Nonlin. Processes Geophys., 22, 205–213, 2015
www.nonlin-processes-geophys.net/22/205/2015/
doi:10.5194/npg-22-205-2015

Nonlinear Processes in Geophysics
Open Access

Improved variational methods in statistical data assimilation

J. Ye[1], N. Kadakia[1], P. J. Rozdeba[1], H. D. I. Abarbanel[1,2], and J. C. Quinn[3]

[1]Department of Physics, University of California, San Diego, La Jolla, CA 92093-0374, USA
[2]Marine Physical Laboratory (Scripps Institution of Oceanography), University of California, San Diego, La Jolla, CA 92093-0374, USA
[3]Intellisis Corporation, 10350 Science Center Drive, Suite 140, San Diego, CA 92121, USA

Correspondence to: H. D. I. Abarbanel (habarbanel@ucsd.edu)

Received: 18 August 2014 – Published in Nonlin. Processes Geophys. Discuss.: 10 October 2014
Revised: 3 March 2015 – Accepted: 23 March 2015 – Published: 7 April 2015

Abstract. Data assimilation transfers information from an observed system to a physically based model system with state variables $\boldsymbol{x}(t)$. The observations are typically noisy, the model has errors, and the initial state $\boldsymbol{x}(t_0)$ is uncertain: the data assimilation is statistical. One can ask about expected values of functions $\langle G(\boldsymbol{X})\rangle$ on the path $\boldsymbol{X} = \{\boldsymbol{x}(t_0), \ldots, \boldsymbol{x}(t_m)\}$ of the model state through the observation window $t_n = \{t_0, \ldots, t_m\}$. The conditional (on the measurements) probability distribution $P(\boldsymbol{X}) = \exp[-A_0(\boldsymbol{X})]$ determines these expected values. Variational methods using saddle points of the "action" $A_0(\boldsymbol{X})$, known as 4DVar (Talagrand and Courtier, 1987; Evensen, 2009), are utilized for estimating $\langle G(\boldsymbol{X})\rangle$. In a path integral formulation of statistical data assimilation, we consider variational approximations in a realization of the action where measurement errors and model errors are Gaussian. We (a) discuss an annealing method for locating the path $\boldsymbol{X}^0$ giving a consistent minimum of the action $A_0(\boldsymbol{X}^0)$, (b) consider the explicit role of the number of measurements at each t_n in determining $A_0(\boldsymbol{X}^0)$, and (c) identify a parameter regime for the scale of model errors, which allows $\boldsymbol{X}^0$ to give a precise estimate of $\langle G(\boldsymbol{X}^0)\rangle$ with computable, small higher-order corrections.

1 Introduction

In a broad spectrum of scientific fields, transferring the information contained in L observed data time series $y_l(t_n)$; $l = 1, \ldots, L$; $n = 0, \ldots, m$ to a physically based model of the processes producing those observations allows the estimation of unknown parameters and unobserved states of the model within an observation window $\{t_0, \ldots, t_m\}$. As a sample of these fields, we note applications in meteorology (Talagrand and Courtier, 1987; Evensen, 2009; Lorenc and Payne, 2007), geochemistry (Eibern and Schmidt, 1999), fluid dynamics (Zadeh, 2008) and plasma physics (Mechhoud et al. , 2013), among many others.

The conditional probability distribution $P(\boldsymbol{X}) = \exp[-A_0(\boldsymbol{X})]$ allows evaluation of expected values of physically interesting functions $G(\boldsymbol{X})$ along the path. $P(\boldsymbol{X})$ is conditioned on the measurements $y_l(t_n)$; $l = 1, \ldots, L$; $n = 0, \ldots, m$. These are L-dimensional, while the model state is D-dimensional; it is usually the case that $D \gg L$. Data assimilation seeks to use the information in $\boldsymbol{y}(t_n)$ to estimate unknown parameters in the model and unobserved states of the model. If this is accomplished, one uses prediction for $t > t_m$ to examine the validity of the model.

The action $A_0(\boldsymbol{X})$ contains terms giving a measurement's influence on $P(\boldsymbol{X})$, terms propagating the model between the measurement times t_n, and a term $-\log[P(\boldsymbol{x}(0))]$ representing the uncertainty in the initial state (Abarbanel, 2013). We discuss the familiar case where additive measurement errors are independent at each t_n and Gaussian with covariance $R_m^{-1}(l, l', t)$; $l, l' = 1, \ldots, L$. The model is a physical differential equation, discretized in space and time and satisfying the D-dimensional stochastic discrete time map $x_a(n+1) = f_a(\boldsymbol{x}(n)) + R_f^{-1/2}(a, b)\eta_b(n)$; $a, b = 1, \ldots, D$ with iid Gaussian noise error $\eta_a(n) \sim \mathcal{N}(0, 1)$. We take $R_m(l, l', n) = R_m(n)\,\delta_{l,l'}$ and $R_f(a, b) = R_f\,\delta_{a,b}$. $R_m(n)$ is zero except near observation times t_n.

With these conditions, the action takes a familiar form:

$$\begin{aligned} A_0(\boldsymbol{X}) &= \sum_{n=0}^{m} \frac{R_m(n)}{2} \sum_{l=1}^{L} \left[x_l(n) - y_l(n) \right]^2 \\ &+ \frac{R_f}{2} \sum_{n=0}^{m-1} \sum_{a=1}^{D} \left[x_a(n+1) - f_a(\boldsymbol{x}(n)) \right]^2 \\ &- \log[P(\boldsymbol{x}(0))]. \end{aligned} \tag{1}$$

The conditional expected value $\langle G(\boldsymbol{X}) \rangle$ of a function $G(\boldsymbol{X})$ on the path $\boldsymbol{X} = \{\boldsymbol{x}(t_0), \ldots, \boldsymbol{x}(t_m)\}$ is given as

$$\langle G(\boldsymbol{X}) \rangle = \frac{\int \mathrm{d}\boldsymbol{X} G(\boldsymbol{X}) \exp[-A_0(\boldsymbol{X})]}{\int \mathrm{d}\boldsymbol{X} \exp[-A_0(\boldsymbol{X})]}. \tag{2}$$

One interesting function $G(\boldsymbol{X})$ is $\boldsymbol{X}$ itself, whose expected value gives us the average path over the measurement window $[t_0, t_m]$. Estimates of the parameters and $P(\boldsymbol{x}(t_m))$ permit prediction for $t > t_m$; in this window, one compares observations and predictions using model output with a user-selected metric. No information is passed to the model in the prediction window. Importantly, good *estimation* of observed state variables (a "good fit") is not sufficient to produce confidence in the quality of the model; good *prediction* is critical.

To approximate the integral $\langle G(\boldsymbol{X}) \rangle$, we follow the stationary path method of Laplace (Laplace, 1774) and seek saddle points in path space $\boldsymbol{X}^q$ labeled by $q = 0, 1 \ldots$ satisfying $\partial A_0(\boldsymbol{X}) / \partial \boldsymbol{X}_\alpha = 0$. The index α is a label for the state and time $\alpha = \{a, n\}$. To determine their importance in evaluating the integral, Eq. (2) paths are sorted by increasing action levels $A_0(\boldsymbol{X}^q)$: $A_0(\boldsymbol{X}^0) \leq A_0(\boldsymbol{X}^1) \leq \cdots$.

2 Evaluating $\langle G(\boldsymbol{X}) \rangle$

We take the the usual data assimilation technique (Talagrand and Courtier, 1987; Evensen, 2009; Lorenc and Payne, 2007) several steps further by

1. showing how to find a consistent path $\boldsymbol{X}^0$ for the minimum action level using an annealing method,
2. demonstrating the importance of the number of measurements L at each observation time t_n, and
3. explaining how to make systematic perturbation corrections to $\langle G(\boldsymbol{X}^0) \rangle$, i.e., $G(\boldsymbol{X}^0)$ evaluated on the minimum action level path.

For nonlinear problems of interest, there may be many paths $\boldsymbol{X}^q$; $q = 0, 1, \ldots$ satisfying the saddle point condition. To assess their contributions to $\langle G(\boldsymbol{X}) \rangle$, we expand $A_0(\boldsymbol{X})$ in the neighborhood of each $\boldsymbol{X}^q$ as (note that all variables are notated in Table 2)

$$\begin{aligned} A_0(\boldsymbol{X}) &= A_0\left(\boldsymbol{X}^q\right) + \left(X - X^q\right)_{\alpha_1} \gamma^2_{\alpha_1 \alpha_2} \left(\boldsymbol{X}^q\right) \left(X - X^q\right)_{\alpha_2} \\ &+ \sum_{r=3} \frac{A^{(r)}(\boldsymbol{X}^q)_{\alpha_1 \ldots \alpha_r}}{r!} \left(X - X^q\right)_{\alpha_1} \cdots \left(X - X^q\right)_{\alpha_r}. \end{aligned} \tag{3}$$

There is an implied sum over all paired Greek indices α_j. A sum over all terms with comparable action level $A_0(\boldsymbol{X}^q)$ then gives an approximation to $\langle G(\boldsymbol{X}) \rangle$.

Changing variables to $U_\alpha = \gamma_{\alpha\beta}(\boldsymbol{X}^q)\, (X - X^q)_\beta$ leads to a contribution to the numerator of $\langle G(\boldsymbol{X}) \rangle$ in Eq. (2) from each $\boldsymbol{X}^q$

$$\frac{\exp\left[-A_0\left(\boldsymbol{X}^q\right)\right]}{\det\gamma\left(\boldsymbol{X}^q\right)} \int \mathrm{d}\mathbf{U} \exp\left(-\mathbf{U}^2 - V\left(\boldsymbol{X}^q\right)\right) \left[G\left(\boldsymbol{X}^q\right) + W\left(\boldsymbol{X}^q\right)\right] \tag{4}$$

where

$$V\left(\boldsymbol{X}^q\right) = \sum_{r=3} \frac{A^{(r)}(\boldsymbol{X}^q)_{\alpha_1 \ldots \alpha_r}}{r!} \left(\gamma\left(\boldsymbol{X}^q\right)^{-1} U\right)_{\alpha_1} \cdots \left(\gamma\left(\boldsymbol{X}^q\right)^{-1} U\right)_{\alpha_r},$$

and

$$W\left(\boldsymbol{X}^q\right) = \sum_{k=1} \frac{G^{(k)}(\boldsymbol{X}^q)_{\alpha_1 \ldots \alpha_k}}{k!} \left(\gamma\left(\boldsymbol{X}^q\right)^{-1} U\right)_{\alpha_1} \cdots \left(\gamma\left(\boldsymbol{X}^q\right)^{-1} U\right)_{\alpha_k}.$$

The contributions to the denominator of Eq. (2) are identical to Eq. (4), with the factor $[G(\boldsymbol{X}^q) + W(\boldsymbol{X}^q)]$ replaced by unity. We sum over the contribution of each $\boldsymbol{X}^q$ to evaluate $\langle G(\boldsymbol{X}) \rangle$.

If the lowest action level $A_0(\boldsymbol{X}^0)$ is much smaller than all others, then $\exp[-A_0(\boldsymbol{X}^0)] \gg \exp[-A_0(\boldsymbol{X}^{q \neq 0})]$ and its contribution to $\langle G(\boldsymbol{X}) \rangle$ totally dominates the integral. We have then that

$$\langle G(\boldsymbol{X}) \rangle = \frac{\int \mathrm{d}\mathbf{U} \exp\left(-\mathbf{U}^2 - V\left(\boldsymbol{X}^0\right)\right) \left[G\left(\boldsymbol{X}^0\right) + W\left(\boldsymbol{X}^0\right)\right]}{\int \mathrm{d}\mathbf{U} \exp\left(-\mathbf{U}^2 - V\left(\boldsymbol{X}^0\right)\right]}, \tag{5}$$

plus exponentially small corrections from the action levels associated with other saddle paths $\boldsymbol{X}^{q \neq 0}$.

3 Annealing to find a consistent minimum action level $\boldsymbol{A_0(X^0)}$

We now turn to an annealing method to find the path $\boldsymbol{X}^0$ where $\exp[-A_0(\boldsymbol{X}^0)] \gg \exp[-A_0(\boldsymbol{X}^{q \neq 0})]$. Within this method, we first examine the importance of the number L of measurements at each observation time t_n. We then present an argument that, in the integral Eq. (4), contributions to $\langle G(\boldsymbol{X}) \rangle$ from the terms $V(\boldsymbol{X}^q)$, $W(\boldsymbol{X}^q)$ behave as inverse powers of R_f as R_f / R_m becomes large. This would leave us with $\langle G(\boldsymbol{X})\rangle \approx G(\boldsymbol{X}^0)$ with corrections that are a power series in R_f^{-1}. The implication is that all statistical data assimilation questions, as R_f/R_m becomes large, can be well approximated by the contribution of the path $\boldsymbol{X}^0$ with the lowest action level $A_0(\boldsymbol{X}^0)$ along with corrections one can evaluate via standard perturbation theory. Variations about $\langle G(\boldsymbol{X}^0)\rangle$ would be small and computable.

The term in the action expressing uncertainty in the initial model state $\boldsymbol{x}(0)$, $P(\boldsymbol{x}(0))$, is often written assuming Gaussian variation about some base state $\boldsymbol{x}_{\text{base}}$, so $-\log[P(\boldsymbol{x}(0)] \propto (\boldsymbol{x}(0)-\boldsymbol{x}_{\text{base}})^2\,\boldsymbol{R}_{\text{base}}/2$, and this has the form of the measurement term evaluated at $n=0$. We incorporate this expression into the term with coefficient R_m in the action and no longer display it. This term, often called a prior, is also seen at the initial condition for the time evolution of the conditional probability distribution $P(\boldsymbol{X})$.

3.1 Annealing details

The annealing method starts with the observation (Quinn, 2010) that the equation for the saddle points $\boldsymbol{X}^q$ simplifies at $R_f=0$. The action at $R_f=0$ has no information about the model, and relies solely on the measurements. The saddle point solution is $x_l(n)=y_l(n); l=1,2,\ldots,L$ for all observation times. The other $(D-L)$ components of the model state vector are undetermined, and the solution is quite degenerate. As we increase R_f, the action levels split, and, depending on R_m, R_f, L and the precise form of the dynamical vector field $\boldsymbol{f}(\boldsymbol{x})$, there will be 1, 2, ... saddle points of $A_0(\boldsymbol{X})$. The saddle points of interest are local minima.

For any $R_f>0$, the search for saddle points of the action requires an unconstrained numerical optimization problem to be solved: minimize $A_0(\boldsymbol{X})$. This is 4DVar in its "weak" formulation (Talagrand and Courtier, 1987; Evensen, 2009). The methods we have utilized for performing this numerical optimization range from Newton and quasi-Newton methods (Press et al., 2012) to more sophisticated interior point methods (Waëchter , 2002). Each search for saddle paths requires an initial guess $\boldsymbol{X}^{(0)}$ that is then iterated via the algorithm to produce $\boldsymbol{X}^{(1)}$, then $\boldsymbol{X}^{(2)}$, and continuing on to produce a final path $\boldsymbol{X}^{(\text{final})}$ for *each* value of R_f.

We begin the annealing process by choosing the initial R_f, denoted R_{f0}, as very small, but nonzero; indeed, if R_{f0} was chosen to be vanishing, the set of optimal paths would be all paths whose measured components match the data, and whose unmeasured components are entirely unconstrained. Such a solution is infinitely degenerate and gives no more insight than the data themselves. Therefore, we begin the search at some $R_{f0}\ll 1$ with the first $L(m+1)$ components of $\boldsymbol{X}^{(0)}$ taken as the observations $\boldsymbol{y}(t_n)$; $n=0,\ 1,\ \ldots,\ m$. The remaining components of $\boldsymbol{X}^{(0)}$ are chosen from a uniform distribution reflecting the dynamical range of the unmeasured state variables.

Because the search algorithm is an iterative process with potentially many basins of attraction, it is not evident which minimum or saddle point we will hit in this initial optimization. Accordingly, we actually start with N_I copies of this procedure, making N_I independent, with initial choices of $\boldsymbol{X}^{(0;r)}$; $r=1,\ 2,\ \ldots,\ N_I$ at R_{f0}; in other words, we initiate many such annealing procedures in parallel. These N_I choices differ in the unmeasured components of the path, independently drawn from a uniform distribution over the range of each variable.

In the next step of the annealing process, the value of R_f is increased to $R_{f0}\,\xi$ where $\xi>1$. Using as our initial paths the final paths $\boldsymbol{X}^{(\text{final};r)}$ from the previous optimization with $R_f=R_{f0}$, we perform the numerical optimization procedure once again, independently for each of the N_I paths. This results in a new set of N_I saddle paths $\boldsymbol{X}^{(\text{final};r)}$; $r=1,2,\ldots,N_I$ at $R_f=R_{f0}\,\xi$.

This process is repeated as many times as desired, increasing R_f from one iteration to the next by a power of ξ. We use the N_I final paths from each value of R_f as the initial paths for the next value of R_f. The output is a set of N_I paths and action values for each value of $\beta=0,\ 1,\ \ldots,\ \beta_{\max}$, and the action level for each of the N_I final paths is plotted for each R_f. Specific examples of this will be presented below.

All of these optimizations, for each r and each β, is a 4DVar calculation for which we have used various algorithms (Press et al., 2012; Waëchter , 2002). In a sense, we can think of this as an ensemble version of 4DVar, with the important aspect that we perform N_I calculations of 4DVar at each β.

As R_f/R_m grows large, the errors in the model diminish relative to the measurement errors and impose high-precision $\boldsymbol{x}(n+1)\approx \boldsymbol{f}(\boldsymbol{x}(n))$ on the model dynamics. The noise in the measurement has been taken to be Gaussian $\mathcal{N}(0,\sigma^2)$, so the measurement error term in the action satisfies a χ^2 distribution (NIST Handbook of Statistical Methods , 2012) with mean $R_m\,\sigma^2(m+1)L/2$ and standard deviation $R_m\,\sigma^2\,\sqrt{(m\,+\,1)L/2}$. We note that this observation has been made by Bennett (2002) when the weak form of 4DVar, here using the action as the objective function, results in accurate representation of the dynamics.

Using properties of the χ^2 distribution, the action level for large R_f approximately approaches a lowest value

$$A_0\left(\boldsymbol{X}^0\right) \rightarrow \frac{R_m\sigma^2}{2}L(m+1)\left[1 \pm \frac{1}{\sqrt{(m+1)L/2}}\right]. \qquad (6)$$

σ^2 is the noise level in $y_l(n)$ and $(m+1)$ is the number of measurement times t_n. This provides a consistency condition on the identification of the path $\boldsymbol{X}^0$ by giving a consistent minimum action value $A_0(\boldsymbol{X}^0)$. If the action levels revealed by our annealing procedure do not give this result for $A_0(\boldsymbol{X}^0)$, it is a sign that the data are inconsistent with the model.

At the beginning of the annealing procedure when $R_f=R_{f0}$, there was a degeneracy in the action values. As R_f increases, this degeneracy is broken and the action levels split. If the action level $A_0(\boldsymbol{X}^0)$ is substantially smaller than the action level on the next saddle path $A_0(\boldsymbol{X}^0) \ll A_0(\boldsymbol{X}^1)$, all expected values $\langle G(\boldsymbol{X})\rangle$ are given by $G(\boldsymbol{X}^0)$, and corrections due to fluctuations about that path. The contribution to the expected values of the path integral for $\langle G(\boldsymbol{X})\rangle$ from $\boldsymbol{X}^1$ with the next action level is exponentially smaller, of order $\exp[-(A_0(\boldsymbol{X}^1) - A_0(\boldsymbol{X}^0))]$.

The annealing procedure discussed above is different from standard simulated annealing (Aguiar e Oliviera et al., 2012). We call this annealing because varying R_f is similar to varying a temperature in a statistical physics problem where R_f is inversely proportional to the temperature. At high temperatures, small R_f, the dynamics among the degrees of freedom $x_a(n)$ is essentially irrelevant, and we have a universal solution where the observed degrees of freedom match the observations. As the temperature is decreased, the dynamics plays a more and more significant role, and allowed paths "freeze" out. Action levels play the role of energy levels in statistical physics, and as the action is positive definite, the paths are directly analogous to instantons in the Euclidian field theory represented by $A_0(\boldsymbol{X})$ (Zinn-Justin, 2002).

The calculations used in the annealing process may seem formidable. We have chosen $N_I = 100$ in the examples we present now, because we wish to make clear the scope of the calculations and their implications. We chose this rather than concentrating now on optimizing the algorithms. This will come with some experience with different model structures. Nonetheless, due to the manner in which the action levels space themselves and split R_f is increased, it might prove less demanding to use a large N_I for the first few values of R_f, and then significantly decrease the number of paths from N_I. Our goal is to trace accurately just the lowest action value state $\boldsymbol{X}^0$ and the next $\boldsymbol{X}^1$ to estimate the scale of the dominance of $A_0(\boldsymbol{X}^0)$ in performing the path integral.

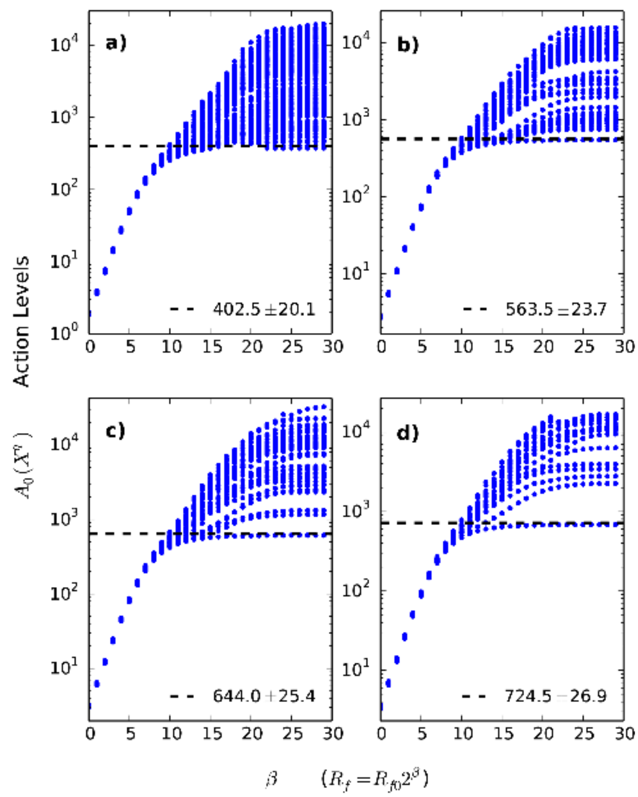

Figure 1. Action levels as a function of R_f for the Lorenz96 model, $D = 20$, $R_{f0} = 0.01$. **(a)** At $L = 5$, we used $y_1(t)$, $y_3(t)$, $y_5(t)$, $y_7(t)$, and $y_9(t)$ in the action; **(b)** at $L = 7$, $y_{11}(t)$ and $y_{13}(t)$ are added; **(c)** at $L = 8$, $y_{15}(t)$ is added; and **(d)** at $L = 9$, $y_{17}(t)$ is added. The expected values of the lowest action level are denoted by black dashed lines.

4 Examples from the Lorenz96 model

We now present the results of a set of calculations on the Lorenz96 (Lorenz, 2006) model used frequently in geophysical data assimilation discussions as a test bed for proposed methods. The model has D degrees of freedom $x_a(t)$ satisfying the differential equation

$$\dot{x}_a(t) = x_{a-1}(t)\,(x_{a+1}(t) - x_{a-2}(t)) - x_a(t) + f + \zeta_a(t); \quad (7)$$

$a = 1, \ldots, D$; $x_{-1}(t) = x_{D-1}(t)$, $x_0(t) = x_D(t)$, $x_{D+1}(t) = x_1(t)$; $\zeta_a(t)$ is $N(0, R_f^{-1})$ Gaussian noise. $f = 8.17$, for which the $x_a(t)$ are chaotic (Kostuk, 2012). We studied $D = 20$ and added f as an additional degree of freedom satisfying $\dot{f} = 0$. The number of model equations in the action is therefore equal to $D + 1$.

We performed a twin experiment in which we solved these equations, with $\zeta_a(t) = 0$ and with an arbitrary choice of initial conditions $\boldsymbol{x}(0)$ using a fourth-order Runge–Kutta solver with $\Delta t = 0.025$ over 160 steps in time. Here, $t_0 = 0$ and $t_m = T = 4$. We then added iid Gaussian noise with zero mean and variance $\sigma^2 = 0.25$ to each time series. $L = 1, 2, \ldots$ of the data series were then represented in the action at each measurement time t_n during our annealing procedure. In our calculations, $R_{f0} = 0.01$ and $\xi = 2$.

In the action, we selected $R_m = 4$, the inverse variance of the noise added to the data in our twin experiment, so the minimum action level we expect is $80.5\,L \pm 8.97\sqrt{L}$. The paths are $(m+1)\,(D+1) = 3381$-dimensional. Our search for minimum paths used a BFGS routine (Press et al., 2012) to which we provided an analytical form of the gradient of $A_0(\boldsymbol{X})$. The search was initialized with $N_I = 100$ times with initial paths from a uniform distribution of values in the interval $[-10, 10]$.

We also investigated using the IPOPT public domain numerical optimization package (Waëchter , 2002) and found that it also worked well, giving essentially the same results as BFGS. Any existing method for 4DVar should work as well for this annealing procedure.

In Fig. 1, we display the computed action levels for $L = 5$, 7, 8 and 9 at each value of $\log R_f \propto \beta$. For $L = 5$, there are many close action levels associated with the extremum

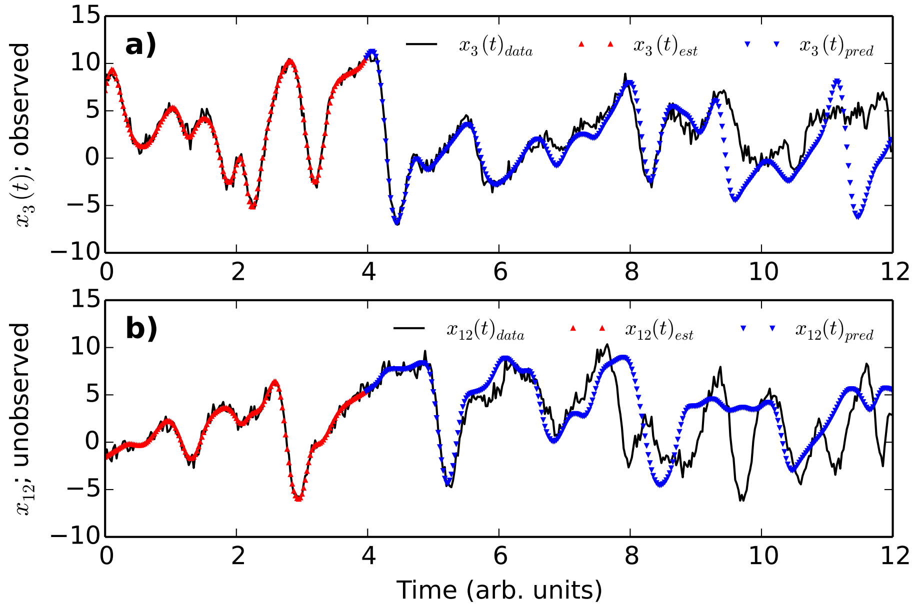

Figure 2. Data, estimated, and predicted time series for the Lorenz96 model (Lorenz, 2006) with $D = 20$, $L = 8$. **(a)** $x_3(t)$ was an observed state variable, and **(b)** $x_{12}(t)$ was unobserved. The data (black), the estimated state variable (red), and the predicted state variable (blue) are shown for each of them.

paths of the action Eq. (1); as L increases, the lowest action level visibly separates from the others. At the bottom of each panel, we indicate the lowest action level value and its standard deviation. The next-lowest action level $A_0(\boldsymbol{X}^1)$, for $L = 5$, 7, 8, and 9, is at 403.8, 749.8, 1161.6, and 2256.1, respectively. This means that for $L = 5$ we would have to sum over the contributions of many paths $\boldsymbol{X}^q$ to evaluate the expected value $\langle G(\boldsymbol{X})\rangle$. At $L = 7$ or higher, $\boldsymbol{X}^0$ dominates the integral. Our estimate for the forcing parameter, set to 8.17, was 8.22 at large β.

We think it important to note that if we had begun our search for the saddle point paths $\boldsymbol{X}^q$ at large values of R_f, we would be almost sure to miss the actual path $\boldsymbol{X}^0$, which gives the lowest action level, since the Hessian matrix of A_0 is ill conditioned when R_f is large; see Fig. 4.6 in Quinn (2010).

Another sense of why beginning a search at large values of R_f may fail to find the action level identified in the annealing approach is that the basin of attraction of the minimum action level is likely to be so small, relative to the size of path space, that "falling into it" through an arbitrary initial path used in any version of a variational procedure is unlikely. The annealing method systematically tracks the known minimum for very small R_f and, in that manner, starts in and remains in the appropriate basin of attraction.

The real test of an estimation procedure in data assimilation is not accuracy in the estimation of measured states and fixed parameters, but accuracy in prediction beyond the observation window. The predictions require accurate estimation of the unobserved model state variables at the end of the observation window. Indeed, one can achieve good "fits" of observed variables that lead to inaccurate predictions for $t > t_m = T$ (Abarbanel, 2013).

As this is a twin experiment, we show in Fig. 2 the data, the estimated state variable, and the predicted state variable for an *observed* variable $x_3(t)$ and for an *unobserved* variable $x_{12}(t)$ for $L = 8$. In a real experiment, we could not compare our estimates for the parameters or the unobserved state variables. Although the estimation procedure for the path $\boldsymbol{X}^0$ with the minimum action value is rather good, both in estimating the 12 unobserved states and the one parameter, there nevertheless exist errors in our knowledge of the full state $(\boldsymbol{x}(t) = 4)$. The predictions thus lose their accuracy in time because of the chaotic nature of the trajectories at $f = 8.17$.

In the Lorenz96 equations, one usually has a single forcing parameter f. To see how well our procedure works for several unknown parameters, we introduced ten different forcing parameters f_a into the Lorenz96 model at $D = 10$: $\dot{x}_a(t) = x_{a-1}(t)\ \ (x_{a+1}(t) - x_{a-2}(t)) - x_a(t) + f_a$, that is, with no fluctuations in the dynamics. Noise was added to the solutions to the Lorenz96 system before the $x_a(t)$ are used as data. For $D = 10$, the lowest action level stands out from the rest at $L = 4$. In Table 1, we show our estimates for the ten forcing parameters for $L = 4$, 5, and 6, as well as the actual value used in the calculations of the data. In these estimates, and for the single forcing parameter reported above for $D = 20$, there is a known source of bias (Kostuk et al., 2012). As one can see in the examples, it is small here.

In the twin experiments we presented noisy data from the known model as $L = 1, 2, \ldots$ data series of observations. To

Table 1. Known and estimated forcing parameters for the Lorenz96 model at $D = 10$, and $L = 4$, 5, and 6.

Known f_a	$L = 4$	$L = 5$	$L = 6$
5.7	5.742	5.737	5.768
7.1	7.096	7.080	7.094
9.6	9.696	9.686	9.654
6.2	6.156	6.174	6.131
7.5	7.605	7.592	7.604
8.4	8.353	8.330	8.349
5.3	5.310	5.278	5.214
9.7	9.679	9.703	9.643
8.5	8.632	8.629	8.626
6.3	6.334	6.336	6.308

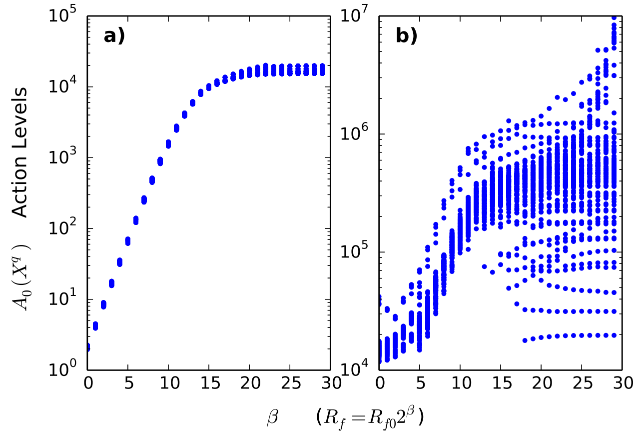

Figure 3. Action levels as a function of R_f for the Lorenz96 model, $D = 10$, $R_{f0} = 0.01$. **(a)** $L = 5$ and **(b)** $L = 6$ when the wrong data are used for the Lorenz96 model. We actually used data from four realizations of the Lorenz63 model (Lorenz, 1963). The structure of the action levels versus R_f shows no trace of the minimum allowed level Eq. (6). This indicates that the data and the model are incompatible. The action levels are also quite large, and, for $L = 6$, numerous and not well separated.

give some sense of what one might expect if the model were *totally wrong*, in the sense that the data we presented came either from a completely different model system or from enormously noisy data, we presented data from a collection of 1963 Lorenz model (Lorenz, 1963) oscillators to a Lorenz96 $D = 10$ model. Twelve time series data are generated by four individual Lorenz63 (Lorenz, 1963) systems with different initial conditions. Gaussian white noise with zero mean and standard deviation $\sigma = 0.5$ are added to each time series. All these "data" $y_l(t)$ are rescaled to lie within $[-10, 10]$.

We then place these signals as "data" in the action with the model taken as Lorenz96 $D = 10$; the single forcing parameter is treated as a time-dependent state variable obeying $\dot{f} = 0$. We use $L = 5$ and $L = 6$ as measurements using the data time series taken in the order $y_1(t)$, $y_3(t)$, $y_5(t)$, $y_7(t)$, $y_9(t)$ for $L = 5$, and $y_1(t)$, $y_3(t)$, $y_5(t)$, $y_7(t)$, $y_9(t)$, $y_2(t)$ for $L = 6$. In Fig. 3 we display the action levels versus $\log[R_f/R_{f0}] = \beta$ associated with this for $L = 5$ and $L = 6$. Results for other values of L are consistent with these. This example provides a graphic illustration of the inconsistency of the data and the model and how this makes its appearance in the annealing procedure.

We also investigated, but do not display here, the action levels when the parameter f is held at a different value in the model than was used for generating the data. In particular, we generated the data with $f = 8.17$ and then held $f = 18$ in evaluating the action levels for a Lorenz96 model with $D = 5$ and $L = 1$ and 2. In each case, no minimum action level split from the collection of $\boldsymbol{X}^q$ and no level was close to the consistent χ^2 condition discussed above. This actually emphasizes the importance of allowing the variational method to include all parameters as state variables with a zero vector field, allowing the estimation of the parameters along with the unmeasured state variables.

We now have seen that a consistent smallest action level can be identified via an annealing process, and the dependence of the action levels on the number of measurements, L, has been demonstrated. We have no formal proof that the path $\boldsymbol{X}^0$ gives a *global* minimum of the action; our criteria of consistency with Eq. (6) and excellent predictions after the observation window are functionally useful features of the procedure.

5 Corrections to the contribution of a saddle path to $\langle G(\boldsymbol{X})\rangle$

We turn back to the evaluation of the path integral for $\langle G(\boldsymbol{X})\rangle$ in Eq. (2). In that integral for our action, we have

$$\frac{1}{2}A_0^{(2)}(\boldsymbol{X}) = \gamma^2_{\alpha\beta}(\boldsymbol{X}) = R_m(n)\delta_{a,l}\delta_{b,l'} + R_f h_{\alpha\beta}(\boldsymbol{X})$$
$$\text{and } A_0^{(r)}(\boldsymbol{X}) = R_f g^{(r-2)}(\boldsymbol{X}),$$

for $r \geq 3$. The functions $h(\boldsymbol{X})$ and $g(\boldsymbol{X})$ are derivable from the form of $A_0(\boldsymbol{X})$. These are to be evaluated at $\boldsymbol{X} = \boldsymbol{X}^q$ for the qth saddle path.

In the form of the integral Eq. (4), we see that the term in the exponential with order r in $\mathbf{U}$ has an order of magnitude about $R_f/(R_m + R_f\, h(\boldsymbol{X}^q))^{r/2}$ for $r \geq 3$. Similarly, in the expansion of $G(\boldsymbol{X})$, the term of order k has a denominator of order $1/(R_m + R_f\, h(\boldsymbol{X}^q))^{k/2}$ for $k \geq 1$. Since only terms with even powers of $\mathbf{U}$ are nonzero in the integral because of symmetry, we have a collection of terms that, as R_f becomes large with respect to R_m, decrease as $1/R_f$ or faster. As a rule of thumb in the calculations we presented for Lorenz96 $D = 20$, we see that for $\beta = 15$ or larger, or $R_f/R_m \approx 100$ or larger, the terms in the path integral beyond the quadratic term in the expansion of the action become small. A stronger estimate would come from evaluating the leading term in $1/R_f$.

Table 2. List of important mathematical notations (in alphabetical order).

Symbol	Description
$A(\mathbf{X})$	Action of path $\mathbf{X}$
D	Dimensionality of the dynamical system
$f_a(\mathbf{x}(n))$	One-step discrete time map of the dynamical system
L	Dimensionality of the measured time series
m	Number of time steps in observation window
$P(\mathbf{X})$	Conditional probability of path $\mathbf{X}$ given measurements $\mathbf{y}$
R_f	Inverse of covariance matrix for model error
R_m	Inverse of covariance matrix for measurement noises
$\mathbf{x}(t)$	State variables of the dynamical system
$\mathbf{x}(0)$	Initial state of the dynamical system
$\mathbf{X}$	Path of the model state composed of state variables $\mathbf{X} = \{\mathbf{x}(t_0), \ldots, \mathbf{x}(t_m)\}$
$\mathbf{X}^q$	Paths satisfying the saddle point condition, ordered by action values $A_0(\mathbf{X}^q)$
$\mathbf{y}(t_n)$	Measured data time series

6 Conclusions

We have examined the path integral formulation of data assimilation Eq. (2) and asked how well a variational approach to the conditional expected values of functions $G(\boldsymbol{X})$ on the path $\boldsymbol{X} = \{\boldsymbol{x}(0), \boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(m)\}$ can approximate the integral. We use Laplace's method, which estimates the path integral by seeking saddle paths of the action where $\partial A_0(\boldsymbol{X})/\partial X_\alpha = 0$. This approximation is widely used in meteorology and other fields, where it is known as weak 4DVar (Talagrand and Courtier, 1987; Evensen, 2009; Lorenc and Payne, 2007). The action is the cost function that is minimized.

When measurement errors and model errors are distributed as iid Gaussian noise, we have described an annealing method in the strength R_f with which the model errors enter the action that permits the following of a collection of action levels from $R_f = 0$, where the saddle paths are known precisely, through a set of increasing values of R_f at each of which a numerical optimization algorithm produces a set of saddle paths that are used as the initial conditions for the solution of the variational approximation at the next larger value of R_f. This permits the identification of a lowest action level associated with a saddle path $\boldsymbol{X}^0$ where $A_0(\boldsymbol{X}^0)$ is split from the action values for the other possible saddle paths $\boldsymbol{X}^{q>0)}$ at each R_f: $A_0(\boldsymbol{X}^0) < A_0(\boldsymbol{X}^{q>0)})$. The contribution of $\boldsymbol{X}^0$ then dominates the integral.

We also explored the dependence of the action levels revealed through the annealing method on the number of measurements L made at each observation time. When L reaches and exceeds a critical value, related to the number of unstable directions in the deterministic chaotic dynamics of the model, the contribution of the path $\boldsymbol{X}^0$ is either certainly the dominant contribution to the conditional expected value of $G(\boldsymbol{X})$, or it is the only path satisfying the saddle path condition. By expanding the integrand of Eq. (2) about $\boldsymbol{X}^0$ we argued that the resulting corrections to the contribution of $\boldsymbol{X}^0$ produced a power series in R_f^{-1}.

In previous work with variational principles for data assimilation, we are unaware of any procedure such as our annealing method using R_f to identify a consistent minimum action. Nor are we aware of a systematic exploration of the dependence of the action levels on the number of measurements at each observation time. (Of course, there is the discussion in Quinn, 2010, which suggested this work.) Finally, we do not know of any previous discussion of the corrections to the variational approximation, which here is shown to consist of small perturbations when the resolution of the model error term in the action is increased; namely, R_f becomes large.

The relation of the annealing method to familiar 4DVar calculations (Talagrand and Courtier, 1987; Evensen, 2009) is simple to state: each set of calculations during the annealing in R_f, at each value of R_f, is a 4DVar calculation. The annealing has the added advantage of allowing one to establish an identified lowest action value path $\boldsymbol{X}^0$ that gives the dominant contribution to the quantities of interest, namely, the conditional expected values of functions $G(\boldsymbol{X})$ on the path.

A similar annealing procedure in the importance of the model errors as represented by R_f can be incorporated into a Monte Carlo calculation of the path integral. The relation between that method of approximating the high-dimensional integral and the variational method we have focused on here is through the Langevin equation in "algorithmic time" described in Abarbanel (2013). In each approach, the statistics of moments about $\boldsymbol{X}^0$ or marginal distributions of selected state variables is achieved by the choice of $G(\boldsymbol{X})$. This paper has compared those evaluations in the variational method to standard 4DVar by using 4DVar at each stage of the annealing adding the identification of the path $\boldsymbol{X}^0$ with the lowest action level. By performing the annealing procedure in a Monte Carlo context, one could compare it to other standard ensemble methods such as the ensemble Kalman filter (EnKF) (Evensen, 2009). However, that is a subject for another investigation of the annealing approach; we have focused here only on variational methods.

We have worked within a framework where the measurement errors and the model errors in the data assimilation are Gaussian, with the inverse covariance of the model errors taken to be of order R_f. We have shown that the path that gives a consistent minimum action level can be traced by an annealing procedure starting with a setting where the dynamical model essentially plays no role, $R_f \approx 0$, then systematically increasing the influence of the model dynamics. As the scale of the model error reaches about 100 times the scale of the measurement error, the path $\boldsymbol{X}^0$ associated with a consistent global minimum action level dominates the path integral with corrections of order $1/R_f$. The important role of the number of measurements L at each measurement time is also demonstrated.

If the noise terms represented by the model errors are not Gaussian, one can still use the annealing method to identify a path with the lowest action level, but showing that perturbation theory about the path $\boldsymbol{X}^0$ giving that lowest action level is a power series in R_f^{-1} may not succeed. The precise way R_f enters the matrix $\gamma^2_{\alpha\alpha'}(\boldsymbol{X}^0)$ determines that.

In this paper, we do not address the typical situation where the number of measurements actually available is less than that needed to allow the ground level of the action $A_0(\boldsymbol{X}^0)$ to lie well below the next level. For a solution to that problem, we have used information from the waveform of the measurements as shown in some detail in Rey et al. (2014).

The results here justify the use of the variational approximation in data assimilation, focus on the role of the number of measurements one requires for accuracy in that approximation, and permit the evaluation of systematic corrections to the approximation when the form of the action is Eq. (1). We anticipate that our use of Gaussian model error and measurement error is a convenience and that other distributions of these errors will permit many of the same set of statements about the value of variational approximations to statistical data assimilation.

Acknowledgements. Partial support has come from the ONR MURI program (N00014-13-1-0205). We are very appreciative of the considered comments of the anonymous referees during the discussion period of this paper. Indeed, one of the first referee's questions led to the idea that the annealing approach could be useful in Monte Carlo estimations of the expected value path integrals.

Edited by: Z. Toth;
Reviewed by: two anonymous referees

References

Abarbanel, H. D.: Predicting the Future: Completing Models of Observed Complex Systems, Springer, New York, 2013.

Aguiar e Oliviera, H., Ingber, L., Petraglia, A., Petraglia, M. R., and Machado, M. A. S.: Stochastic Global Optimization and Its Applications with Fuzzy Adaptive Simulated Annealing, Vol. 35, Springer, New York, 2012.

Bennett, A. F.: Inverse Modeling of the Ocean and Atmosphere, Cambridge University Press, 2002.

Eibern, H. and Schmidt, H.: A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modeling, J. Geophys. Res.-Atmos., 104, 18583–18598, 1999.

Evensen, G.: Data Assimilation: The Ensemble Kalman Filter, Springer, New York, 2009.

Kostuk, M.: Synchronization and statistical methods for the data assimilation of HVc neuron models, PhD thesis in Physics, University of California, San Diego, available at: http://escholarship.org/uc/item/2fh4d086 (last access: 2 October 2014), 2012.

Kostuk, M., Toth, B., Meliza, C., Margoliash, D., and Abarbanel, H.: Dynamical estimation of neuron and network properties II: path integral Monte Carlo methods, Biol. Cybern., 106, 155–167, 2012.

Laplace, P. S.: Memoir on the probability of causes of events, Mémoires de Mathématique et de Physique, Tome Sixième, 1774.

Lorenc, A. C. and Payne, T.: 4D-Var and the butterfly effect: statistical four-dimensional data assimilation for a wide range of scales, Q. J. Roy. Meteorol. Soc., 133, 607–614, 2007.

Lorenz, E. N.: Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130–141, 1963.

Lorenz, E. N.: Predictability – a problem partly solved, in: Predictability of Weather and Climate, edited by: Palmer, T. and Hagedorn, R., Cambridge University Press, 40–58, 2006.

Mechhoud, S., Witrant, E., Dugard, L., and Moreau, D.: Combined distributed parameters and source estimation in tokamak plasma heat transport, in: 2013 European Control Conference (ECC), 17–19 July, Zurich, 47–52, 2013.

NIST/SEMATECH e-Handbook of Statistical Methods: http://www.itl.nist.gov/div898/handbook/ (last access: 12 January 2014), 2012.

Press, W. H., Teokulsky, S. A., Vetterling, W. T., and Flannery, B. P.: Numerical Recipes in C: the Art of Scientific Computing, Cambridge University Press, 2012.

Quinn, J. C.: A path integral approach to data assimilation in stochastic nonlinear systems, PhD thesis in Physics, University of California, San Diego, available at: http://escholarship.org/uc/item/0bm253qk (last access: 2 October 2014), 2010.

Rey, D., Eldridge, M., Kostuk, M., Abarbanel, H. D., Schumann-Bischoff, J., and Parlitz, U.: Accurate state and parameter estimation in nonlinear systems with sparse observations, Phys. Lett. A, 378, 869–873, 2014.

Talagrand, O. and Courtier, P.: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation, I: Theory, Q. J. Roy. Meteorol. Soc., 113, 1311–1328, 1987.

Waëchter, A.: An Interior Point Algorithm for Large-Scale Nonlinear Optimization with Applications in Process Engineering, PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2002.

Zadeh, K. S.: Parameter estimation in flow through partially saturated porous materials, J. Comput. Phys., 227, 10243–10262, 2008.

Zinn-Justin, J.: Quantum Field Theory and Critical Phenomena, Oxford University Press, 2002.