The author sincerely appreciates the 2nd referee's careful reading of the manuscript; the encouraging suggestions are also appreciated. The author's responses to the referee's comments are as follows:

1. *I find the abstract to be a bit vague and think you may want to state more clearly what is done in the paper. For example mention that you implement 4DVAR. The way its written makes it unclear if youre referring to things covered in the paper or in the literature.*

As the reviewer pointed out, the abstract did not properly represent what the main text described. I have modified the abstract to emphasise 4D-Var and the divergence term. (see abstract in the revised ms)

2. *You mention later on that the methods involving the divergence of f would be difficult to implement for large dimension and 4DVAR due to the need to compute the derivative of div f. I think it would be interesting if you could mention this when you introduce your models, or show the gradient of the cost function for 4DVAR and highlight the need of this extra term. E.g. after Eq 25 you could have a "where delJ_tube = ... (26)"*

According to the reviewer's suggestion, a note on the gradient has been added. (Page 9 Lines 1–7)

3. *P5L23 'You can see'* $\rightarrow$ *'It is clear'*

Done. (Page 10 Line 7)

4. *Yaxis labels in Fig 1 are missing and a number at the origin is not clear.*

Done. (see Fig. 1 in the revised ms)

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~~ it should be incorporated into cost functions for discrete-time estimation problems. The divergence of the drift term is essential ~~, but for the path probability assignments in combination~~ in weak-constraint 4D-Var (w4D-Var), but it is not necessary in Markov chain Monte Carlo with the Euler ~~time-discretization scheme, it is not necessary. The latter property will help simplify the implementation of nonlinear data assimilation for large systemswith sampling methods such as the Metropolis-adjusted Langevin algorithm~~scheme. Although the former property may cause difficulties when implementing w4D-Var in large systems, a new technique is proposed for estimating the divergence term and its derivative.

1 Introduction

In traditional weak-constraint 4D-Var ~~setting (e.g., Zupanski, 1997; Trémolet, 2006)~~settings (e.g. Zupanski, 1997; Trémolet, 2006), a quadratic cost function is defined as the negative logarithm of the probability for each ~~Brownian~~ sample path, which is suitable for path sampling ~~(e.g., Zinn-Justin, 2002). The optimization~~ (e.g. Zinn-Justin, 2002). The optimisation problem is naively described as finding the most probable path by ~~minimizing~~ minimising the quadratic cost function. However, the term ~~"~~ ‘the most probable path~~"~~ ’ does not make sense in this context, ~~since~~ because 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~~ optimisation problem properly, one should introduce a macroscopic variable ~~ϕ~~ $\phi=\phi(t)$ that represents a smooth curve, and introduce a measure ~~μ that accounts how dense the paths that lie~~ that accounts for how densely the paths are populated in the ϵ~~-neighbor centered~~ -neighbourhood centred 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.~~ ‘the tube’. Then the problem is defined as finding the most probable tube ϕ, which represents the maximum a posteriori (MAP) estimate of the path distribution. Mathematicians pioneering the theory of SDE ~~(e.g., Ikeda and Watanabe, 1981; Zeitouni, 1989) were already~~ (e.g. Ikeda and Watanabe, 1981; Zeitouni, 1989) have been aware of this subtle point since the 1980s, and established the proper form of the cost function as the Onsager–Machlup (OM) functional (Onsager and Machlup, 1953) for the ~~most probable tube.~~ tube.

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

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.

Before moving on to its applications, here we review the concept of the OM functional. To make presentation simple, we assume that $D = 1$ and $\sigma = 1$, and concentrate on the formulation of the prior distribution in the subsequent two sections 1.1 and 1.2.

1.1 OM functional for path sampling

The model equation (1) is discretised with the Euler scheme (with the drift term at the previous time) as

$$x_n = x_{n-1} + f(x_{n-1})\delta_t + \xi_{n-1}, \quad 1 \le n \le N, \tag{4}$$

where δ_t is the time increment, and each ξ_{n-1} obeys $\mathcal{N}(0, \delta_t)$. Equation (4) can be considered as a nonlinear mapping $F_1 : \xi \mapsto x$, from the noise vector $\xi = (\xi_0, \xi_1, \cdots, \xi_{N-1})^T$ to the state vector $x = (x_1, x_2, \cdots, x_N)^T$. The inverse of the mapping is linearised as

$$\begin{bmatrix} \delta\xi_0 \\ \delta\xi_1 \\ \vdots \\ \delta\xi_{N-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 - \delta_t f'(x_1) & 1 & & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & -1 - \delta_t f'(x_{N-1}) & 1 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \vdots \\ \delta x_N \end{bmatrix}, \tag{5}$$

whose Jacobian is $DF_1^{-1} = |d\xi/dx| = 1$.

It is also discretised with the trapezoidal scheme (with the drift term at the midpoint) as

$$x_n = x_{n-1} + \frac{f(x_n) + f(x_{n-1})}{2}\delta_t + \xi_{n-1}, \quad 1 \le n \le N, \tag{6}$$

which defines a mapping $F_2 : \xi \mapsto x$. The inverse of the mapping is linearised as

$$\begin{bmatrix} \delta\xi_0 \\ \delta\xi_1 \\ \vdots \\ \delta\xi_{N-1} \end{bmatrix} = \begin{bmatrix} 1-\frac{\delta_t}{2}f'(x_1) & 0 & \cdots & 0 & 0 \\ -1-\frac{\delta_t}{2}f'(x_1) & 1-\frac{\delta_t}{2}f'(x_2) & & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & -1-\frac{\delta_t}{2}f'(x_{N-1}) & 1-\frac{\delta_t}{2}f'(x_N) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \vdots \\ \delta x_N \end{bmatrix}, \tag{7}$$

whose Jacobian is $DF_2^{-1} = |d\xi/dx| = \prod_{n=1}^{N}[1-(\delta_t/2)f'(x_n)] \fallingdotseq \exp\left[-(\delta_t/2)\sum_{n=1}^{N} f'(x_n)\right]$.

Generally, we can assign a measure μ_0 to a cylinder set $\hat{\Omega} \equiv \hat{\Omega}_0 \times \hat{\Omega}_1 \times \cdots \times \hat{\Omega}_{N-1}$ in the noise space using a density g as follows.

$$\mu_0(\hat{\Omega}) = \int_{\hat{\Omega}_0} d\xi_0 \int_{\hat{\Omega}_1} d\xi_1 \cdots \int_{\hat{\Omega}_{N-1}} d\xi_{N-1} g(\xi_0, \xi_1, \cdots, \xi_{N-1}) = \int_{\hat{\Omega}} g(\xi)\lambda(d\xi) = \int_{\hat{\Omega}} \mu_0(d\xi), \tag{8}$$

where λ is the Lebesgue measure on $\mathbb{R}^N$. In our case, we can regard that a small area $d\xi$ in the noise space is equipped with a measure:

$$\mu_0(d\xi) = g(\xi)\lambda(d\xi), \quad g(\xi) \equiv \frac{1}{(2\pi\delta_t)^{N/2}} \mathrm{e}^{-\frac{1}{2\delta_t}\sum_{n=1}^{N}\xi_{n-1}^2}. \tag{9}$$

Suppose we have a cylinder set $\Omega \equiv \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$ in the state space, where each $\Omega_n \subset \mathbb{R}^1$ is on time slice $t = n\delta_t$. Now, the mapping F_1 (or F_2) induces a measure through the change-of-variables from ξ to x with respect to the measure μ_0 as

$$\mu_i(\Omega) = \int_{\Omega_1} dx_1 \int_{\Omega_2} dx_2 \cdots \int_{\Omega_N} dx_N (g \circ F_i^{-1})(x_1, x_2, \cdots, x_N) DF_i^{-1} = \int_{\Omega} \mu_i(dx), \quad i = 1, 2. \tag{10}$$

In our case, each mapping assigns the following measure to a small area dx in the corresponding state space:

$$\mu_1(dx) \equiv g(F_1^{-1}(x))DF_1^{-1}\lambda(dx) = \frac{1}{(2\pi\delta_t)^{N/2}} \mathrm{e}^{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left(\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right)^2} \lambda(dx), \tag{11}$$

$$\mu_2(dx) \equiv g(F_2^{-1}(x))DF_2^{-1}\lambda(dx) = \frac{1}{(2\pi\delta_t)^{N/2}} \mathrm{e}^{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left[\left(\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1/2})\right)^2 + f'(x_n)\right]} \lambda(dx), \tag{12}$$

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

These measures μ_1 and μ_2 represent the occurrence probability of the noise seen from the state space, and thus can be used for path sampling.

1.2 OM functional for mode estimate

If we perform path sampling with sufficient number, in theory we can find the mean of distribution via averaging the samples, or the mode of distribution via organising them into a histogram. Still, in some practical applications, we must efficiently find the mode of distribution via variational methods; computationally, this approach is much cheaper than path sampling. For that purpose, we are tempted to use a quadratic cost function for the minimisation. However, we can illustrate a simple example against minimising the path probability (11) to obtain the mode of distribution. Suppose we have a discrete-time stochastic system in $\mathbb{R}^1$, starting from $x_0 = 0$, and we move forward two time steps:

$$x_1 = x_0 + x_0^2\delta_t + \xi_0 = \xi_0, \qquad x_2 = x_1 + x_1^2\delta_t + \xi_1 = \xi_0 + \xi_0^2\delta_t + \xi_1, \tag{13}$$

where ξ_0 and ξ_1 obey independent normal distributions $\mathcal{N}(0, \delta_t)$. It may be seen as a discrete version of $dx_t = x_t^2 dt + dw_t$. It is easy to notice that the mode of distribution (x_1, x_2) is not $(0, 0)$ owing to the nonlinear term $\xi_0^2\delta_t$. However, according to the path probability (11):

$$\mu_1(dx_1 dx_2) \propto \exp\left[-\frac{\delta_t}{2}\left(\left(\frac{x_1 - x_0}{\delta_t} - x_0^2\right)^2 + \left(\frac{x_2 - x_1}{\delta_t} - x_1^2\right)^2\right)\right]\lambda(dx_1 dx_2),$$

the best trajectory is $(x_1, x_2) = (0, 0)$, which has no noise $(\xi_0, \xi_1) = (0, 0)$. We expect a path with the highest probability at $(x_1, x_2) = (0, 0)$, but it is not the route where the paths are most concentrated.

Motivated by this example, we shall investigate a proper strategy to find the route that maximises the density of paths. In this regard, we ask how densely the paths populate in the small neighbourhood of a curve $\phi = \phi(t)$ in state space.

Assuming that f and ϕ are twice continuously differentiable, we evaluate the density of paths in the ϵ-neighbourhoods around a curve ϕ connecting points $\{\phi_n, n = 1, 2, \cdots N\}$ with the following integral:

$$I_{\epsilon,\delta_t}(\phi) = \int_{\phi_1-\epsilon}^{\phi_1+\epsilon} dx_1 \int_{\phi_2-\epsilon}^{\phi_2+\epsilon} dx_2 \cdots \int_{\phi_N-\epsilon}^{\phi_N+\epsilon} dx_N \frac{1}{(2\pi\delta_t)^{N/2}} \exp\left\{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left(\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right)^2\right\} \tag{14}$$

$$= \int_{-\epsilon}^{\epsilon} dy_1 \int_{-\epsilon}^{\epsilon} dy_2 \cdots \int_{-\epsilon}^{\epsilon} dy_N \frac{1}{(2\pi\delta_t)^{N/2}} \exp\left\{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left(\frac{y_n - y_{n-1}}{\delta_t} + \frac{\phi_n - \phi_{n-1}}{\delta_t} - f(y_{n-1} + \phi_{n-1})\right)^2\right\} \tag{15}$$

$$= \int_{-\epsilon}^{\epsilon} dy_1 \int_{-\epsilon}^{\epsilon} dy_2 \cdots \int_{-\epsilon}^{\epsilon} dy_N \frac{1}{(2\pi\delta_t)^{N/2}} \exp\left\{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left(\frac{y_n - y_{n-1}}{\delta_t}\right)^2\right\}$$
$$\times \exp\left\{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left[\left(\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(y_{n-1} + \phi_{n-1})\right)^2 + 2\left(\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(y_{n-1} + \phi_{n-1})\right)\left(\frac{y_n - y_{n-1}}{\delta_t}\right)\right]\right\}. \tag{16}$$

By regarding y_n in Eq. (16) as being generated according to the probability $\frac{1}{(2\pi\delta_t)^{N/2}} \mathrm{e}^{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left(\frac{y_n - y_{n-1}}{\delta_t}\right)^2}$, we can interpret the integration as a weighted ensemble averaging of a random function up to a numerical constant. The sequence y_n can be set as a random walk $y_0 = 0$, $y_n = \sum_{k=1}^{n} \xi_k$, where ξ_k are independent normal random variables obeying $\mathcal{N}(0, \delta_t)$. For simplicity, we rather assume that ξ_k takes values $\pm\sqrt{\delta_t}$ with 0.5 probability for either one, because Donsker's theorem ensures it has the same probability law as the former for a large ensemble. We suppose $\sqrt{\delta_t} < \epsilon$ so that no step of the random walk escapes from the ϵ-neighbourhood. Accordingly, the integral is expressed as the ensemble average with respect to random walks confined in the tube $[-\epsilon,\epsilon]$:

$$I_{\epsilon,\delta_t}(\phi) \propto \mathbb{E}_y\left[\mathrm{e}^{-J(\phi,y)}\middle|(\forall n)\ \ |y_n|<\epsilon\right], \tag{17}$$

$$J(\phi,y) \equiv -\frac{\delta_t}{2}\sum_{n=1}^{N}\left[\left(\frac{\phi_n-\phi_{n-1}}{\delta_t}-f(y_{n-1}+\phi_{n-1})\right)^2+2\left(\frac{\phi_n-\phi_{n-1}}{\delta_t}-f(y_{n-1}+\phi_{n-1})\right)\left(\frac{y_n-y_{n-1}}{\delta_t}\right)\right] \tag{18}$$

where $\mathbb{E}_y$ denotes the ensemble averaging of the random walks denoted by y, each of which follows the route $(y_0,y_1,\cdots,y_N)$, and satisfies $|y_n|<\epsilon$ for all n.

Because y_{n-1} is small, we can apply the expansion:

$$f(y_{n-1}+\phi_{n-1}) = f(\phi_{n-1})+f'(\phi_{n-1})y_{n-1}+O(y^2), \tag{19}$$

where f' is the derivative of f. Let us accept that the following average containing the higher order terms $O(y^2)$ converges (see Eq. (B20)).

$$\mathbb{E}_y\left[\mathrm{e}^{\sum_{n=1}^{N}O(y^2)(y_n-y_{n-1})}\middle|(\forall n)\ \ |y_n|<\epsilon\right]\xrightarrow{\epsilon\to 0} 1. \tag{20}$$

As shown in Appendix B, the remaining terms in the exponent $-J(\phi,y)$ are less than $O(\epsilon)$ except the following one.

$$\sum_{n=1}^{N}f'(\phi_{n-1})y_{n-1}(y_n-y_{n-1}) = \sum_{n=1}^{N}f'(\phi_{n-1})\left[\frac{1}{2}(y_{n-1}-y_n)+\frac{1}{2}(y_{n-1}+y_n)\right](y_n-y_{n-1}) \tag{21}$$

$$= \sum_{n=1}^{N}f'(\phi_{n-1})\frac{1}{2}(y_{n-1}-y_n)(y_n-y_{n-1})+\sum_{n=1}^{N}f'(\phi_{n-1})\frac{1}{2}(y_n^2-y_{n-1}^2) \tag{22}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}f'(\phi_{n-1})\xi_n^2+\frac{1}{2}\sum_{n=1}^{N-1}\left[f'(\phi(t_{n-1}))-f'(\phi(t_{n-1}+\delta_t))\right]y_n^2+\frac{1}{2}f'(\phi_{N-1})y_N^2 \tag{23}$$

$$= -\frac{\delta_t}{2}\sum_{n=1}^{N}f'(\phi_{n-1})+O(\epsilon^2). \qquad \because \xi_n=\pm\sqrt{\delta_t},\quad f'(\phi(t_{n-1}))-f'(\phi(t_{n-1}+\delta_t))=O(\delta_t),\quad y_n^2<\epsilon^2. \tag{24}$$

Consequently, we obtain the asymptotic expression for the ensemble average when ϵ is small and $\delta_t<\epsilon^2$:

$$I_{\epsilon,\delta_t}(\phi) \propto \mathbb{E}_y\left[\mathrm{e}^{-\frac{\delta_t}{2}\sum_{n=1}^{N}\left[\left(\frac{\phi_n-\phi_{n-1}}{\delta_t}-f(\phi_{n-1})\right)^2+f'(\phi_{n-1})\right]+O(\epsilon)+\sum_{n=1}^{N}O(y^2)(y_n-y_{n-1})}\middle|(\forall n)\ \ |y_n|<\epsilon\right] \tag{25}$$

$$\to \mathrm{e}^{-\frac{1}{2}\int_0^T\left[\left(\dot\phi(t)-f(\phi(t))\right)^2+f'(\phi(t))\right]dt}. \tag{26}$$

Appendix B2 shows that a similar form is available for multi-dimensional processes, except the term $f'(\phi(t))$ is promoted to $\mathrm{div}\, f(\phi(t))$.

Importantly, the control variable for the optimisation has changed from x to ϕ.

1.3 Probabilistic description of data assimilation

Using the OM functional derived in sections 1.1 and 1.2 as model error term, we shall develop a probabilistic description of data assimilation.

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

$$P(x|y) \propto P(x)P(y|x) = P(x|x_0)P(x_0)P(y|x) = \prod_{n=1}^{N} P(x_n|x_{n-1})P(x_0) \prod_{m\in M} P(y_m|x_m). \tag{27}$$

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

$$P(x_0) \propto \exp\left(-\frac{|x_0 - x_b|^2}{2\sigma_b^2}\right), \tag{28}$$

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

$$P(y_m|x_m) \propto \exp\left(-\frac{|y_m - x_m|^2}{2\sigma_o^2}\right). \tag{29}$$

~~Now, we move on to approximation with a discrete time step. The change-of-measure argument (Appendix~~

Based on the argument in section ~~B1) or the path integral argument (e.g., Zinn-Justin, 2002) on a path of this stochastic process shows that~~ 1.1, Eq. (~~1~~4) has the transition probability at discrete time steps

$$P(x_n|x_{n-1}) \propto \exp\left(-\frac{\delta_t}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2\right), \tag{30}$$

called the Euler scheme, which uses the drift $f(x_{n-1})$ at the previous time step. ~~This~~ Section 1.1 also shows that this transition probability has another expression~~(see the derivation of Eq. (B8) in Appendix B1 or Zinn-Justin (2002)):~~ :

$$P(x_n|x_{n-1}) \propto \exp\left(-\frac{\delta_t}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2 - \frac{\delta_t}{2}\operatorname{div} f(x_n)\right), \tag{31}$$

$$f(x_{n-\frac{1}{2}}) \equiv \frac{f(x_n) + f(x_{n-1})}{2}, \quad \operatorname{div} f(x) \equiv \sum_{i=1}^{D}\frac{\partial f^i}{\partial x^i}(x), \tag{32}$$

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): where “~~(e.g. Zinn-Justin, 2002):

$$P(x|x_0) \propto \exp\left(-\delta_t \sum_{n=1}^{N}\frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2\right) \tag{33}$$

$$\leftrightharpoons \exp\left(-\delta_t \sum_{n=1}^{N}\left[\frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\operatorname{div} f(x_n)\right]\right), \tag{34}$$

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~~section ~~B2~~1.2, we can also define the probability $P(U_\phi|\phi_0)$ for a smooth tube that represents its ~~neighboring paths $U_\phi = \{\omega \mid (\forall n)|\phi_n - x_n(\omega)| < \epsilon\}$:~~ neighbouring paths $U_\phi = \{\omega|(\forall n)|\phi_n - x_n(\omega)| < \epsilon\}$:

$$P(U_\phi|\phi_0) \propto \exp\left(-\delta_t \sum_{n=1}^{N}\left[\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-1})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-1})\right]\right). \tag{35}$$

The scaling argument for a smooth curve in Appendix A allows us to use the drift term $\cancel{f(\phi_{n-1})}$ $f(\phi_{n-\frac{1}{2}})$ instead in Eq. (~~36):~~ 35):

$$P(U_\phi|\phi_0) \propto \exp\left(-\delta_t \sum_{n=1}^{N}\left[\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-\frac{1}{2}})\right]\right). \tag{36}$$

The corresponding posterior probabilities are thus given as follows: ~~for a Brownian path~~

$$P_{\text{path}}(x|y) \propto \exp\left(-J_{\text{path}}(x|y)\right), \tag{37}$$

$$J_{\text{path}}(x|y) \equiv \frac{1}{2\sigma_b^2}|x_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|x_m - y_m|^2 + \delta_t\sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-1})\right|^2\right) \tag{38}$$

$$\fallingdotseq \frac{1}{2\sigma_b^2}|x_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|x_m - y_m|^2 + \delta_t\sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{x_n - x_{n-1}}{\delta_t} - f(x_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(x_n)\right) \tag{39}$$

for a sample path, and

$$P_{\text{tube}}(U_\phi|y) \propto P(U_\phi|\phi_0)P(\phi_0)P(y|U_\phi) \propto \exp\left(-J_{\text{tube}}(\phi|y)\right), \tag{40}$$

$$J_{\text{tube}}(\phi|y) \equiv \frac{1}{2\sigma_b^2}|\phi_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|\phi_m - y_m|^2 + \delta_t\sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-\frac{1}{2}})\right) \tag{41}$$

$$\fallingdotseq \frac{1}{2\sigma_b^2}|\phi_0 - x_b|^2 + \sum_{m\in M}\frac{1}{2\sigma_o^2}|\phi_m - y_m|^2 + \delta_t\sum_{n=1}^{N}\left(\frac{1}{2\sigma^2}\left|\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-1})\right|^2 + \frac{1}{2}\mathrm{div}\, f(\phi_{n-1})\right) \tag{42}$$

for a smooth tube. Note that different pairs of ~~time-discretization~~ time-discretisation 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. (38), (39), (41), and (42).

2 Method

2.1 Four schemes for OM

In the argument in ~~Section~~sections ~~1~~1.1 and 1.2, 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)~~. As a proof-of-concept described in these sections, we will test all the cases with conceivable combinations of the timing of the drift term $f(x_t)$ and the presence or absence of the divergence term. Including those shown in Eqs. (38), (39), (41), and (42), as well as those that are potentially incorrect, the possible candidates for the ~~discretization~~ discretisation schemes of the OM functional ~~would be as follows:~~ are as follows, where the symbol ψ represents either ϕ for a smooth curve or x for a sample path.

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

$$\widetilde{OM}_{\mathrm{E}} \equiv \frac{1}{2\sigma^2}\left|\frac{\psi_n - \psi_{n-1}}{\delta_t} - f(\psi_{n-1})\right|^2; \tag{43}$$

2. Euler scheme with divergence term (ED): ~~where $f(x_{n-\frac{1}{2}}) = (f(x_n) + f(x_{n-1}))/2$;~~

$$\widetilde{OM}_{\mathrm{ED}} \equiv \frac{1}{2\sigma^2}\left|\frac{\psi_n - \psi_{n-1}}{\delta_t} - f(\psi_{n-1})\right|^2 + \frac{1}{2}\operatorname{div} f(\psi_{n-1}); \tag{44}$$

3. Trapezoidal scheme (T):

$$\widetilde{OM}_{\mathrm{T}} \equiv \frac{1}{2\sigma^2}\left|\frac{\psi_n - \psi_{n-1}}{\delta_t} - f(\psi_{n-\frac{1}{2}})\right|^2; \tag{45}$$

4. Trapezoidal scheme with divergence term (TD) ~~(e.g., Ikeda and Watanabe, 1981; Apte et al., 2007; Dutra et al., 2014): where $f(x_{n-\frac{1}{2}}) = (f(x_n) + f(x_{n-1}))/2$~~(e.g. Ikeda and Watanabe, 1981; Apte et al., 2007; Dutra et al., 2014):

$$\widetilde{OM}_{\mathrm{TD}} \equiv \frac{1}{2\sigma^2}\left|\frac{\psi_n - \psi_{n-1}}{\delta_t} - f(\psi_{n-\frac{1}{2}})\right|^2 + \frac{1}{2}\operatorname{div} f(\psi_{n-\frac{1}{2}}), \tag{46}$$

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

~~By using the cost function adopted in~~

2.2 Data assimilation algorithms

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

$$x^{(k+1)} = x^{(k)} - \alpha\nabla J_{\text{path}} + \sqrt{2\alpha}\,\xi, \quad \alpha > 0,\ \xi \sim \mathcal{N}(0,1)^{D(N+1)},\ \nabla J = \left(\frac{\partial J}{\partial x}\right)^T \tag{47}$$

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

$$\phi^{(k+1)} = \phi^{(k)} - \alpha\nabla J_{\text{tube}}, \quad \alpha > 0. \tag{48}$$

Note that if the OM functional of type $\widetilde{OM}_{\text{ED}}$ is used, the gradient is of the form:

$$\begin{aligned}
\nabla_{\phi_n} J_{\text{tube}} &= \frac{1}{\sigma_b^2}(\phi_0 - x_b)\delta_{0,n} + \sum_{m\in M}\frac{1}{\sigma_o^2}(\phi_m - y_m)\delta_{m,n} \\
&\quad + \frac{1}{\sigma^2}\left(\frac{\phi_n - \phi_{n-1}}{\delta_t} - f(\phi_{n-1})\right) \qquad (n > 0) \\
&\quad + \frac{\delta_t}{\sigma^2}\left(-\frac{1}{\delta_t} - \left(\frac{\partial f}{\partial \phi_n}(\phi_n)\right)^T\right)\left(\frac{\phi_{n+1} - \phi_n}{\delta_t} - f(\phi_n)\right) + \frac{\delta_t}{2}\frac{\partial}{\partial \phi_n}\operatorname{div} f(\phi_n) \qquad (n < N),
\end{aligned} \tag{49}$$

where $\left(\frac{\partial f}{\partial \phi_n}(\phi_n)\right)^T$ is an adjoint integration starting from the subsequent term, which is typical in gradient calculations in 4D-Var. In comparison, the term $\frac{\partial}{\partial \phi_n}\operatorname{div} f(\phi_n)$ requires the second derivative of f, which is not typical in 4D-Var, and could be difficult to implement in large dimensional systems.

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 correct answer'. The reference solution that approximates the ~~right~~ correct answer is provided by a ~~naive~~ particle smoother (PS) ~~(e.g., Doucet et al., 2000)~~(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:

$$P(x) \simeq \frac{1}{M}\sum_{m=1}^{M}\prod_{n=0}^{N}\delta(x_n - \chi_n^{(m)}), \tag{50}$$

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

2. Reweight it according to Bayes' ~~s~~ theorem:

$$P(y|x) \propto \exp\left(-\frac{1}{2\sigma_o^2}|y - x_N|^2\right), \tag{51}$$

$$P(x|y) = \frac{P(x)P(y|x)}{\int dx P(x)P(y|x)} = \sum_{m=1}^{M}\prod_{n=0}^{N}\delta(x_n - \chi_n^{(m)})\frac{w^{(m)}}{\sum_{m=1}^{M} w^{(m)}}, \tag{52}$$

$$w^{(m)} \equiv \exp\left(-\frac{1}{2\sigma_o^2}|y - \chi_N^{(m)}|^2\right). \tag{53}$$

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

$$dx_t = \tanh(x_t)dt + dw_t, \quad x_{t=0} \sim \mathcal{N}(0, 0.16), \tag{54}$$

subject to an observation y:

$$y | x_{t=5} \sim \mathcal{N}(x_{t=5}, 0.16), \quad y = 1.5. \tag{55}$$

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

Figure 1 shows the probability densities of paths ~~normalized~~ normalised 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~~ It is clear 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, as shown in Fig. 2. These schemes correspond to candidates in Eqs. (38) and (39). The expected path by ED bends ~~toward~~ towards a larger x, which should be caused by an extra penalty for a larger x. The expected path by T bends ~~toward~~ towards a smaller x, which should be caused by the lack of ~~penalty for~~ a penalty for a larger x.

The results of 4D-Var, which represents ~~the maximum a posteriori (MAP) estimatesof the tube~~the MAP estimates, are shown in Fig. 3. ED and TD provide the proper MAP estimate~~of the tube~~. These schemes correspond to candidates in Eqs. (41) and (42). The expected paths by E and T bend ~~toward~~ towards a smaller ϕ, which should be caused by the lack of ~~penalty for~~ a penalty for a larger ϕ.

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

$$\begin{cases} dx_1 &= (-x_2 - x_3)dt + \sigma dw_1, \\ dx_2 &= (x_1 + a x_2)dt + \sigma dw_2, \\ dx_3 &= (b + x_1 x_3 - c x_3)dt + \sigma dw_3, \end{cases} \tag{56}$$

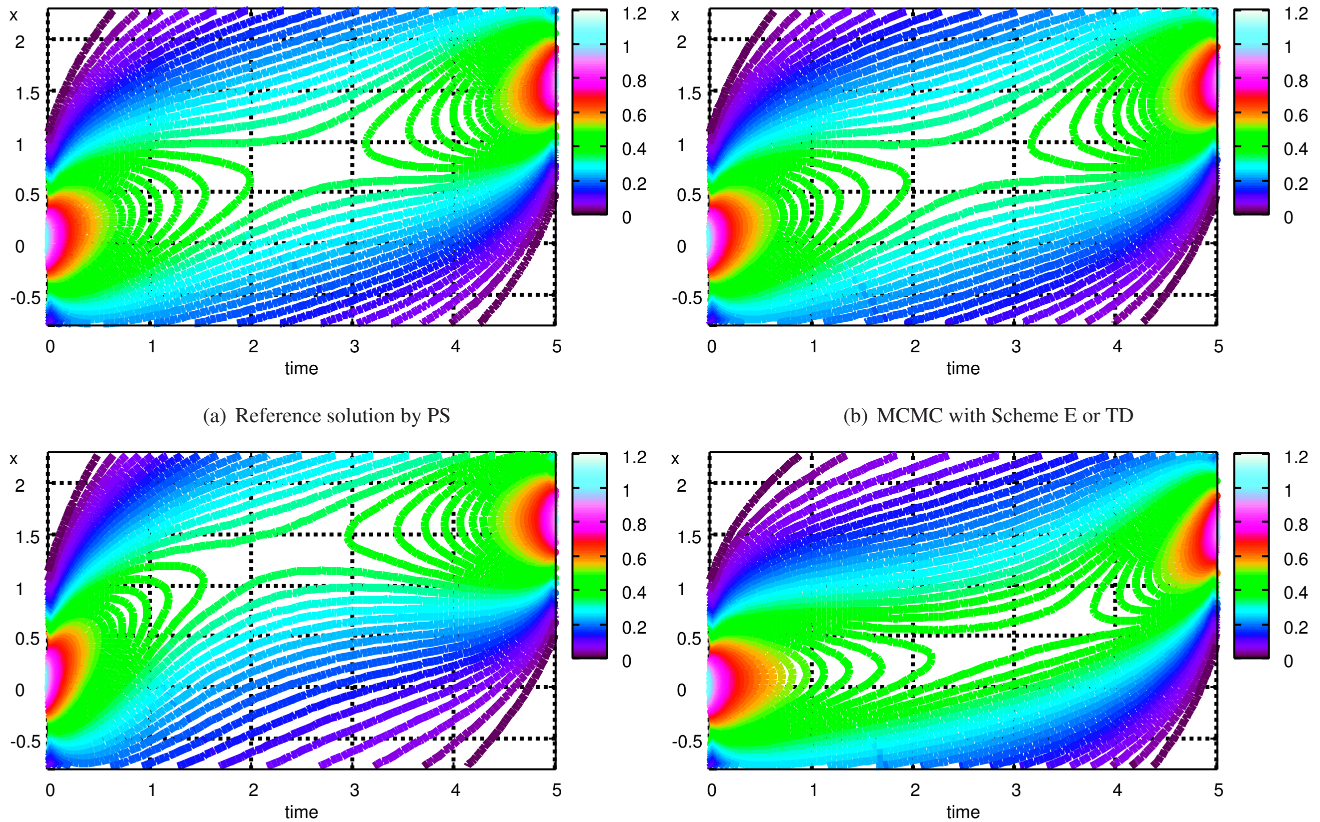
$$x_{t=0} \sim \mathcal{N}(x_b, 0.04I), \tag{57}$$

subject to an observation y:

$$y | x_{t=0.4} \sim \mathcal{N}(x_{t=0.4}, 0.04I), \tag{58}$$

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, $\operatorname{div} f(x) = x_1 + a - c$ 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 ~~$\hat{\phi} = \operatorname{argmax}_\phi P[\phi | Y]$~~ $\hat{\phi} = \operatorname{argmax}_\phi P(\phi$ where U represents the tube ~~centered~~ centred at ϕ with radius 0.03.



(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.

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~~.

3.3 ~~Toward~~ Towards 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)~~symb 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~~ utilising Hutchinson's trace estimator (Hutchinson, 1990) (See Appendix C). It is also

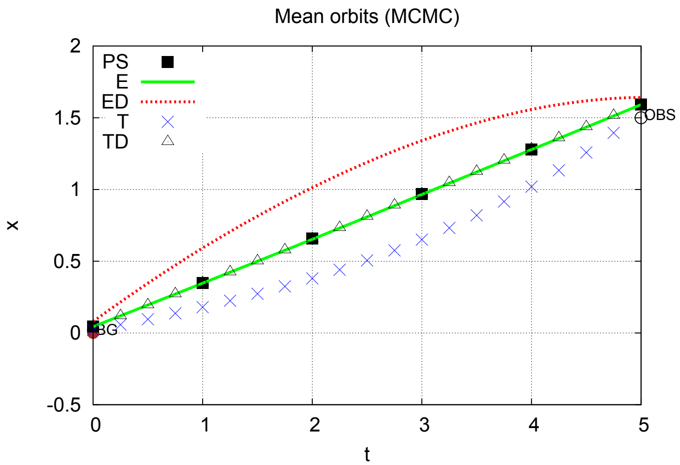


Figure 2. Expected path derived by MCMC (hyperbolic model).

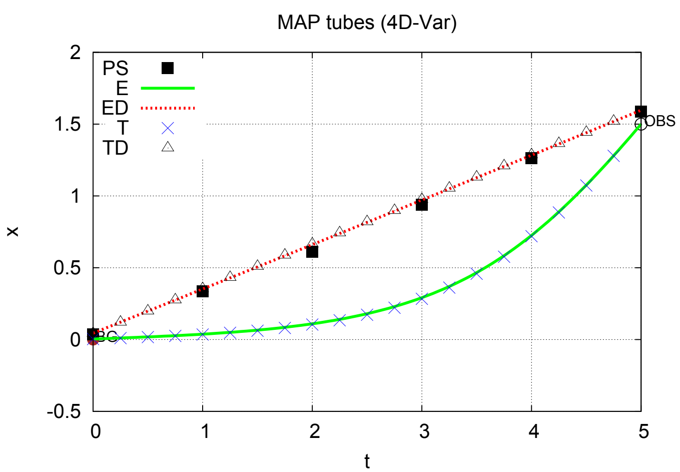


Figure 3. Most probable tube derived by 4D-Var (hyperbolic model).

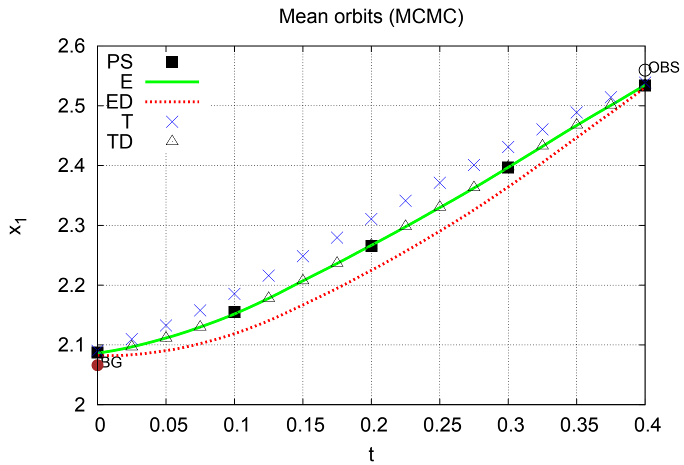


Figure 4. Expected path derived by MCMC (Rössler model).

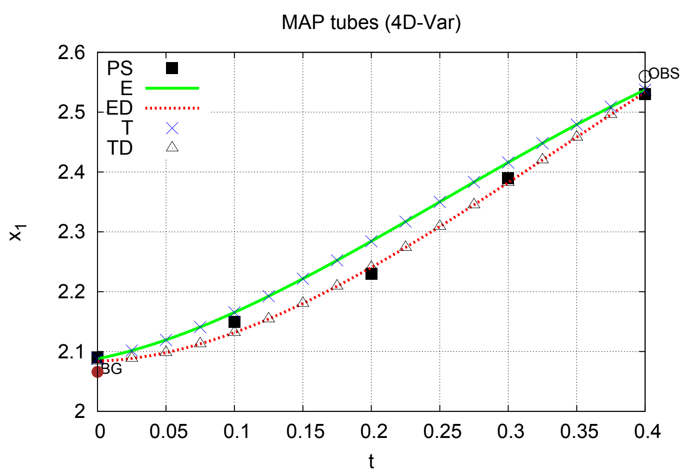


Figure 5. Most probable tube derived by 4D-Var (Rössler model).

Table 1. Applicable OM schemes

		with div (f)	without div (f)
Sampling by MCMC	Euler scheme		✓
	trapezoidal scheme	✓	
MAP estimate by 4D-Var	Euler scheme	✓	
	trapezoidal scheme	✓	

clear that the Euler scheme without the divergence term is more convenient for ~~implementation of~~ implementing path sampling, because it does not require cumbersome calculation of the divergence term.

4 Conclusions

We examined several ~~discretization~~ discretisation 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

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

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~~ lists the discretisation schemes which were found to be applicable, i.e. those expected to converge to the same result as the reference solution. 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)~~(e.g. Apte et al., 2007; Malsom and Pinski, 2016;

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

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

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~~ utilising 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 $x_{n-\frac{1}{2}}$~~ $f(\psi_{n-1})$ term around $\psi_{n-\frac{1}{2}}$ in scheme E gives

$$\widetilde{OM} \simeq \sum_{n=1}^{N} \delta_t \left\{ \sigma^{-2} \left[\frac{\psi_n - \psi_{n-1}}{\delta_t} - f(\psi_{n-\frac{1}{2}}) - (\psi_n - \psi_{n-1}) \frac{\partial f}{\partial x}(\psi_{n-\frac{1}{2}}) \right]^2 + \mathrm{div}\,(f) \right\}$$

$$= \delta_t \left\{ \sigma^{-2} (\text{noise} + \text{shift})^2 + \text{divergence} \right\}.$$

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

where we assume order-one fluctuations: ~~$\sigma = O(1)$.~~ $\sigma = O(1)$, and the symbol ψ represents either ϕ for a smooth curve or x for a sample path.

For a sample path of the stochastic process, the scaling ~~$x_n - x_{n-1} = O(\delta_t^{\frac{1}{2}})$~~ $\psi_n - \psi_{n-1} = O(\delta_t^{\frac{1}{2}})$, which 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 $x_n - x_{n-1} = O(\delta_t)$~~ In the case of a smooth curve, there is no stochastic term, and thus $\psi_n - \psi_{n-1}$ is the product of a bounded function $f(\psi_{n-1})$ and δ_t, which results in a value with $O(\delta_t)$. This 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 a trapezoidal scheme

Consider two stochastic processes (cf., ~~Section~~ 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,

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

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

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

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

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

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

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

and from Eq. (B6)

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

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~~ weight is assigned 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)~~(Ikeda and Watanabe, 1981; 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~~ This means that if a drift f is applied to the Wiener process~~w_t~~, and the reference frame is shifted by ϕ, the process $x_t - \phi(t)$ which has the drift $f(\cdot + \phi) - \dot{\phi}$ ~~, we can apply~~ is obtained. The weight relative to the Wiener measure can be calculated by Girsanov's formula ~~to get~~ ~~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 where $\mathbb{E}[\cdots \mid \|w\|_T < \epsilon]$ converges to 1 as $\epsilon \to 0$.~~

~~In particular, the exponentiated average of the cross term $f \cdot dw$ in~~ as follows.

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

where the expectation is taken with respect to the Wiener process w conditioned to $\|w\|_T \equiv \sup_{0<t<T} |w_t| < \epsilon$. We are going to evaluate the terms containing w_t in the exponent on the RHS of Eq. (B11)~~tends away from 1 as follows (Ikeda and Watanabe, 1981):~~ :

1. If we assume ϕ is a twice continuously differentiable function, then by applying Ito's product rule to $\dot{\phi}(t) w_t$, and using $(\forall t)\ |w_t| < \epsilon$,

   $$\left|\int_0^T \dot{\phi}(t) dw_t\right| = \left|\dot{\phi}(T) w_T - \int_0^T w_t \ddot{\phi}(t) dt\right| \leq A_1 \epsilon, \tag{B12}$$

   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~~ A_1 is a positive constant independent of ϵ.

2. If we assume f is a twice continuously differentiable function, then by using $(\forall t)\ |w_t| < \epsilon$,

   $$\left|\int_0^T f(w_t+\phi(t))\dot{\phi}(t) dt - \int_0^T f(\phi(t))\dot{\phi}(t) dt\right| \leq A_2 \epsilon, \tag{B13}$$

   where A_2 is a positive constant independent of ϵ.

3. In the similar manner as in 2,

   $$\left|\int_0^T |f(w_t+\phi(t))|^2 dt - \int_0^T |f(\phi(t))|^2 dt\right| \leq A_3 \epsilon, \tag{B14}$$

   where A_3 is a positive constant independent of ϵ.

4. The evaluation of $\int_0^T f(w_t + \dot{\phi}(t)) dw_t$ is as follows.

   (a) By applying Taylor's expansion to $f(w_t + \phi(t))$,

   $$\int_0^T f(w_t+\phi(t)) \cdot dw_t = \int_0^T f(\phi(t)) \cdot dw_t + \int_0^T (w_t \cdot \nabla) f(\phi(t)) \cdot dw_t + \int_0^T O(w^2) \cdot dw_t. \tag{B15}$$

(b) By applying Ito's product rule to $w_t f(\phi(t))$, and using $(\forall t)\ |w_t| < \epsilon$,

$$\int_0^T f(\phi(t)) \cdot dw_t = w_T f(\phi(T)) - \int_0^T \sum_{i,j} w_t^i \frac{\partial f_i}{\partial x_j}(\phi(t))\dot{\phi}_j(t)dt = O(\epsilon). \tag{B16}$$

(c) Regarding the second term on the RHS of Eq. (~~B25).~~ ~~For a rigorous derivation, see the proof of Theorem IV 9.1 of Ikeda and Watanabe (1981)or Section~~B15), we see that

$$\begin{aligned}
&\int_0^T (w_t \cdot \nabla) f(\phi(t)) \cdot dw_t + \frac{1}{2}\int_0^T \nabla \cdot f(\phi(t))dt \\
&= \int_0^T \sum_{i,j} \frac{\partial f_i}{\partial x_j}(\phi(t)) w_t^j dw_t^i + \frac{1}{2}\int_0^T \sum_{i,j} \delta_{ij} \frac{\partial f_i}{\partial x_j}(\phi(t))dt \\
&= \int_0^T \sum_{i,j} \frac{\partial f_i}{\partial x_j}(\phi(t)) \left(w_t^j dw_t^i + \frac{1}{2}\delta_{ij} dt \right) = \int_0^T \sum_{i,j} \frac{\partial f_i}{\partial x_j}(\phi(t)) d\zeta_t^{ji},
\end{aligned} \tag{B17}$$

where $\zeta_t^{ji} = \int_0^t w_s^j \circ dw_s^i$ (Stratonovich integral).

By applying Evaluations 1–4 to Eq. ~~2 of Zeitouni (1989).~~(B11), we obtain

$$\begin{aligned}
I_\epsilon(\phi) = \exp &\left(\int_0^T \left| f(\phi(t)) - \dot{\phi}(t) \right|^2 dt - \frac{1}{2}\int_0^T \nabla \cdot f(\phi(t))dt \right) \\
&\times \mathbb{E}\left[\exp\left(O(\epsilon) + O(\epsilon^2) + \int_0^T \sum_{i,j} \frac{\partial f_j}{\partial x_i}(\phi(t)) d\zeta_t^{ji} + \int_0^T O(|w|^2) \cdot dw_t \right) \middle| \|w\|_T < \epsilon \right],
\end{aligned} \tag{B18}$$

On pages 450–451 in Ikeda and Watanabe (1981), it is shown that

$$\mathbb{E}\left[\exp\left(c\int_0^T \sum_{i,j} \frac{\partial f_j}{\partial x_i}(\phi(t)) d\zeta_t^{ji} \right) \middle| \|w\|_T < \epsilon \right] \xrightarrow{\epsilon \to 0} 1 \quad (\forall c), \tag{B19}$$

$$\mathbb{E}\left[\exp\left(c\int_0^T O(|w|^2) \cdot dw_t \right) \middle| \|w\|_T < \epsilon \right] \xrightarrow{\epsilon \to 0} 1 \quad (\forall c), \tag{B20}$$

and it is obvious that

$$\mathbb{E}\left[\exp\left(cO(\epsilon) + cO(\epsilon^2) \right) \middle| \|w\|_T < \epsilon \right] \xrightarrow{\epsilon \to 0} 1 \quad (\forall c). \tag{B21}$$

They also showed that if

$$\mathbb{E}\left[\exp\left(ca_j \right) \middle| \|w\|_T < \epsilon \right] \xrightarrow{\epsilon \to 0} 1 \quad (\forall c) \tag{B22}$$

for $j = 1, 2, \cdots, J$, then

$$\mathbb{E}\left[\exp\left(\sum_{j=1}^{J} a_j\right)\middle|\|w\|_T < \epsilon\right] \xrightarrow{\epsilon\to 0} 1. \tag{B23}$$

By applying this to Eqs. (B20), (B19), and (B21), we deduce from Eq. (B18) that

$$I_\epsilon(\phi) \xrightarrow{\epsilon\to 0} \exp\left(\int_0^T \left|f(\phi(t)) - \dot{\phi}(t)\right|^2 dt - \frac{1}{2}\int_0^T \nabla\cdot f(\phi(t))dt\right). \tag{B24}$$

From evaluation 4, we also have that

$$\mathbb{E}\left[\exp\left(\int_0^T f(w_t + \phi(t))\cdot dw_t\right)\middle|\|w\|_T < \epsilon\right] \xrightarrow{\epsilon\to 0} \exp\left[-\frac{1}{2}\int_0^T \operatorname{div} f(\phi(t))dt\right]. \tag{B25}$$

~~Notice that~~ Eq. (B25) serves as an evaluation formula for the divergence term along ϕ via ensemble calculation if we interpret the expectation as an ensemble average:

$$\ln\mathbb{E}\left[\exp\left(\int_0^T f(w_t + \phi(t))\cdot dw_t\right)\middle|\|w\|_T < \epsilon\right] \xrightarrow{\epsilon\to 0} -\frac{1}{2}\int_0^T \operatorname{div} f(\phi(t))dt. \tag{B26}$$

The ensemble can be generated by using ~~the~~ a Wiener process limited to the small area $\|w\|_T < \epsilon$. Taking the derivative of Eq. (B26) with respect to $\phi_i(t)$, we also ~~get~~ obtain the formula for evaluating the derivative of the divergence term along ϕ, as follows.

$$\frac{\mathbb{E}\left[\nabla f(\phi + w)\cdot dw \exp\left(\int_0^T f(\phi + w)\cdot dw\right)\middle|\|w\|_T < \epsilon\right]}{\mathbb{E}\left[\exp\left(\int_0^T f(\phi + w)\cdot dw\right)\middle|\|w\|_T < \epsilon\right]} \xrightarrow{\epsilon\to 0} -\frac{1}{2}\nabla(\operatorname{div} f)dt, \tag{B27}$$

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 (B26) and (B27) 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. (42) and (41) ~~utilize~~ utilise 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~~ realisation of the cost function is given as

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

where b is a small number. Notice that $\hat{J}_{\text{tube}}(\phi|y)$ 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. (48), $\phi^{(k+1)} = \phi^{(k)} - \alpha\nabla\hat{J}_{\text{tube}}$, 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.

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.