2}\|\sigma^{-1}(\dot{\varphi} - b)\|^2\,ds + g(\varphi(t_{i+1})) \right\}.$$

This sometimes works but on some problems it can be a disastrous choice.

But a related choice works extremely well.

Focusing again on a single observation interval, consider the function

$$\Phi^{\epsilon}(t,x) = \mathbf{E}_{t,x}\left[e^{-\frac{1}{\epsilon}g(X^{\epsilon}(t_{i+1}))}\right].$$

Assuming $X^{\epsilon}(t_i) = x$, it's not hard to see that if we choose

$$v^{\epsilon} = -\epsilon\frac{\sigma^{T}\Phi^{\epsilon}_{x}}{\Phi^{\epsilon}}$$

we get $R^{\epsilon} = 1$. In other words this choice of v is the best possible.

Φ^{ϵ} solve a linear second order parabolic PDE with terminal condition $\Phi^{\epsilon}(t_{i+1},x) = \mathrm{e}^{-\frac{1}{\epsilon}g(x)}$.

Of course there's no hope of finding a global solution of the PDE in more than a few dimensions.

Focusing again on a single observation interval, consider the function

$$\Phi^\epsilon(t,x) = \mathbf{E}_{t,x}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon(t_{i+1}))}\right].$$

Assuming $X^\epsilon(t_i) = x$, it's not hard to see that if we choose

$$v^\epsilon = -\epsilon\frac{\sigma^T\Phi^\epsilon_x}{\Phi^\epsilon}$$

we get $R^\epsilon = 1$. In other words this choice of v is the best possible.

Φ^ϵ solve a linear second order parabolic PDE with terminal condition $\Phi^\epsilon(t_{i+1},x) = \mathrm{e}^{-\frac{1}{\epsilon}g(x)}$.

Of course there's no hope of finding a global solution of the PDE in more than a few dimensions.

Instead we'll consider the $\epsilon \to 0$ limit of the log transform of Φ^ϵ,

$$G^\epsilon = -\epsilon \log \Phi^\epsilon$$

which solves the second order Hamilton-Jacobi Equation

$$-G^\epsilon_t - bG^\epsilon_x + \frac{1}{2}\left(\sigma^T G^\epsilon_x\right)^2 - \frac{\epsilon}{2}\sigma\sigma^T G^\epsilon_{xx} = 0, \qquad G^\epsilon(t_{i+1}, x) = g(x) \tag{3}$$

In terms of G^ϵ

$$v^\epsilon = -\sigma^T G^\epsilon_x.$$

So we can set

$$v^0 = -\sigma^T G_x$$

where G is the viscosity solution of

$$-G_t - bG_x + \frac{1}{2}\left(\sigma^T G_x\right)^2 = 0, \qquad G(t_{i+1}, x) = g(x)$$

and hope for the best.

Instead we'll consider the $\epsilon \to 0$ limit of the log transform of Φ^ϵ,

$$G^\epsilon = -\epsilon \log \Phi^\epsilon$$

which solves the second order Hamilton-Jacobi Equation

$$-G_t^\epsilon - bG_x^\epsilon + \frac{1}{2}\left(\sigma^T G_x^\epsilon\right)^2 - \frac{\epsilon}{2}\sigma\sigma^T G_{xx}^\epsilon = 0, \qquad G^\epsilon(t_{i+1}, x) = g(x) \tag{3}$$

In terms of G^ϵ

$$v^\epsilon = -\sigma^T G_x^\epsilon.$$

So we can set

$$v^0 = -\sigma^T G_x$$

where G is the viscosity solution of

$$-G_t - bG_x + \frac{1}{2}\left(\sigma^T G_x\right)^2 = 0, \qquad G(t_{i+1}, x) = g(x)$$

and hope for the best.

G has the control representation

$$G(t,x) = \inf_{\substack{\varphi\in\mathcal{AC}([t,T]),\\ \varphi(t)=x}} \left\{ \int_t^T \frac{1}{2}\|\sigma^{-1}(\dot{\varphi} - b)\|^2\, ds + g\left(\varphi(T)\right) \right\}.$$

Notice that $\gamma_1 = G(0, x_0)$.

G is the rate appearing in the Laplace Principle.

Furthermore, where G is differentiable,

$$b(t,x) + \sigma(t,x)\, v^0(t,x) = \dot{\hat{\varphi}}_{t,x}(t)$$

where

$$\hat{\varphi}_{t,x} = \arg \min_{\substack{\varphi\in\mathcal{AC}([t,T]),\\ \varphi(t)=x}} \left\{ \int_t^T \frac{1}{2}\|\sigma^{-1}(\dot{\varphi} - b)\|^2\, ds + g\left(\varphi(T)\right) \right\}.$$

For v^0 we have that

$$d\hat{X}^\epsilon(t) = \dot{\hat{\varphi}}_{t,\hat{X}^\epsilon(t)}\,dt + \sqrt{\epsilon}\,\sigma(\hat{X}^\epsilon(t))\,dW(t).$$

Roughly this says take a step in the most likely direction, compute the new most likely direction from your new position, take another step.

This procedure can be carried out at reasonable cost and, as we prove, the estimator has very favorable error properties.

Theoretical points:

A result in

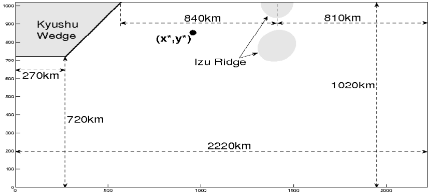
> Rare event simulation for small noise diffusions (with E. Vanden-Eijnden), CPAM (2012)

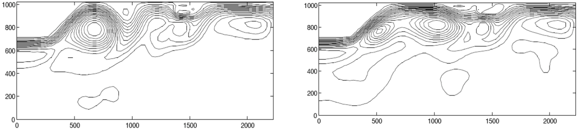
implies that for this method (for one observation step) $R \to 1$ as $\epsilon \to 0$.

So vanishing statistical error in contrast to the exponentially growing error before.

That paper also address the errors introduced by numerical discretization.

Application to a simple Kuroshio model:





Data assimilation in the small noise regime and application to the Kuroshio (with E. Vanden-Eijnden), *Monthly Weather Review* (2012)

$$\partial_t X + \frac{\partial}{\partial x}(uX) + \frac{\partial}{\partial y}(vX) + f\left(\frac{f_x}{f} - \frac{r_x}{r}\right)u \\ + f\left(\frac{f_y}{f} - \frac{r_y}{r}\right)v = \nu \Delta X + \sigma \eta$$

$$X = \frac{\partial}{\partial x}\left(\frac{1}{r}\psi_x\right) + \frac{\partial}{\partial y}\left(\frac{1}{r}\psi_y\right)$$

X is the vorticity and ψ is the volume transport streamfunction.

40 observations of $\psi(x^*, y^*)$ are taken every 2.63 days (with mean 0, standard deviation 0.1, Gaussian errors).

The observations are taken from a segment of a long run undergoing a transition from the small meander to the large meander.

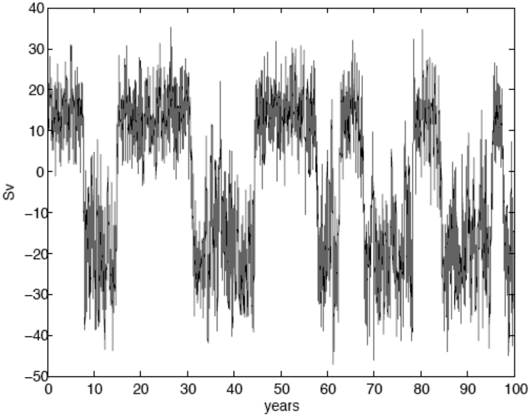
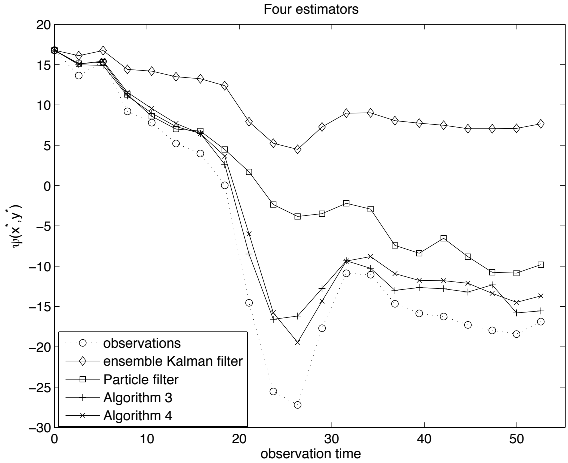
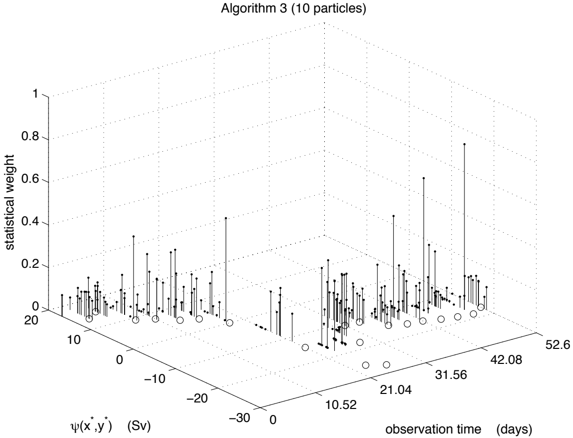
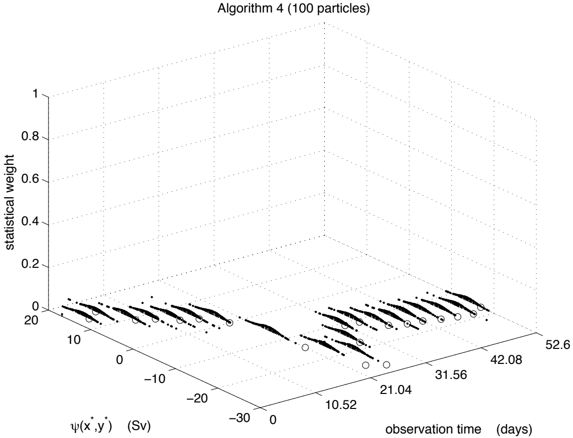


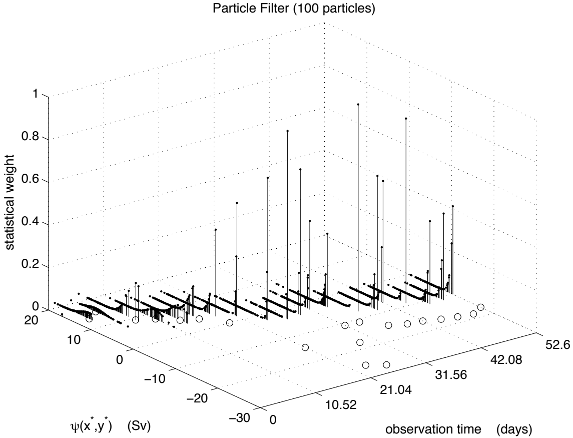
Figure : Projected view of transitions.

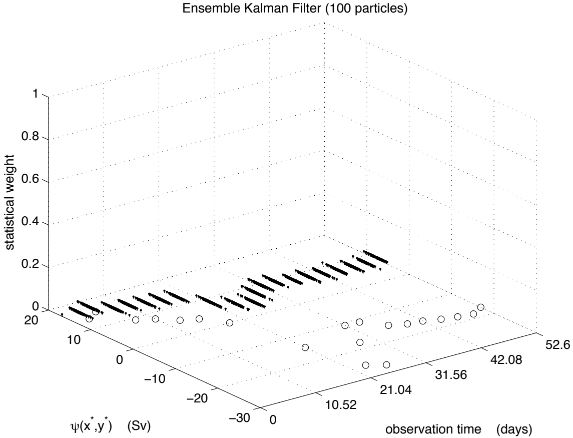


The particle filter, the ensemble Kalman filter, and Algorithm 4 are all run with 100 particles while Algorithm 3 is run with 10 particles.


Algorithm 3 (10 particles)
statistical weight
1
0.8
0.6
0.4
0.2
0
20
10
0
−10
−20
−30
$\psi(x^*,y^*)$ (Sv)
0
10.52
21.04
31.56
42.08
52.6
observation time (days)


Algorithm 4 (100 particles)
statistical weight
1
0.8
0.6
0.4
0.2
0
20
10
0
−10
−20
−30
ψ(x*,y*) (Sv)
0
10.52
21.04
31.56
42.08
52.6
observation time (days)


Particle Filter (100 particles)
statistical weight
1
0.8
0.6
0.4
0.2
0
20
10
0
−10
−20
−30
ψ(x*,y*) (Sv)
0
10.52
21.04
31.56
42.08
52.6
observation time (days)


Ensemble Kalman Filter (100 particles)
statistical weight
1
0.8
0.6
0.4
0.2
0
20
10
0
−10
−20
−30
ψ(x*,y*)   (Sv)
0
10.52
21.04
31.56
42.08
52.6
observation time   (days)