Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

The Onsager–Machlup functional for data assimilation

Nozomi Sugiura[1]

[1]Research Institute for Global Change, JAMSTEC, Yokosuka, Japan

Correspondence to: Nozomi Sugiura (nsugiura@jamstec.go.jp)

Abstract. When taking the model error into account in data assimilation, one needs to evaluate the prior distribution represented by the Onsager–Machlup functional. Numerical experiments have clarified how one should put it into discrete form in the maximum a posteriori estimates and in the assignment of probability to each path. In the maximum a posteriori estimates, the divergence of the drift term is essential, but for the path probability assignments in combination with the Euler time-discretization scheme, it is not necessary. The latter property will help simplify the implementation of nonlinear data assimilation for large systems with sampling methods such as the Metropolis-adjusted Langevin algorithm.

1 Introduction

In traditional weak-constraint 4D-Var setting (e.g., Zupanski, 1997; Trémolet, 2006), a quadratic cost function is defined as the negative logarithm of the probability for each Brownian path, which is suitable for path sampling (e.g., Zinn-Justin, 2002). The optimization problem is naively described as finding the most probable path by minimizing the quadratic cost function. However, the term "the most probable path" does not make sense in this context, since the paths are not countable. One should notice that the concern is not about ranking the individual path probabilities, but about seeking the route with the densest path population. To define the optimization problem properly, one should introduce a macroscopic variable ϕ that represents a smooth curve, and introduce a measure μ that accounts how dense the paths that lie in the ϵ-neighbor centered at ϕ are, which can be termed as "the tube." Then, the density of paths is formulated as $p(\text{tube}) = \mu(\text{paths in the tube})p(\text{curve})$. For a stochastic differential equation (SDE) with drift term f and additive noise, the second term of the rhs takes a quadratic form $C_1 \exp\left(-\frac{1}{2}\int |\dot{\phi}(t) - f(\phi(t))|^2 dt\right)$, while the first term μ is in the form of $C_2 \exp\left(-\frac{1}{2}\int \mathrm{div}\, f(\phi(t))dt\right)$, which accounts for how densely the paths are populated in the tube. Mathematicians pioneering the theory of SDE (e.g., Ikeda and Watanabe, 1981; Zeitouni, 1989) were already aware of this subtle point since the 1980s, and established the proper form of cost function as the Onsager–Machlup (OM) functional (Onsager and Machlup, 1953) for the most probable tube.

The aim of this work is to organize the existing knowledge about the OM functional in a form that can be applicable to model error representations in data assimilation, i.e., numerical evaluation of nonlinear smoothing problems.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Throughout this article, we consider nonlinear smoothing problems of the form

$$dx_t = f(x_t)dt + \sigma dw_t, \tag{1}$$

$$x_0 \sim \mathcal{N}(x_b, \sigma_b^2 I), \tag{2}$$

$$(\forall m \in M) \quad y_m | x_m \sim \mathcal{N}(x_m, \sigma_o^2 I), \tag{3}$$

where t is time, x is a D-dimensional stochastic process, w is a D-dimensional Wiener process, $x_b \in \mathbb{R}^D$ is the background value of the initial condition, $\sigma_b > 0$ is the standard deviation of the background value, $y_m \in \mathbb{R}^D$ is observational data at time t_m, $x_m = x_{t_m}$, $t_m = m\delta_t$, M is the set of observation times, $\sigma_o > 0$ is the standard deviation of the observational data, and $\sigma > 0$ is the noise intensity.

Following the derivation in Section 2.3 of Law et al. (2015), we can assign each path a posterior probability

$$P(x|y) \propto P(x)P(y|x) = P(x|x_0)P(x_0)P(y|x) = \prod_{n=1}^{N} P(x_n|x_{n-1})P(x_0) \prod_{m \in M} P(y_m|x_m). \tag{4}$$

According to Eq. (2), the prior probability for the initial condition is given as

$$P(x_0) \propto \exp\left(-\frac{|x_0 - x_b|^2}{2\sigma_b^2}\right), \tag{5}$$

where $|x_0 - x_b|^2$ represents the squared Euclidean norm $\sum_{i=1}^{D}(x_0^i - x_b^i)^2$. According to Eq. (3), the likelihood of the state x_m, given observation y_m, is

$$P(y_m|x_m) \propto \exp\left(-\frac{|y_m - x_m|^2}{2\sigma_o^2}\right). \tag{6}$$

Now, we move on to approximation with a discrete time step. The change-of-measure argument (Appendix B1) or the path integral argument (e.g., Zinn-Justin, 2002) on a path of this stochastic process shows that Eq. (1) has the transition probability at discrete time steps

$$P(x_n|x_{n-1}) \propto \exp\left(-\frac{\delta_t}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2\right), \tag{7}$$

called the Euler scheme, which uses the drift $f(x_{n-1})$ at the previous time step. This transition probability has another expression (see the derivation of Eq. (B8) in Appendix B1 or Zinn-Justin (2002)):

$$P(x_n|x_{n-1}) \propto \exp\left(-\frac{\delta_t}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2 - \frac{\delta_t}{2}\operatorname{div} f(x_{n-\frac{1}{2}})\right), \tag{8}$$

$$f(x_{n-\frac{1}{2}}) \equiv \frac{f(x_n) + f(x_{n-1})}{2}, \quad \operatorname{div} f(x) \equiv \sum_{i=1}^{D} \frac{\partial f^i}{\partial x^i}(x), \tag{9}$$

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

which can be called the trapezoidal scheme because the integral is evaluated with the drift terms at both ends of each interval. The transition probability leads to the prior probability $P(x|x_0)$ of a path $x = \{x_n\}_{0\le n\le N}$ as follows (e.g., Zinn-Justin, 2002):

$$P(x|x_0) \propto \exp\left(-\delta_t \sum_{n=1}^{N} \frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2\right) \tag{10}$$

$$\leftrightharpoons \exp\left(-\delta_t \sum_{n=1}^{N}\left[\frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(x_{n-\frac{1}{2}})\right]\right), \tag{11}$$

where "$\leftrightharpoons$" sign indicates that, if δ_t is sufficiently small, the equations on the both sides are compatible.

On the other hand, based on the argument in Appendix B2, we can also define the probability $P(U_\phi|\phi_0)$ for a smooth tube that represents its neighboring paths $U_\phi = \{\omega \mid (\forall n)|\phi_n - x_n(\omega)| < \epsilon\}$:

$$P(U_\phi|\phi_0) \propto \exp\left(-\delta_t \sum_{n=1}^{N}\left[\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-\frac{1}{2}})\right]\right). \tag{12}$$

The scaling argument for a smooth curve in Appendix A allows us to use the drift term $f(\phi_{n-1})$ instead in Eq. (12):

$$P(U_\phi|\phi_0) \propto \exp\left(-\delta_t \sum_{n=1}^{N}\left[\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-1})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-1})\right]\right). \tag{13}$$

The corresponding posterior probabilities are thus given as follows:

$$P_{\text{path}}(x|y) \propto \exp\left(-J_{\text{path}}(x|y)\right), \tag{14}$$

$$J_{\text{path}}(x|y) \equiv \frac{1}{2\sigma_b^2}|x_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|x_m - y_m|^2 + \delta_t \sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2\right) \tag{15}$$

$$\leftrightharpoons \frac{1}{2\sigma_b^2}|x_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|x_m - y_m|^2 + \delta_t \sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(x_{n-\frac{1}{2}})\right) \tag{16}$$

for a Brownian path and

$$P_{\text{tube}}(U_\phi|y) \propto P(U_\phi|\phi_0)P(\phi_0)P(y|U_\phi) \propto \exp\left(-J_{\text{tube}}(\phi|y)\right), \tag{17}$$

$$J_{\text{tube}}(\phi|y) \equiv \frac{1}{2\sigma_b^2}|\phi_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|\phi_m - y_m|^2 + \delta_t \sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-\frac{1}{2}})\right) \tag{18}$$

$$\leftrightharpoons \frac{1}{2\sigma_b^2}|\phi_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|\phi_m - y_m|^2 + \delta_t \sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-1})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-1})\right) \tag{19}$$

for a smooth tube. Note that different pairs of time-discretization schemes of the OM functional, $\frac{1}{2\sigma^2}\left(\frac{dx}{dt} - f(x)\right)^2 + \frac{1}{2}\mathrm{div}\,(f)$, are nominated for paths and for tubes in Eqs. (15), (16), (18), and (19).

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

2 Method

2.1 Four schemes for OM

In the argument in Section 1, the prior probability has a form $P(x|x_0) \propto \exp\left(-\delta_t \sum_{n=1}^{N} \widetilde{OM}\right)$, where $\widetilde{OM}$ is the OM functional (Onsager and Machlup, 1953). Including those shown in Eqs. (15), (16), (18), and (19), the possible candidates for the discretization schemes of the OM functional would be as follows:

1. Euler scheme (E) (e.g., Zinn-Justin, 2002; Dutra et al., 2014):

$$\widetilde{OM}_{\mathrm{E}} \equiv \frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2; \tag{20}$$

2. Euler scheme with divergence term (ED):

$$\widetilde{OM}_{\mathrm{ED}} \equiv \frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2 + \frac{1}{2}\mathrm{div}\, f(x_{n-1}), \tag{21}$$

   where $f(x_{n-\frac{1}{2}}) = (f(x_n) + f(x_{n-1}))/2$;

3. Trapezoidal scheme (T):

$$\widetilde{OM}_{\mathrm{T}} \equiv \frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2; \tag{22}$$

4. Trapezoidal scheme with divergence term (TD) (e.g., Ikeda and Watanabe, 1981; Apte et al., 2007; Dutra et al., 2014):

$$\widetilde{OM}_{\mathrm{TD}} \equiv \frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(x_{n-\frac{1}{2}}), \tag{23}$$

   where $f(x_{n-\frac{1}{2}}) = (f(x_n) + f(x_{n-1}))/2$.

By using the cost function adopted in one of the above schemes in the model error term, we can apply a data assimilation algorithm, either Markov-chain Monte Carlo (MCMC) (e.g., Metropolis et al., 1953) or four-dimensional variational data assimilation (4D-Var) (e.g., Zupanski, 1997). Among versions of MCMC, we focus on the Metropolis-adjusted Langevin algorithm (MALA) (e.g., Roberts and Rosenthal, 1998; Cotter et al., 2013). MALA samples the paths $x^{(k)} = \{x_n(\omega_k)\}_{0\le n\le N}$ according to the distribution P_{path} by iterating ($\alpha > 0$):

$$x^{(k+1)} = x^{(k)} - \alpha \nabla J_{\mathrm{path}} + \sqrt{2\alpha}\xi, \quad \xi \sim \mathcal{N}(0,1)^{D(N+1)}, \quad \nabla J = \left(\frac{\partial J}{\partial x}\right)^T \tag{24}$$

with the Metropolis rejection step for adjustment to get an ensemble of sample paths according to the posterior probability, while 4D-Var seeks the center of the most probable tube $\phi = \{\phi_n\}_{0\le n\le N}$ by iterating ($\alpha > 0$):

$$\phi^{(k+1)} = \phi^{(k)} - \alpha \nabla J_{\mathrm{tube}}. \tag{25}$$

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

To investigate the applicability of the four candidate schemes, we use them in these algorithms.

The results should be checked with "the right answer." The reference solution that approximates the right answer is provided by a naive particle smoother (PS) (e.g., Doucet et al., 2000), which does not involve the explicit computation of prior probability. When we have observations only at the end of the assimilation window, the PS algorithm is as follows:

1. Generate samples of initial and model errors, integrate M copies of the model, and use them to obtain a Monte-Carlo approximation of the prior distribution:

$$P(x) \simeq \frac{1}{M}\sum_{m=1}^{M}\prod_{n=0}^{N}\delta(x_n - \chi_n^{(m)}), \tag{26}$$

where $\chi_n^{(m)}$ is the state of member m at time n.

2. Reweight it according to Bayes's theorem:

$$P(y|x) \propto \exp\left(-\frac{1}{2\sigma_o^2}|y - x_N|^2\right), \tag{27}$$

$$P(x|y) = \frac{P(x)P(y|x)}{\int dx P(x)P(y|x)} = \sum_{m=1}^{M}\prod_{n=0}^{N}\delta(x_n - \chi_n^{(m)})\frac{w^{(m)}}{\sum_{m=1}^{M} w^{(m)}}, \tag{28}$$

$$w^{(m)} \equiv \exp\left(-\frac{1}{2\sigma_o^2}|y - \chi_N^{(m)}|^2\right). \tag{29}$$

3 Results

3.1 Example A (hyperbolic model)

In our first example, we solve the nonlinear smoothing problem for the hyperbolic model (Daum, 1986), which is a simple problem with one-dimensional state space, but which has a nonlinear drift term. We want to find the probability distribution of the paths described by

$$dx_t = \tanh(x_t)dt + dw_t, \quad x_{t=0} \sim \mathcal{N}(0, 0.16), \tag{30}$$

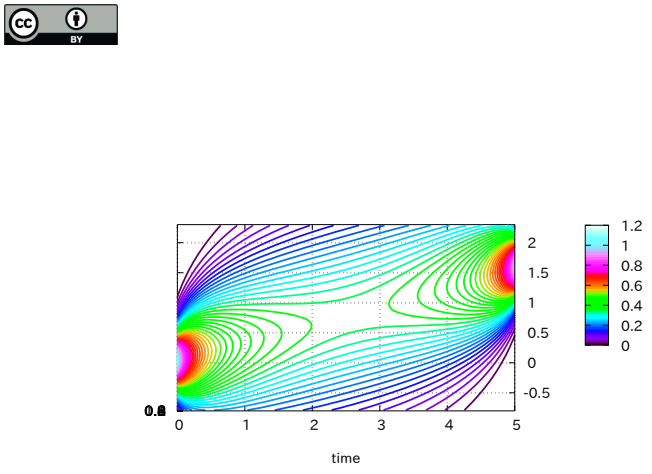
subject to an observation y:

$$y|x_{t=5} \sim \mathcal{N}(x_{t=5}, 0.16), \quad y = 1.5. \tag{31}$$

The setting follows Daum (1986). In this case, $\mathrm{div}\, f(x) = 1/\cosh^2(x)$ imposes a penalty for small x.

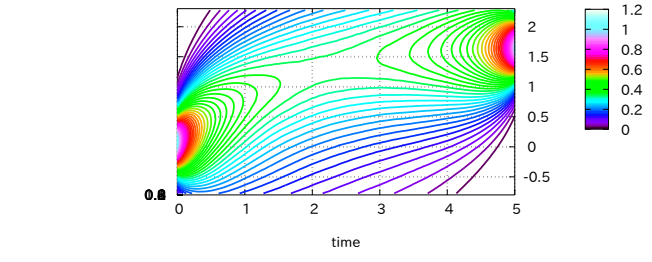
Figure 1 shows the probability densities of paths normalized on each time slice, $P_{t=n}(\phi) = \int P(U_\phi|y)d\phi_{t\neq n}$, derived by MCMC and PS. PS is performed with 5.1×10^{10} particles. You can see that MCMC with E or TD provides the proper distribution matched with that of PS; this is also clear from the expected paths yielded by these experiments, shown in Fig. 2. These schemes correspond to candidates in Eqs. (15) and (16). The expected path by ED bends toward larger x, which should be

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017



(a) Reference solution by PS

(b) MCMC with Scheme E or TD



(c) MCMC with Scheme ED

(d) MCMC with Scheme T

Figure 1. Probability density of paths derived by MCMC and PS for the hyperbolic model.

caused by an extra penalty for larger x. The expected path by T bends toward smaller x, which should be caused by the lack of penalty for larger x.

The results of 4D-Var, which represents the maximum a posteriori (MAP) estimates of the tube, are shown in Fig. 3. ED and TD provide the proper MAP estimate of the tube. These schemes correspond to candidates in Eqs. (18) and (19). The expected paths by E and T bend toward smaller ϕ, which should be caused by the lack of penalty for larger ϕ.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

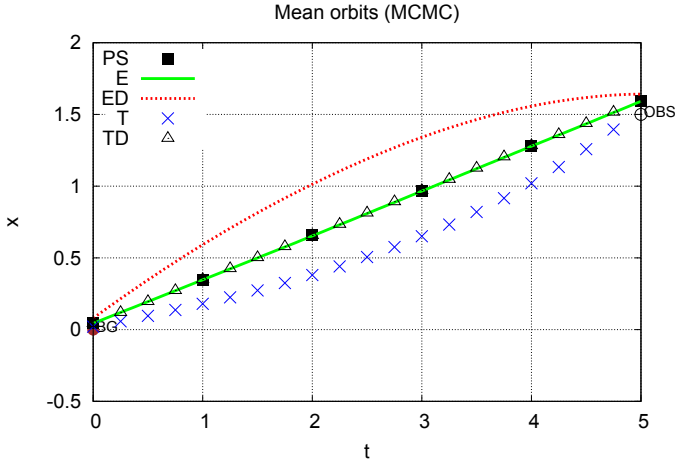


Figure 2. Expected path derived by MCMC (hyperbolic model).

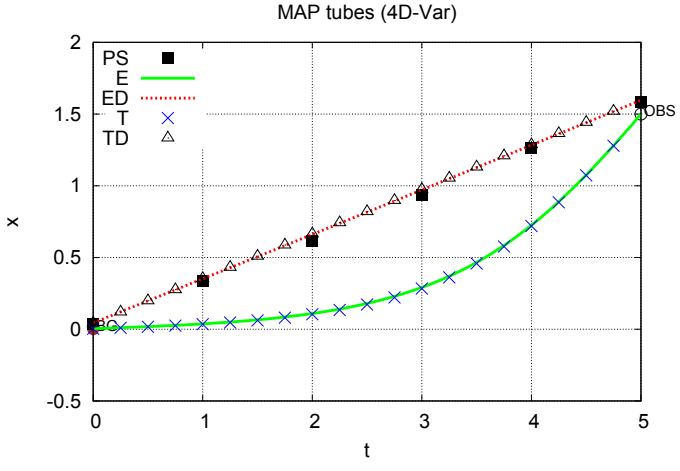


Figure 3. Most probable tube derived by 4D-Var (hyperbolic model).

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

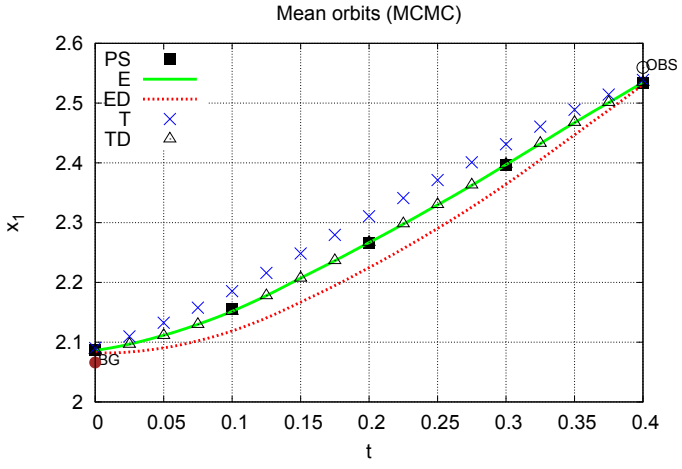


Figure 4. Expected path derived by MCMC (Rössler model).

3.2 Example B (Rössler model)

In our second example, we solve the nonlinear smoothing problem for the stochastic Rössler model (Rössler, 1976). We want to find the probability distribution of the paths described by

$$\begin{cases} dx_1 &= (-x_2 - x_3)dt + \sigma dw_1, \\ dx_2 &= (x_1 + ax_2)dt + \sigma dw_2, \\ dx_3 &= (b + x_1 x_3 - cx_3)dt + \sigma dw_3, \end{cases} \tag{32}$$

$$x_{t=0} \sim \mathcal{N}(x_b, 0.04I), \tag{33}$$

subject to an observation y:

$$y|x_{t=0.4} \sim \mathcal{N}(x_{t=0.4}, 0.04I), \tag{34}$$

where $(a,b,c) = (0.2, 0.2, 6)$, $\sigma = 2$, $x_b = (2.0659834, -0.2977757, 2.0526298)^T$, and $y = (2.5597086, 0.5412736, 0.6110939)^T$. In this case, $\mathrm{div}\, f(x) = x_1 + a - c$ imposes a penalty for large x_1.

The results by MCMC and 4D-Var for the Rössler model are shown in Figs. 4 and 5, respectively. The state variable x_1 is chosen for the vertical axes. PS is performed with 3×10^{12} particles. The curve for PS in Fig. 5 indicates $\hat{\phi} = \mathrm{argmax}_\phi P[\phi|Y]$, where U represents the tube centered at ϕ with radius 0.03.

Figure 4 shows that, just as for the hyperbolic model, E and TD provide the proper expected path. Figure 5 shows that ED and TD provide the proper MAP estimate of the tube.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

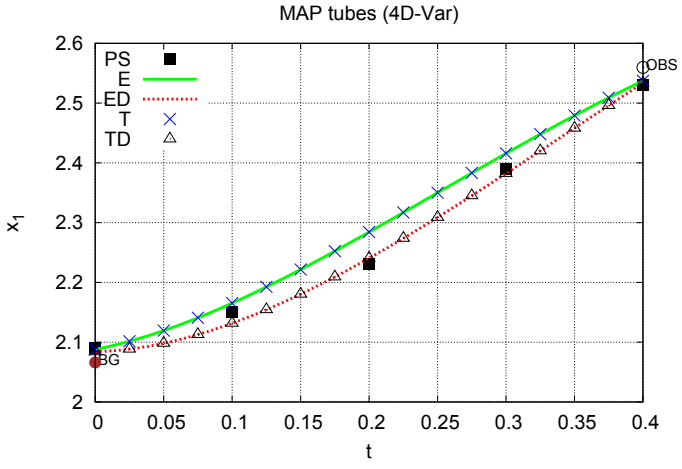


Figure 5. Most probable tube derived by 4D-Var (Rössler model).

3.3 Toward application to large systems

When one computes the cost value $J(x)$, the negative logarithm of the posterior probability, in data assimilation, the value $f(x)$ is explicitly computed via the numerical model, while $\mathrm{div}\, f(x)$ is not. If the dimension D of the state space is large, and f is complicated, the algebraic expression of $\mathrm{div}\, f(x)$ can be difficult to obtain. The gradient of the cost function $\nabla J(x)$ contains the derivative of $f(x)$, which can be implemented as the adjoint model via numerical differentiation (e.g., Giering and Kaminski, 1998). However, schemes with the divergence term require the calculation of the second derivative of $f(x)$, for which the algebraic expression can be even more difficult to obtain. Still, there may be a way to circumvent this difficulty by utilizing Hutchinson's trace estimator (Hutchinson, 1990) (See Appendix C). It is also clear that the Euler scheme without the divergence term is more convenient for implementation of path sampling, because it does not require cumbersome calculation of the divergence term.

4 Conclusions

We examined several discretization schemes of the OM functional, $\frac{1}{2\sigma^2}\left(\frac{dx}{dt} - f(x)\right)^2 + \frac{1}{2}\mathrm{div}\,(f)$, for the nonlinear smoothing problem

$$dx_t = f(x_t)dt + \sigma dw_t,$$
$$x_0 \sim \mathcal{N}(x_b, \sigma_b^2 I), \quad (\forall m \in M)\ y_m | x_m \sim \mathcal{N}(x_m, \sigma_o^2 I)$$

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Table 1. Applicable OM schemes

		with $\mathrm{div}\,(f)$	without $\mathrm{div}\,(f)$
Sampling by MCMC	Euler scheme		✓
	trapezoidal scheme	✓	
MAP estimate by 4D-Var	Euler scheme	✓	
	trapezoidal scheme	✓	

by matching the answers given by MCMC and 4D-Var with that given by PS, taking the hyperbolic model and the Rössler model as examples. Table 1 shows which of the discretization schemes were found to be applicable. These results are consistent with the literature (e.g., Apte et al., 2007; Malsom and Pinski, 2016; Dutra et al., 2014; Stuart et al., 2004).

This justifies, for instance, the use of the following cost function for the MAP estimate given by 4D-Var:

$$J = \frac{|\phi_0 - x_b|^2}{2\sigma_b^2} + \sum_{m\in M} \frac{|\phi_m - y_m|^2}{2\sigma_o^2} + \delta_t \sum_{n=1}^{N} \left(\frac{1}{2\sigma^2} \left| \frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-1}) \right|^2 + \frac{1}{2} \mathrm{div}\, f(\phi_{n-1}) \right),$$

where n is the time index, δ_t is the time increment, x_b is the background value, σ_b is the standard deviation of the background value, y is the observational data, σ_o is the standard deviation of the observational data, and σ is the noise intensity. However, the divergence term above should be excluded for the assignment of path probability in MCMC.

For application in large systems, the Euler scheme without the divergence term is preferred for path sampling because it does not require cumbersome calculation of the divergence term. In 4D-Var, the divergence term can be incorporated into the cost function by utilizing Hutchinson's trace estimator.

Code availability. The codes for data assimilation are available at https://github.com/nozomi-sugiura/OnsagerMachlup/.

Appendix A: Scaling of the terms

Taylor expansion of the $f(x_{n-1})$ term around $x_{n-\frac{1}{2}}$ in scheme E gives

$$\widetilde{OM} \simeq \sum_{n=1}^{N} \delta_t \left\{ \sigma^{-2} \left[\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}}) - (x_n - x_{n-1}) \frac{\partial f}{\partial x}(x_{n-\frac{1}{2}}) \right]^2 + \mathrm{div}\,(f) \right\}$$
$$= \delta_t \left\{ \sigma^{-2} (\mathrm{noise} + \mathrm{shift})^2 + \mathrm{divergence} \right\}.$$

$$\mathrm{noise} \equiv \frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}}), \mathrm{shift} \equiv (x_n - x_{n-1}) \frac{\partial f}{\partial x}(x_{n-\frac{1}{2}}), \mathrm{divergence} \equiv \mathrm{div}\,(f),$$

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

where we assume order-one fluctuations: $\sigma = O(1)$.

For a sample path of the stochastic process, the scaling $x_n - x_{n-1} = O(\delta_t^{\frac{1}{2}})$ leads to

$$\widetilde{OM} = \sum \delta_t \left\{ \sigma^{-2} \left(\underbrace{\text{noise}^2}_{\delta_t^{-1}} + \underbrace{\text{noise} \times \text{shift}}_{1} + \underbrace{\text{shift}^2}_{\delta_t} \right) + \underbrace{\text{divergence}}_{1} \right\}. \tag{A1}$$

The shift term induces a Jacobian that coincides with the divergence term in TD (Zinn-Justin, 2002).

For a smooth tube, the scaling $x_n - x_{n-1} = O(\delta_t)$ leads to

$$\widetilde{OM} = \sum \delta_t \left\{ \sigma^{-2} \left(\underbrace{\text{noise}^2}_{1} + \underbrace{\text{noise} \times \text{shift}}_{\delta_t} + \underbrace{\text{shift}^2}_{\delta_t^2} \right) + \underbrace{\text{divergence}}_{1} \right\}. \tag{A2}$$

The shift term is negligible, but the divergence term is not.

Appendix B: Divergence term

B1 Divergence term in trapezoidal scheme

Consider two stochastic processes (cf., Section 6.3.2 of Law et al. (2015)):

$$dx_t = f(x_t)dt + dw_t, \quad x(0) = x_0, \tag{B1}$$

$$dx_t = dw_t, \quad x(0) = x_0, \tag{B2}$$

where (B1) has measure μ and (B2) has measure μ_0 (Wiener measure). By the Girsanov theorem, the Radon–Nikodym derivative of μ with respect to μ_0 is

$$\frac{d\mu}{d\mu_0} = \exp\left[-\int_0^T \left(\frac{1}{2}|f(x)|^2 dt - f(x) \cdot dx \right) \right]. \tag{B3}$$

If we define $F(x_T) - F(x_0) = \int_{x_0}^{x_T} f(x) \circ dx$ with the Stratonovich integral, then by Ito's formula,

$$dF = f \cdot dx + \frac{1}{2}\text{div}\,(f)dt. \tag{B4}$$

Eliminating $f \cdot dx$ in Eq. (B3) using Eq. (B4), we get

$$\frac{d\mu}{d\mu_0} = \exp\left[-\int_0^T \frac{1}{2}|f(x)|^2 dt + F(x_T) - F(x_0) - \frac{1}{2}\int_0^T \text{div}\,(f)dt \right]. \tag{B5}$$

Substituting $F(x_T) - F(x_0) = \int_0^T f \circ \frac{dx}{dt} dt$,

$$\frac{d\mu}{d\mu_0} = \exp\left[-\int_0^T \frac{1}{2}|f(x)|^2 dt + \int_0^T f \circ \frac{dx}{dt} dt - \frac{1}{2}\int_0^T \text{div}\,(f)dt \right]. \tag{B6}$$

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

If we write the Wiener measure formally as $\mu_0 = \exp\left[-\frac{1}{2}\int_0^T \left|\frac{dx}{dt}\right|^2 dt\right]$, we get from Eq. (B3)

$$\mu = \exp\left[-\int_0^T \frac{1}{2}\left|\frac{dx}{dt} - f(x)\right|^2 dt\right] \tag{B7}$$

and from Eq. (B6)

$$\mu = \exp\left[-\int_0^T \frac{1}{2}\left(\left|\frac{dx}{dt} - f(x)\right|^2 + \operatorname{div}(f)\right) dt\right], \tag{B8}$$

where the integrals should be interpreted in the Ito sense and in the Stratonovich sense, respectively.

B2 Divergence term for smooth tube

When you assign weight to smooth tubes, there should always be a divergence term, for the following reason.

Let x be a diffusion process that follows the stochastic differential equation

$$dx_t = f(x_t)dt + dw_t, \tag{B9}$$

where w is a Wiener process. To investigate paths near a smooth curve ϕ, let us consider the following stochastic process $x_t - \phi(t)$ (Zeitouni, 1989):

$$d(x_t - \phi(t)) = (f(x_t - \phi(t) + \phi(t)) - \dot{\phi}(t))dt + dw_t. \tag{B10}$$

Since the process $x_t - \phi(t)$ is shifted from the Wiener process w_t by the drift $f(\cdot + \phi) - \dot{\phi}$, we can apply Girsanov's formula to get

$$\frac{P(\|x-\phi\|_T < \epsilon)}{P(\|w\|_T < \epsilon)} = \mathbb{E}\left[\exp\left(\int_0^T \left(f(w_t + \phi(t)) - \dot{\phi}(t)\right)\cdot dw_t - \frac{1}{2}\int_0^T \left|f(w_t+\phi(t)) - \dot{\phi}(t)\right|^2 dt\right) \mid \|w\|_T < \epsilon\right] \tag{B11}$$

$$= \mathbb{E}\left[\exp\left(\int_0^T \left(f(\phi(t)) + w_t \cdot \nabla f(\phi(t)) + O(w^2) - \dot{\phi}(t)\right)\cdot dw_t - \frac{1}{2}\int_0^T \left|f(w_t+\phi(t)) - \dot{\phi}(t)\right|^2 dt\right) \mid \|w\|_T < \epsilon\right], \tag{B12}$$

where $\|w\|_T \equiv \sup_{0<t<T}|w_t|$. Application of Ito's lemma to $\int \sum_{i,j} w_j \frac{\partial f_i}{\partial x_j} dw^i$ leads to

$$\frac{P(|x-\phi| < \epsilon)}{P(|w| < \epsilon)} = \exp\left[-\frac{1}{2}\int_0^T \left|f(\phi(t)) - \dot{\phi}(t)\right|^2 dt - \frac{1}{2}\int_0^T \operatorname{div} f(\phi(t))dt\right] \mathbb{E}[\cdots \mid \|w\|_T < \epsilon], \tag{B13}$$

where $\mathbb{E}[\cdots \mid \|w\|_T < \epsilon]$ converges to 1 as $\epsilon \to 0$.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

In particular, the exponentiated average of the cross term $f \cdot dw$ in Eq. (B11) tends away from 1 as follows (Ikeda and Watanabe, 1981):

$$\mathbb{E}\left[\exp\left(\int_0^T f(w_t+\phi(t))\cdot dw_t\right) \mid \|w\|_T < \epsilon\right] \xrightarrow{\epsilon\to 0} \exp\left[-\frac{1}{2}\int_0^T \operatorname{div} f(\phi(t))dt\right], \tag{B14}$$

where w is the Wiener process and ϕ is a smooth curve. Roughly speaking, the weight of the tube should be modified from the appearance frequency of the smooth path ϕ, $\exp\left(-\int \frac{1}{2}|f-\dot{\phi}|^2 dt\right)$, by taking into account the weight in Eq. (B14).

For a rigorous derivation, see the proof of Theorem IV 9.1 of Ikeda and Watanabe (1981) or Section 2 of Zeitouni (1989).

Notice that Eq. (B14) serves as an evaluation formula for the divergence term along ϕ via ensemble calculation if we interpret the expectation as ensemble average:

$$\ln\mathbb{E}\left[\exp\left(\int_0^T f(w_t+\phi(t))\cdot dw_t\right) \mid \|w\|_T < \epsilon\right] \xrightarrow{\epsilon\to 0} -\frac{1}{2}\int_0^T \operatorname{div} f(\phi(t))dt. \tag{B15}$$

The ensemble can be generated by using the Wiener process limited to the small area $\|w\|_T < \epsilon$. Taking derivative of Eq. (B15) with respect to $\phi_i(t)$, we also get the formula for evaluating the derivative of the divergence term along ϕ as follows.

$$\frac{\mathbb{E}\left[\nabla f(\phi+w)\cdot dw \exp\left(\int_0^T f(\phi+w)\cdot dw\right) \mid \|w\|_T<\epsilon\right]}{\mathbb{E}\left[\exp\left(\int_0^T f(\phi+w)\cdot dw\right) \mid \|w\|_T<\epsilon\right]} \xrightarrow{\epsilon\to 0} -\frac{1}{2}\nabla(\operatorname{div} f)dt, \tag{B16}$$

where $(\nabla f(\phi+w), dw) = \sum_j \frac{\partial f_j(\phi+w)}{\partial \phi_i} dw_j$ can be calculated using the adjoint model $\nabla f(\phi+w)$. Although these evaluation formulas (B15) and (B16) illustrate the meaning of the divergence term, they seem too expensive to be used in the 4D-Var iterations.

Appendix C: Estimator for the divergence term

Cost functions in Eqs. (19) and (18) utilize the derivative of the drift term $f(x)$, and thus the gradient of the term contains the second derivative of $f(x)$, whose algebraic form is difficult to obtain in high-dimensional systems. Here, we propose an alternative form using Hutchinson's trace estimator (Hutchinson, 1990), which approximates the trace of matrix $\mathbb{E}[\xi^T A \xi] = \operatorname{tr}(A)$ using a stochastic vector whose components are independent identically distributed stochastic variables that take value ± 1 with probability 0.5.

A realization of the cost function is given as

$$\begin{aligned}\hat{J}_{\text{tube}}(\phi|y) = {} & \frac{1}{2\sigma_b^2}|\phi_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|\phi_m - y_m|^2 \\ & + \delta_t \sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{\phi_n-\phi_{n-1}}{\delta_t} - f(\phi_{n-1})\right|^2 + \frac{1}{2}\xi_{n-1}^T b^{-1}\left[f(\phi_{n-1}+b\xi_{n-1}) - f(\phi_{n-1})\right]\right),\end{aligned} \tag{C1}$$

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

where b is a small number. Notice $\hat{J}_{\text{tube}}(\phi|y)$ is a stochastic variable that satisfies

$$\mathbb{E}\left[\hat{J}_{\text{tube}}(\phi|y)\right] = J_{\text{tube}}(\phi|y). \tag{C2}$$

If the adjoint of f is at hand, the gradient of the stochastic cost function is given as

$$\begin{aligned}
\nabla_{\phi_n}\hat{J}_{\text{tube}}(\phi|y) = {} & \frac{1}{\sigma_b^2}(\phi_0 - x_b)\delta_{0,n} + \sum_{m\in M}\frac{1}{\sigma_o^2}(\phi_m - y_m)\delta_{m,n} \\
& + \frac{1}{\sigma^2}\left(\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-1})\right) \qquad (n>0) \\
& + \frac{\delta_t}{\sigma^2}\left(-\frac{1}{\delta_t} - \left(\frac{\partial f}{\partial \phi_n}(\phi_n)\right)^T\right)\left(\frac{\phi_{n+1} - \phi_n}{\delta_t} - f(\phi_n)\right) \qquad (n<N) \\
& + \frac{\delta_t}{2}\left[\left(\frac{\partial f}{\partial \phi_n}(\phi_n + b\xi_n)\right)^T b^{-1}\xi_n - \left(\frac{\partial f}{\partial \phi_n}(\phi_n)\right)^T b^{-1}\xi_n\right]. \qquad (n<N)
\end{aligned} \tag{C3}$$

The iterations similar to Eq. (25), $\phi^{(k+1)} = \phi^{(k)} - \alpha\nabla\hat{J}_{\text{tube}}$, will work.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. This work was partly supported by MEXT KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas JP15H05819.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-44
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 27 July 2017

References

Apte, A., Hairer, M., Stuart, A. M., and Voss, J.: Sampling the posterior: An approach to non-Gaussian data assimilation, Physica D: Nonlinear Phenomena, 230, 50–64, https://doi.org/10.1016/j.physd.2006.06.009, 2007.

Cotter, S. L., Roberts, G. O., Stuart, A., White, D., et al.: MCMC methods for functions: modifying old algorithms to make them faster, Statistical Science, 28, 424–446, 2013.

Daum, F.: Exact finite-dimensional nonlinear filters, IEEE Transactions on Automatic Control, 31, 616–622, 1986.

Doucet, A., Godsill, S., and Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering, Statistics and computing, 10, 197–208, 2000.

Dutra, D. A., Teixeira, B. O. S., and Aguirre, L. A.: Maximum a posteriori state path estimation: Discretization limits and their interpretation, Automatica, 50, 1360–1368, 2014.

Giering, R. and Kaminski, T.: Recipes for adjoint code construction, ACM Transactions on Mathematical Software (TOMS), 24, 437–474, 1998.

Hutchinson, M. F.: A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines, Communications in Statistics-Simulation and Computation, 19, 433–450, 1990.

Ikeda, N. and Watanabe, S.: Stochastic Differential Equations and Diffusion Processes, vol. 24 of *North-Holland Mathematical Library*, chap. VI.9, North-Holland, 1981.

Law, K., Stuart, A., and Zygalakis, K.: Data Assimilation, Springer, 2015.

Malsom, P. J. and Pinski, F. J.: Role of Ito's lemma in sampling pinned diffusion paths in the continuous-time limit, Phys. Rev. E, 94, 042 131, https://doi.org/10.1103/PhysRevE.94.042131, http://link.aps.org/doi/10.1103/PhysRevE.94.042131, 2016.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equation of state calculations by fast computing machines, The Journal of Chemical Physics, 21, 1087–1092, 1953.

Onsager, L. and Machlup, S.: Fluctuations and irreversible processes, Physical Review, 91, 1505, 1953.

Roberts, G. O. and Rosenthal, J. S.: Optimal scaling of discrete approximations to Langevin diffusions, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60, 255–268, 1998.

Rössler, O.: An equation for continuous chaos, Physics Letters A, 57, 397–398, https://doi.org/http://dx.doi.org/10.1016/0375-9601(76)90101-8, http://www.sciencedirect.com/science/article/pii/0375960176901018, 1976.

Stuart, A. M., Voss, J., and Wilberg, P.: Conditional Path Sampling of SDEs and the Langevin MCMC Method, Commun. Math. Sci., 2, 685–697, http://projecteuclid.org/euclid.cms/1109885503, 2004.

Trémolet, Y.: Accounting for an imperfect model in 4D-Var, Quarterly Journal of the Royal Meteorological Society, 132, 2483–2504, https://doi.org/10.1256/qj.05.224, http://dx.doi.org/10.1256/qj.05.224, 2006.

Zeitouni, O.: On the Onsager–Machlup functional of diffusion processes around non C^2 curves, The Annals of Probability, pp. 1037–1054, 1989.

Zinn-Justin, J.: Quantum Field Theory and Critical Phenomena, chap. 4.6, Oxford University Press, 4th edn., 2002.

Zupanski, D.: A general weak constraint applicable to operational 4DVAR data assimilation systems, Monthly Weather Review, 125, 2274–2292, 1997.