Stochastic Integration for the Heterogeneous Correlation Modeling Using a Diffusion Equation

OLIVIER PANNEKOUCKE AND LAURENT VEZARD
CNRM/GAME, Météo-France/CNRS, Toulouse, France

(Manuscript received 23 September 2009, in final form 3 February 2010)

ABSTRACT

In this note, a stochastic integration scheme is proposed as an alternative to a deterministic integration scheme, usually employed for the diffusion operator in data assimilation. The stochastic integration scheme is no more than a simple interpolation of the initial condition in lieu of the deterministic integration. Furthermore, this also presents a potential in high performance computing. For the classic preconditioned minimizing problem, the stochastic integration is employed to implement the square root of the background error covariance matrix, while its adjoint is obtained from the adjoint code of the square root code. In a first part, the stochastic integration method and its weak convergence are detailed. Then the practical use of this approach in data assimilation is described. It is illustrated in a 1D test bed, where it is shown to run smoothly for background error covariance modeling, with nearest-neighbor interpolations, and $O(100)$ particles.

1. Introduction

The purpose of data assimilation is to provide the most likely numerical representation $x_a$ (the analysis) of the atmosphere, or of the ocean, from known observations $y_o$. Nevertheless, observations are known to be noisy; that is, if $x$ denotes the true numerical state then $y_o = H(x) + e_o$, with $e_o$ as the observation error and $H$ the observation operator that maps the model space into the observation space. Moreover, the observational network is heterogeneous in space and in time, with not enough observed area, which leads to a closure problem. A practical solution consists in adding a background $x_b$ (generally the most recent forecast) and to compute the analysis state as a correction $\delta x_a$ of the background, so that $x_a = x_b + \delta x_a$.

The variational approach for resolving the issue consists in minimizing the cost function:

$$J(v) = v^T v + (d - HB^{1/2}v)^T R^{-1} (d - HB^{1/2}v),$$

where $B = \mathbb{E}(e_b e_b^T)$ is the background error covariance matrix with $e_b = x_b - x$, $R = \mathbb{E}(e_o e_o^T)$ is the observation-error covariance matrix, $d = y_o - H(x)$ is the innovation vector, and $v = B^{-1/2} \delta x$. The minimum of $J$ is reached in $v_a = B^{1/2} \delta x_a$. The square root matrix $B^{1/2}$ is defined so that $B = B^{1/2} B^{T/2}$. The variational problem is solved by iterations along a minimizing process. Classic minimizing algorithms (e.g., conjugated gradient) require the gradient of the cost function:

$$VJ(v) = v - B^{1/2} H^T R^{-1} (d - HB^{1/2}v),$$

where it clearly appears that $B^{1/2}$ and its adjoint $B^{T/2}$ are needed. In practice, $B$ is flow dependent, with heterogeneous correlation functions. The local length scale can stand for the local heterogeneity and can be defined as the scale for which the tangential parabola of the correlation function at the origin is equal to 0.5 (Daley 1991; Pannekoucke et al. 2008). Ensemble methods provide a way to estimate this matrix (Fisher 2003), but the usually small size of the ensemble leads to noisy statistics due to sampling effects. The huge size of $B$ implies that it must be modeled.

Different models exist for correlation matrix and among them, the diagonal assumption in spectral space (Courtier et al. 1998), which leads to homogeneous correlation...
functions. Amid the heterogeneous correlation models, one can cite the diagonal assumption in wavelet space (Fisher 2003), the recursive filter approach (Purser et al. 2003a,b), and the diffusion equation (Weaver and Courtier 2001).

The formulation based on the diffusion equation is now considered. It appears that along the minimization process of the cost function \( f \), a time integration of the diffusion equation is required for each computation of \( B^{1/2} \) (or \( B^{-1/2} \)) applied on an input vector. Since one deterministic integration is numerically costly, it would be desirable to find alternative ways. In this note, a stochastic time integration is considered as a possible alternative in lieu of the deterministic integration. The advantage of this approach is to replace a deterministic time integration by simple independent interpolations. This requires the offline computation of interpolation positions. The method is illustrated within a one-dimensional test bed.

The note is organized as follow. Section 2 recalls the correlation model based on the diffusion equation and its link with probabilities. The relationship between stochastic differential equations (SDEs) and elliptic partial differential equations (EPDEs) is described in section 3. This section also gives details about the numerical stochastic resolution of SDE for different boundary conditions. A one-dimensional illustration is then proposed in section 4, with emphasis on the technical implementation. The present work focuses on the horizontal covariance modeling, and details about 3D models can be found in Weaver and Courtier (2001) or in Pannekoucke (2009) for a nonseparable formulation. This is justified by the fact that in 3D models, it is classic to represent the horizontal component separately from the vertical component.

2. Covariance modeling with the diffusion operator

On the plane \( \mathbb{R}^2 \) with coordinates \( x = (x_i)_{i \in \{1, 2\}} \), the heterogeneous diffusion equation of a field \( u \) is given by

\[
\partial_t u = \partial_i (\nu_i \partial_i u),
\]

where \( \nu(x) = [\nu_i(x)] \) is the local diffusion tensor. In this equation, the Einstein summation convention is used and \( \partial_i \) denotes the partial derivative along the \( x_i \) coordinate. The Gaussian function is a particular solution of the homogeneous diffusion equation (i.e., when the diffusion tensors are constant).

In fact, it corresponds to a Green solution (also called, in this situation, the heat kernel):

\[
G^{\nu}(x, x') = \frac{1}{|2\pi \nu|^{1/2}} \exp \left\{ -\frac{1}{2} (x - x')^T \nu^{-1} (x - x') \right\},
\]

with \( \Gamma = 2\nu t \) and \(| \cdot | \) denotes the matrix determinant, so that

\[
u(x, t) = \int_{\mathbb{R}^2} G^{\nu}(x, x') u(x', t = 0) \, dx'_1 \, dx'_2.
\]

This has led some authors to take advantages of this Gaussian solution to design models of background error correlation functions. Thus, following Weaver and Courtier (2001), the square root of the background error covariance matrix is formulated as a product (for monovariate formulation):

\[
B^{1/2} = \Sigma L^{1/2} W^{-1/2},
\]

where \( \Sigma \) is the diagonal matrix of standard deviations and \( L \) is the time integration over \( t = 0 \) to \( t = 1 \) (this is the convention adopted in the present note) of the heterogeneous diffusion Eq. (1), so that \( \Gamma = 2\nu \) (Pannekoucke and Massart 2008). Thus, \( L^{1/2} \) denotes the half time integration. Here \( W \) and \( \Lambda \) are two normalizations matrix, the former is a metric tensor that depends on the spatial discretization (Weaver and Courtier 2001; Pannekoucke and Massart 2008), and the latter is so that \( C^{1/2} = \Lambda L^{1/2} W^{-1/2} \) leads to a correlation matrix \( C = C^{1/2} C^{1/2} \).

A crucial aspect of this model is related to the choice of the local diffusion tensor \( \nu(x) \). A solution to the estimation of these tensors has been proposed by Pannekoucke and Massart (2008), and it relies on the ensemble methods to compute \( \nu(x) \) from the local length scale (Pannekoucke et al. 2008), under a local homogeneous assumption of the tensor field: the matrix \( \Gamma(x) \) is first estimated from an ensemble of forecasts, then the local diffusion tensor is set as \( \nu(x) = \Gamma(x)/2 \). This is the diffusion tensor field used in the heterogeneous diffusion in Eq. (1).

Moreover, an efficient numerical implementation of the diffusion in Eq. (1) is needed to construct \( L^{1/2} \). The numerical solution can be obtained with a time and space discretization of Eq. (1) (Weaver and Courtier 2001) or also by using a spectral approach instead of a space discretization (Pannekoucke and Massart 2008).

The next section introduces another way to integrate Eq. (1) by mean of stochastic processes. Before describing this tool, it is relevant to observe the particular form of Eq. (2). In a homogeneous case, the Gaussian form of the heat kernel is a probability density function and thus, Eq. (2) can be written as the following expectation:

\[
u(x, t) = \mathbb{E}[u(X', 0)],
\]

where \( X' \) is a random variable with the density of probability \( p_{\nu}(X') = G^{\nu}(x, X') \), and where \( \mathbb{E} \) stands for the expectation operator. At this stage, the idea is to approximate the last distribution by an empirical one [i.e.,
process in Eq. (5) can also be written as the SDE: a new trajectory for the particle. The Ito\hat{\text{}} integral leads to a new trajectory for the particle. Because of the randomness, each realization of 1D EPDEs can be obtained as the expectation under the Dirac distribution. In that context, the expectation samples of the random variable $x^i$ can be found by interpolating in the discrete case. Each sample is obtained as a path of a stochastic process that remains to be found. Finally, the benefit is that the time integration is simply reduced to local interpolations of the initial condition at appropriate positions. This property is the key idea that motivates this note.

3. A stochastic integration scheme for EPDEs

The aim of this section is to recall how the solution of 1D EPDEs can be obtained as the expectation under a stochastic process (to be determined). A rigorous, but easily grasped, mathematical description can be found in Oksendal (2003), with an extension to $n$-dimensional cases. Thus, in a first time, the required background on stochastic calculus is depicted and then illustrated, next section, within a simple one-dimensional test bed.

a. Link between stochastic calculus and EPDEs

An Itô diffusion process is a stochastic process $X_t$ defined by

$$X_t = X_0 + \int_0^t a(X_s) \, ds + \int_0^t b(X_s) \, dB_s,$$  

where $a(x)$ and $b(x)$ are two bounded functions. The first integral corresponds to a classic integration while the second is an Itô integral related to a Brownian process $B_t$, (i.e., a stochastic process so that, for a duration $\tau$, $B_{\tau+\tau} = B_\tau$ is a Gaussian variable with a mean equal to zero and a variance to $\tau$). The Itô integral is defined as the limit when $\delta t \to 0$ of the sum $\sum b(X_{t_i}) (B_{t_{i+1}} - B_{t_i})$, as in the Riemann integral. In this expansion, the first integral of Eq. (5) is the deterministic part of the process, while the second integral is the Martingale term (Oksendal 2003), so that $\mathbb{E}\left[\int_0^t b(X_s) \, dB_s\right] = 0$. Physically, $X_t$ is the coordinate, at time $t$, of the random movements of a particle (e.g., pollen particles floating in water under the microscope). Because of the randomness, each realization leads to a new trajectory for the particle. The Itô process in Eq. (5) can also be written as the SDE:

$$dX_t = a(X_t) \, dt + b(X_t) \, dB_t,$$  

Later, a process $X_t$, starting from $x$ at $t = 0$ is denoted by $X_t^x$. The extension of the deterministic chain rule for stochastic processes is a fundamental tool in stochastic calculus. Let $f$ be a function, then the stochastic process associated with the change of variable $f(X_t)$, is given by the Itô formula:

$$d[f(X_t)] = (Af)(X_t) \, dt + (b \partial_x f)(X_t) \, dB_t,$$  

where $A$ is the generator of the stochastic process $X_t$, defined by

$$A f(x) = (a \partial_x f)(x) + \frac{1}{2} (b^2 \partial^2_x f)(x).$$  

Note that a stochastic chain rule gives rise to extra terms compared with a deterministic chain rule. At this stage, the first link between the stochastic process and the diffusion equation is to write this second-order differential operator within a prognostic relation.

The form of Eq. (7) indicates that $Y_t$ is also an Itô diffusion, whose integral form is

$$f(X_t) = f(X_0) + \int_0^t (Af)(X_s) \, ds + \int_0^t (b \partial_x f)(X_s) \, dB_s.$$  

Therefore, the expectation $\mathbb{E}$ of the previous equation is formally written [Dynkin’s formula; see Oksendal (2003)] as

$$\mathbb{E}[f(X_t^x)] = f(x) + \int_0^t \mathbb{E}(Af)(X_s^x) \, ds.$$  

Then, it appears that

$$u(x, t) = \mathbb{E}[f(X_t^x)],$$  

differentiable with respect to $t$ and that its derivative is

$$\partial_t u = \mathbb{E}[Af](X_t^x).$$  

The last step is to permute the generator $A$ and the expectation. This is done thanks to the Kolmogorov’s backward equation theorem (Oksendal 2003) stating that $u$ is the solution of the EPDE:

$$\partial_t u = Au,$$  

with the initial condition $f(x) = u(x, 0)$.

Conversely, as previously mentioned, the solution $u(x, t)$ of the second-order partial differential equation in Eq. (10), with the initial condition $u(x, 0)$, can be obtained
as the expectation, at time \( t \), of the random variable \( u(X_t, 0) \), where \( X_t \) is the stochastic process of generator \( \mathcal{A} \). In practice, the solution \( u(x, t) \) is reached as follows. From a given generator \( \mathcal{A} \), one has to construct an Itô diffusion \( X_t \), deduced from Eqs. (8) and (6), and integrate Eq. (10) as the expectation \( u(x, t) = \mathbb{E}[u(X_t^x, 0)] \), where \( X_t^x \) start from \( x \) at time \( t = 0 \).

**b. Numerical solution and convergence properties**

The solution \( u(x, t) = \mathbb{E}[u(X_t^x, 0)] \) of Eq. (10), is numerically approximated as the empirical mean:

\[
u(x, t) \approx \frac{1}{N} \sum_k u(\hat{X}_{k,t}^x, 0), \tag{11}\]

where \( (\hat{X}_{k,t}^x) \), for \( k \in [1, N] \), is an ensemble of \( N \) particles randomly generated and starting in \( x \). Each path \( \hat{X}_{k,t}^x \) is numerically generated from the discrete version of Eq. (6), for a given time scheme. In what follows, an Euler–Maruyama time scheme (Kloeden and Platen 1992) is used. The integration over the period \([0, t]\) is achieved along the discrete time \( t_i = i \delta t \), with \( i \in \{0, n\} \) where \( \delta t = \frac{t}{n} \) (\( n \) is the number of time discretization). Thus, the Euler–Maruyama scheme is written as

\[
\begin{align*}
\hat{X}(0) &= x, \\
\hat{X}(t_{i+1}) &= \hat{X}(t_i) + a[\hat{X}(t_i)] \delta t + b[\hat{X}(t_i)] \sqrt{\delta t} \xi(t_i),
\end{align*}
\tag{12}\]

where \( \xi(t_i) \) is a sample of a random Gaussian variable \( \zeta \) with zero mean and with a standard deviation of one.

The convergence of Eq. (11) depends on the time step \( \delta t \) and on the numbers of particles \( N \). The solution \( u(x, t) = \mathbb{E}[u(X_t^x, 0)] \) only requires the expectation of the random variable \( u(X_t^x, 0) \), thus the accuracy of the numerical solution is considered only through a weak convergence (Kloeden and Platen 1992). This weak convergence, of the approximation \( \hat{X}_t \), toward \( X_t \), is given by

\[
\left| \mathbb{E}[u(X_t^x, 0)] - \mathbb{E}[u(\hat{X}_{k,t}^x, 0)] \right| \leq K_1 (\delta t)^\alpha,
\]

with \( \alpha = 1 \) for the Euler–Maruyama scheme, and \( K_1 \) a positive constant that depends on \( t \). Furthermore, the weak convergence of the empirical mean to the expectation is in the classic form \( O(N^{-1/2}) \) according to the central limit theorem, implying almost surely the inequality:

\[
\left| \mathbb{E}[u(X_t^x, 0)] - \frac{1}{N} \sum_k u(\hat{X}_{k,t}^x, 0) \right| \leq K_2 \frac{1}{\sqrt{N}}. \tag{13}\]

Therefore, with a triangular inequality, the convergence of Eq. (11) almost surely is

\[
\left| \mathbb{E}[u(X_t^x, 0)] - \frac{1}{N} \sum_k u(\hat{X}_{k,t}^x, 0) \right| \leq K_1 (\delta t)^\alpha + K_2 \frac{1}{\sqrt{N}}. \tag{14}\]

In practice, the functions \( u(x, 0), a(x) \) and \( b(x) \) are known by their values on a grid. Out of the grid, an interpolation method is required. In this paper, only two kinds of interpolation are considered: the nearest neighbor and the linear interpolation. This is equivalent to assume the functions piecewise constant in the former interpolation method, and continuously piecewise linear in the latter. Of course, such interpolations introduce discrepancies with the continuous case, and the right-hand-side bound in Eq. (14) must be completed with a third term, which depends on the spatial discretization. In the nearest-neighbor case, this error is in \( O(\delta x) \), while it is in \( O(\delta x^2) \) in the linear case. As a consequence, the numerical heat kernel in the homogeneous case is no more Gaussian.

As suggested in the previous section, the solution \( u(x, t) \) simply results from an empirical average of local particles: in Eq. (1) the value \( u(x, t) \) can be estimated from an ensemble of particles whose paths are limited within the neighborhood around \( x \). The width of this area depends on the local diffusion magnitude. From common results on normal law and considering a 1D homogeneous diffusion of diffusion coefficient \( \nu = L^2/2 \), there is a probability bigger than 99.99\% that, at time \( t = 1 \), a particle stays inside the segment of radius \( 3.9L \) around the starting point. Similar thresholds can be found for \( n \)-dimensional cases thanks to the chi-square law (e.g., in 2D, the equivalent threshold is \( 4.3L \)). These bounds can serve for high performance computing by distributing the computation of Eq. (11) over the global domain into several independent computations over local areas with an appropriate halo. For this purpose, the width of the halo can be fixed thanks to a length scale \( L \) chosen as the largest length scale over the local area (including anisotropic effects). The numerical cost can be reduced by replacing the normal law \( \zeta \) with a random variable of the same mean and variance, such as the Bernoulli variable of values \(-1\) and \(1\), with equal probability. The weak convergence is unchanged. Also, samples of the Bernoulli variable (a switch on a uniform law) are cheaper to obtain than normal law [often generated with a Box–Muller algorithm; Kloeden and Platen (1992)].

**c. Boundary conditions**

The domain considered here is an open bounded set \( D \subset \mathbb{R}^2 \) with an external boundary \( \partial D \). Periodic, homogeneous Neumann and homogeneous Dirichlet boundary conditions are detailed, since they are used in data assimilation (Weaver and Courtier 2001).
The homogeneous Neumann condition corresponds to the case where the derivative is null either on the boundary (or as part of the boundary). The Neumann conditions are also called “wall conditions.” The last denomination is well illustrated in stochastic calculus: homogeneous Neumann conditions match the case where the particles are reflected on the boundary. The reflection homogeneous Neumann conditions correspond to the case where the solution of Eq. (1), with the Dirichlet condition on the boundary, the solution must verify

\[ u(x, t) = \Phi_{D,D}(x), \]

with \( t \wedge \tau = \min(t, \tau) \), and where particles \( x \) on the boundary are evaluated as \( u(x, t) = \Phi_{D,D}(x) \). The homogeneous condition so that \( \Phi_{D,D} \) is a null function.

Thus, after each time step in Eq. (12), a test is required to verify if the particle is still inside the domain or if it has left it. Then, an appropriate reflection or stopping is applied, depending on the boundary condition. Of course, if the domain is periodic, the particle path always stays inside the domain and there no testing is necessary.

4. Applications in data assimilation

The previous formalism is now applied, for data assimilation purposes, in a simple one-dimensional test bed. The aim of this section is to document the use of the stochastic method for the time integration of the correlation model based on the diffusion operator.

The domain is a segment of length \( 2\pi a \), where \( a = 6480 \) km, and it is discretized with 241 points. Three kinds of boundary conditions are considered: periodic, homogeneous Dirichlet, and homogeneous Neumann. Since the domain is similar to the great earth’s circle in the periodic case, then the horizontal coordinate \( x \) is considered either in degree or in kilometer. The step size is \( \delta x = 166 \) km. In this simplified framework, the diffusion equation of a field \( u \) is

\[ \delta_x \mu = \delta_x [\nu(x) \delta_x \mu], \]

where \( \nu(x) \) is a field of local diffusion coefficients. Under a local homogeneous approximation, the diffusion field is linked to the length scale field \( L(x) \) by \( \nu(x) = L^2(x)/2 \) (Pannekoucke and Massart 2008). Two cases are considered: an homogeneous case, where \( L(x) = L_h \), and an heterogeneous case, where \( L(x) = L_h[1 + \cos(x/a)/2] \), with \( L_h = 250 \) km (\( L_h/\delta x = 1.5 \)). Thereafter, the heterogeneous (heterogeneous) covariance matrix modeled with the homogeneous (heterogeneous) diffusion is denoted \( B_{hn}(B_{hn}) \). Moreover, the matrices computed by the mean of the stochastic process and using \( N \) paths by points are denoted with a subscript index \( N \).

First of all, the stochastic process with the generator of Eq. (15) is determined with respect to section 3. Then, the stochastic time integration is tested under various forms and boundary conditions to construct \( B_{N}^{1/2} \) and its adjoint \( B_{N}^{1/2} \). Of course, these matrices are not explicitly stored in computer memory but implemented so that computations are equivalent in their overall effect to simulation of these operators being applied multiplicatively to input vectors.

a. Stochastic process associated with the 1D diffusion

The mathematical tool described in the previous section is applied here to integrate the diffusion in Eq. (15). The current equation is not under the practical form in Eq. (8). This can be done by rewriting Eq. (15) as

\[ \delta_x \mu = \delta_x \nu(x) \delta_x \mu + \nu(x) \delta_x^2 \mu, \]

where the generator \( A f = \delta_x \nu(x) \delta_x f + \nu(x) \delta_x^2 f \) is clearly visible. Then the unknown functions \( a \) and \( b \) are easily found by identification with Eq. (8), the result being

\[
\begin{align*}
a(x) &= \delta_x \nu(x), \\
b(x) &= \sqrt{2\nu(x)}. 
\end{align*}
\]

The acting Itô diffusion is finally written as

\[ dX_t = \delta_x \nu(X_t) \, dt + \sqrt{2\nu(X_t)} \, dB_t. \]

See the appendix for the \( n \)-dimensional case.

b. Implementation of \( B^{1/2} \) and \( B^{1/2} \)

In practice, \( B^{1/2} \) is not stored in the computer memory but is evaluated on a vector as \( B^{1/2} \mathbf{v} \). The evaluation by \( B^{1/2} \) is achieved through the product in Eq. (3). Products have to be performed by matrix-vector, with diagonal matrices \( W^{-1/2} \), \( \Lambda \) and \( \Sigma \), are reduced to element-wise products.

In 1D framework, the diagonal values of \( W^{-1} \) are \( \delta x \) for points inside the domain. For points on the boundary, the diagonal value is \( \delta x/2 \) for periodic conditions and \( \delta x/2 \) for other cases.
Under a local homogeneity assumption, a good approximation for the normalization $L$ can be deduced from the local diffusion tensor (Pannekoucke and Massart 2008). In a 2D, the normalization applied at position $x$ is $L(x)^2 = 2\pi \sqrt{2} \nu(x)$. In a 1D, it is $L(x)^2 = \sqrt{2\pi} L(x)$, where $L(x)$ is the local length scale in $x$. In the following, either the exact or the previous approximated normalization are used.

The principal difficulty rests with the computation of the operator $L^{1/2}$. In this paper, the operator is a half-time stochastic integration (from $t = 0$ to $t = \frac{1}{2}$) of the heterogeneous diffusion equation. It requires the computation at time $t = \frac{1}{2}$ of $N$ particles per grid points, governed by an Itô diffusion. Each path is integrated according to Eq. (12), with appropriate boundary conditions as described in section 3c. These endpoints, at half-time, are computed once and for all, and are stored in computer memories. Then endpoints are used at each computation of $L^{1/2} v$ in the empirical mean evaluation in Eq. (11), where the initial condition $u(x, 0)$ is replaced by interpolations of $v$ at the endpoint paths.

The adjoint $B_N^{T/2}$ can be carried out from the direct code of $B_N^{1/2}$, by using classic adjoint coding rules (Giering and Kaminski 1998). The endpoints involved in the computation of $B_N^{1/2}$ are reused in the computation of $B_N^{T/2}$. In this way, $B_N^{T/2}$ leads to the exact transpose of the matrix $B_N^{1/2}$.

Note that the computation of the background-error covariance matrix as a product of $B_N = B_N^{1/2} B_N^{T/2}$ ensures the symmetry and the positiveness of the covariance model.

c. Numerical experiments and results

Several points for the use of the stochastic approach need to be documented: the sensitivity to the interpolation method; the convergence versus the number of particles, the ratio $L_h/dx$, and the time step; the validity of the approximated normalization; and the use of Dirichlet–Neumann boundary conditions. These are now detailed.

1) SENSITIVITY WITH INTERPOLATION METHODS

As explained in section 3b, an interpolation is needed in the numerical resolution of the dynamics in Eq. (12). The discrepancies between the background error covariance matrix resulting from the linear interpolation and the nearest-neighbor interpolation are illustrated in Fig. 1. In these simulation, the diffusion is heterogeneous with periodic boundary conditions and $\delta t = 1/200$. The number of particles ($N = 10^5$) is large enough to damp the sampling noise. The correlation functions obtained...
by using the exact normalization are presented in Fig. 1a (only 1 out of 10 is represented). Note that these correlation functions can be considered as analysis increments associated to one observation assimilation experiments. In that case, a single observation is located at the position of the maximum of each correlation function. The correlation function related to the positions $90^\circ$ and $180^\circ$ are in bold. It appears that correlation functions are quasi-Gaussian. The difference between correlations obtained with nearest-neighbor interpolation minus those obtained with linear interpolation is represented in Fig. 1b (again only 1 out of 10 is represented). The difference related to the positions $90^\circ$ and $180^\circ$ are also in bold. The linear interpolation leads to greater correlation values than the nearest-neighbor interpolation. This is confirmed by the length scale diagnosis, reported in Fig. 1c, where the length scale of $B_{N}$ in the linear interpolation case (dash–dotted line) is larger than in the nearest-neighbor case (dashed line). Note that both interpolations over estimate the true length scale field (solid line).

Results from this sensitivity test show that the correlation functions obtained with a nearest-neighbor interpolation are similar to those of the functions obtained with the linear interpolation. Moreover, the nearest-neighbor interpolation is less costly than the linear interpolation. Therefore, the nearest-neighbor interpolation appears as a good compromise for data assimilation constraints.

2) CONVERGENCE VERSUS THE NUMBER OF PARTICLES, THE LENGTH SCALE MAGNITUDE, AND THE TIME STEP

The sampling noise effect, arising from the number of particles $N$, is investigated for the homogeneous diffusion with periodic boundary conditions. For such a case, the theoretical correlation function is known and equal to $\rho(x) = \exp(-x^2/L_h^2)$. The theoretical correlation matrix $B_{ho}$ is compared with simulated matrices $B_{N}$ for $N$ in $(100, 400, 1600, 6400)$. The $B_{N}$ is computed with the nearest-neighbor interpolation and with the integration time step sets to $\delta t = 1/200$. Moreover, an exact normalization is employed for $L_h$. The discrepancy between $B_{ho}$ and $B_{N}$ is quantified by the relative error $e_{N} = 100\||B_{N} - B_{ho}||/\||B_{ho}||\)$, where $\||\|$ is the Hilbert norm and $\|B\| = \sqrt{\text{Tr}(BB^T)}$, where Tr denotes the matrix trace. This relative error is represented by a solid line in Fig. 2. It appears that $B_{N}$ do not converge toward $B_{ho}$. This is related to the error due to the interpolations and to the time scheme, leading to a solution $B_{ho}$ distinct from the theoretical one. However, the error decreases when the ratio $L_h/\delta x$ increases: for $N = 6400$, the error is close to 4% for $L_h = 250$ km (see Fig. 2a), while it is less than 2% for $L_h = 500$ km ($L_h/\delta x = 3$) and $L_h = 1000$ km ($L_h/\delta x = 6$) (see Figs. 2b,c). The errors due to the interpolation and to the time
discretization (due to $\delta t = 1/200$) are damped with the ratio $L_h/\delta x$.

The convergence of $B_N$ toward $B_{\infty}$ is theoretically in $O(N^{-1/2})$ since the convergence of $B_N^{1/2}$ to $B_{\infty}^{1/2}$ is in $O(N^{-1/2})$ [according to Eq.(13)]. Confirmation can be seen in Fig. 2 where the dash–dotted line represents the relative error $e_N = 100(||B_N - B_{\infty}||/||B_{\infty}||)$, where $B_{N=\infty}$ is approximated by $B_N$ for a large enough $N_{x\infty}$ (here $N_{x\infty} = 10^3$). The slope of $e_N$ is asymptotically $-1/2$ (represented by the dashed line). But the error increases with $L_h/\delta x$ for a small ensemble of particles: for $N = 100$, the error is less than 10% for $L_h = 250$ km (see Fig. 2a), while it is close to 15% for $L_h = 1000$ km (see Fig. 2c). Same results are found for heterogeneous case (not shown here).

Some sensitivity experiments (not shown here) on the time step $\delta t = 1/2n$ ($n$ is the number of discretization of $[0, 1/2]$, with $N = 400$, indicate that correlations are overestimated for small $n$ (of the order of 10) but converge rapidly for large $n$ (of the order of 50). Note that in any case, the stochastic integration is stable, even with the extreme case $\delta t = 1/2$, and the resulting correlation functions are quasi-Gaussian.

With previous results, a practical ensemble size of $N = 100$ (around 10% error) or $N = 400$ (around 5% error) is accurate enough [that is in $O(100)$]. But this should be confirmed for real case applications since the error depends on the ratio $L_h/\delta x$. Small values for $\delta t = 1/2n$ increase the cost of the integration of particle paths, but this is not limiting as these trajectories are computed one time, only to retain the endpoints; thus, large $n$ can be used. In most of experiments performed in the note $n$ is set to 100.

3) DISCUSSION ON THE NORMALIZATION

As mentioned in section 4b, the normalization $A$ can be approximated from the local diffusion tensor. Under this approximation, the local normalization at the point $x$ is $A(x) = \sqrt{2\pi L(x)}$. This normalization is tested with a periodic boundary condition, for the heterogeneous diffusion operator, with nearest-neighbor interpolation, $N = 400$ and $\delta t = 1/200$. The resulting variance is reproduced in Fig. 3. The approximated normalization overestimates the exact normalization that would lead to a homogeneous variance field equal to 1 everywhere. Indeed, the mean of the resulting variance is 3% lower than 1. But the standard deviation of the variance is 0.05, which is small enough to propose a global inflation that forces the global variance to be equal to 1 in mean.

The approximated normalization is thus still efficient in this stochastic framework. However, some boundary effects can be encountered and are discussed hereafter.

4) ILLUSTRATION WITH NONPERIODIC BOUNDARY CONDITIONS

In data assimilation, bounded domains are common (e.g., in local area model or in ocean modeling with coast treatment). Homogeneous Dirichlet and Neumann boundary conditions are classically considered in covariance modeling with the diffusion operator (Weaver and Courtier 2001). They are now tested for a homogeneous diffusion operator for $N = 400$ and $\delta t = 1/200$. The boundary conditions are set as follow: homogeneous Dirichlet at $x = 0^\circ$ and homogeneous Neumann at $x = 360^\circ$. The correlation functions obtained with the approximated normalization $A$ are reported in Figs. 4a,b. As mentioned in section 4c(1), each function can be viewed as an increment analysis obtained from a single observation assimilation. It appears that these functions are not correlation functions since the variance is not a constant $\sqrt{2\pi L(x)}$ in a homogeneous case, and do not lead to correlation functions. These results can be compared

[FIG. 3. Test of the normalization approximated from the local diffusion tensor in the heterogeneous case. See text for details.]
when applying the exact normalization (see Figs. 4c,d). The correlations resulting from the Dirichlet condition are similar to those obtained with the Neumann condition, but the major difficulty is to find a good approximation for the normalization (Weaver and Courtier 2001). Recently, Mirouze and Weaver (2010) have improved the approximated normalization for these two boundary conditions (not tested here). As there are few differences between the correlation functions obtained with Dirichlet or with Neumann, the homogeneous Dirichlet condition is to be favored within this stochastic approach. Indeed, the Dirichlet condition is cheaper than the Neumann condition: there is no reflection to implement and the contribution of particles stopped at the boundary is null in Eq. (11).

5. Conclusions

In this note, a stochastic integration of the diffusion equation has been proposed and tested to be used in the model of covariance matrix related on the diffusion operator. The idea is to replace the deterministic integration by local interpolations of the initial condition. This method is naturally adapted for parallel computation since the computation only depends on combination of the initial condition taken on independent areas.

The stochastic integration is achieved thanks to a particle method whose path dynamic is described by a stochastic differential equation deduced from the heterogeneous diffusion equation.

It has been shown, within a one-dimensional test bed, that this approach can be considered in a data assimilation framework for background error covariance modeling. It rests on the expansion of the covariance matrix as the product of a matrix (the square root matrix) and its transposition. The transpose of the square root matrix is achieved through the adjoint code of the square root matrix direct code. It can be inferred from experiments that the nearest-neighbor interpolation (at a low cost compared with other interpolation methods), and a relatively small ensemble of particles $O(100)$, are enough to build quasi-Gaussian correlation functions.

However, there is still some work to achieve for the two-dimensional case where the number of particle might be much larger then the simple one-dimensional case. This study is considered as a preliminary change of

---

Fig. 4. Illustration of the correlation functions in (a),(c) the homogeneous Dirichlet case and (b),(d) the homogeneous Neumann case. With (a),(b) an approximated normalization and (c),(d) an exact normalization. See text for details.
perspective in the refinement of the diffusion formulation, and it requires further developments and optimizations to be ready to use.

Acknowledgments. Isabelle Mirouze, Sebastien Massart, and Jean-Antoine Maziejewski are warmly thanks for their careful reading of the manuscript. The authors are grateful to Christophe Baehr for fruitful discussions on stochastic calculus and numerous advice on this note.

APPENDIX

Stochastic Process for $n$-Dimensional Heterogeneous Diffusion

In the general $n$-dimensional case, with $x = (x_i)_{i \in [1,n]}$, Eq. (1) can be written as

$$\partial_t u = \partial_i \nu_{ij} \partial_j u + \nu_{ij} \partial_i^2 u.$$  \hfill (A1)

The aim is to express the $n$-dimensional Itô diffusion $dX_t = (dX_{t,i})_{i \in [1,n]}$, where

$$dX_{t,i} = a_i(X_t) \, dt + \sigma_{ij}(X_t) \, dB_{t,i},$$

with $B_t$ an $n$-dimensional Brownian process (i.e., a stochastic process so that, for a duration $\tau$, $B_{t+\tau} - B_t$ is a Gaussian vector with mean equal to zero and covariance matrix $\tau I$).

Following section 3 and its $n$-dimensional extension, the generator is thus $Af = \partial_i \nu_{ij} \partial_j f + \nu_{ij} \partial_i^2 f$. The unknown functions $a_i$ and $\sigma_{ij}$ are fixed by identification according to the $n$-dimensional version of Eq. (8), which results in the following:

$$\begin{align*}
a_i(x) &= \partial_i \nu_{ij}(x), \\
\sigma(x) &= \sqrt{2} \nu^{1/2}(x).
\end{align*}$$  \hfill (A2)

REFERENCES


Fisher, M., 2005: Background error covariance modelling. *Proc. ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, Reading, United Kingdom, ECMWF, 45–63.


