Approximation of a general data assimilation scheme by stochastic diffusion processes

KONSTANTIN BELYAEV
Research Centre
Bureau of Meteorology
700 Collins Street, Melbourne, 3000
AUSTRALIA

DETLEV MÜLLER
45 Neumühlen
22763 Hamburg
GERMANY

CLEMENTE A.S. TANAJURA
National Laboratory for Scientific Computing
Ministery of Science and Techonology
Av. Getúlio Vargas 333, Petrópolis, RJ
BRAZIL

Abstract: - Data assimilation schemes as commonly used in numerical circulation models are approximated in terms of stochastic processes. This approximation consists of random sequences of Markov chains, which converge to a diffusion-type process. The conditions for this convergence are investigated. The optimisation problem and a numerical experiment are discussed.

Key words: Sequences of Markov Chains, Diffusion Stochastic Processes; Data Assimilation Methods.

1 Introduction

In numerical modelling of geophysical systems, such as the ocean, the atmosphere or the climate system, data assimilation is a common and important approach. It produces the initial conditions for weather and climate models and complements monitoring by correcting model variables in the direction of observations. Such a correction of the model output by observational data is generally based on a scheme of the following type:
A system of partial differential equations, usually represented in finite-difference or finite-element form is considered on the time-interval (t_0, T). In general, t_0 may be associated with 0, while *T*, provided it is large enough, may be considered as infinity. The interval (t_0, T) can be broken into intervals (t_0, t_1), (t_1, t_2), …, (t_n, t_{n+1})… On any time-interval (t_n, t_{n+1}) model starts at moment t_n with initial state-vector θ_n and integrates forward until moment t_{n+1}, when it produces (predicts) the state-vector θ_{n+1}^m. Here and in the following the superscript *m* indicates that the system-state has been obtained by model integration is considered in contrast to other sources of information. During the time-interval (t_n, t_{n+1}) a series of observations enters as the vectors $\xi_1^n, \ldots, \xi_k^n$ independently of the model. Here, the indices *(k,n)* denote the *k*- vector on *n*-th time- interval. Usually, vector ξ_i^n is a subset of vector θ_n since only part of the model-state may be observed. Now the model-output θ_{n+1}^m is corrected by observations according to

$$\theta_{n+1} = \theta_{n+1}^m + \sum_{i=1}^{k} \alpha_i (\xi_i^n - \xi_i^m)$$

Here, ξ_i^m denote the model variables, corresponding to observations, calculated at observation time. The weight functions α_i will generally also depend on time and are either known a priori, or are determined by

some appropriate algorithm. The corrected state-vector θ_{n+1} is now taken as the new initial condition and the model integration resumes.

Under the name data assimilation methods various versions of the aforementioned scheme are commonly used in geophysics. In particular, the Kalman filter approach determines the optimal weight-coefficient α_i utilising the statistical properties of observational data [1]. Alternatively, the observations may not be considered as random. In this case, the optimal weights may be determined according to variational or adjoint data assimilation technique [1]. The integration-intervals can be considered as given (e.g., the model produces the forecast every 24 h), or random, i.e., the correction is applied if the model temperature difference between ocean and atmosphere exceeds a certain, a priori chosen limit.

Despite these and other differences, all these methods follow essentially the scheme mentioned above. Alternatively, it turns out interesting to review these techniques from a different point of view. At the "moment of assimilation" the time- series of variables θ_n undergoes a jump of its trajectories. What will happen if the interval between consecutive assimilations approaches zero along with the values of the jump? How does the limiting behaviour of the trajectories depend on the number of state variables and their distributions? Under which conditions does the limiting distribution of θ_n exist as *T* goes to infinity and what is it? If it exists, it is called the stationary distribution. These and similar questions not only attract interest from the theoretical point of view, but also for practical reasons. For instance, knowing the limiting behaviour of the time series for $\Delta t_k = t_{k+1} - t_k \to 0$, it becomes easy to calculate various parameters needed for the weather forecast, while known stationary distribution enhances the reliability of climate prediction. In addition, the knowledge of this limit may simplify the optimisation problem for weight coefficients, which generally are the extremum of given function, e.g., the error variance.

The aim of the present paper is to prove that under appropriate conditions the trajectories θ_n as function of time converge to the trajectories of the stochastic diffusion process. These are continuous functions satisfying the Fokker-Planck equation. This characteristic provides a tool for the determination of properties of the limiting trajectories, such as their maxima. Furthermore, the optimisation problem of the best weight coefficients-satisfying the unbiased and minimum variance estimate- is solved. Finally, for an illustration of the feasibility and usefulness of this method a numerical experiment is performed. The present work is based on the classical theorem of convergence of Markov chains to diffusion processes [3].

2 Main definitions and notations

Throughout the paper, the following notations are used. Let the system of equations

$$\frac{\partial \theta(t)}{\partial t} = \Lambda(\theta, t) \qquad (1)$$

be considered on the time-interval (t_0, T). Without loss of generality, t_0 in further references will be associated with 0, while T may be both finite and infinite. In (1), $\theta(t)$ represents the random state-vector of dimension *r* defined on a given probability space, while $\Lambda(x,t)$ denotes non-random, generally non-linear operator, acting in R^r, which does not explicitly involve temporal derivatives. Symbol {′} denotes the transpose of a corresponding vector and/or matrix, the symbols | | or || || represent a vector or matrix norm, respectively. A sequence of time-series is considered, such that in each series the interval (*0,T*) is broken into time-points ($0 = t_{0,n}, t_{1,n}, t_{2,n}, \ldots t_{k,n}, \ldots$), where the first index denotes the ordering number of the corresponding point, while the second index refers to the time-series. It is supposed that in each series on the interval $\Delta t_{k,n} = t_{k+1,n} - t_{k,n}$, the number of random vectors ($\xi_1^n, \ldots \xi_l^n, \ldots$), *l* = 0, 1,… ν^n with dimension *q*, $q \le r$ are entered, where ν^n is also a random integer variable with given distribution $p_l^n = P(\nu^n = l)$, independent of vectors ξ_l^n. Knowing the solution of system (1), $\theta_k^{n,m}(t)$, for the initial vector, θ_k^n on entire interval $\Delta t_{k,n}$, the observed variables ξ_l^n as well as a realisation of random index ν^n, the newly constrained variables are introduced by the formula

$$\zeta_k^n = \sum_{l=0}^{\nu^n} \alpha_{l,n} (\xi_l^n - \theta_l^{n,m}) \qquad (2)$$

In (2), the matrixes $\alpha_{l,n}$ with dimension *r* x *r* referred to as weight coefficients are supposed to be known and depend on the time-series.

For the consistency of (2), it is necessary that the vectors ξ_l^n and $\theta_k^{n,m}(t)$ have the same dimension *q*, which may differ from the dimension of the vector $\theta(t)$ of dimension *r*. However, without loss of

generality, it will be assumed that ζ_k^n has the same dimension as θ_k^n, setting the "dummy" components of vectors ξ_l^n and $\theta_k^{n,m}(t)$ to zero. Ultimately, the new state variables θ_{k+1}^n are defined by the formula

$$\theta_{k+1}^n = \theta_k^n + \int_{t_k}^{t_{k+1}} \Lambda(\theta_k^n, \tau) d\tau + \zeta_k^n \qquad (3)$$

Now this θ_{k+1}^n is taken as the initial conditions in (1) for the continuation of the integration.

In this manner, it becomes possible to obtain the sequence of trajectories $\theta^n(t)$ defined over entire interval (*0,T*). Starting from some known random vector θ_0^n, the solution of (1) on each interval $\Delta t_{k,n} = t_{k+1,n} - t_{k,n}$ with breaks at moments $t_{k,n}$ can be evaluated. The main goal of the present paper is to determine the limiting behaviour of solution (3) when $n \to \infty$.

3 Main formulations

The time-lattice $\Delta t_{k,n} = t_{k+1,n} - t_{k,n}$ is considered as non-equidistant, real values in each series. The following statements are introduced.

A1. The intervals $\Delta t_{k,n} \to 0$ approach zero uniformly with respect to *k*, i.e. $\underbrace{\max}_{k} \Delta t_{k,n} \to 0, n \to \infty$.

A2. The probability distribution for random variables ν^n satisfies the conditions: $p_{l,n} = P(\nu^n = l) = p_l \Delta t_{k,n} + o(\Delta t_{k,n}), l > 0$, for any *k*, and $\mu = \sum_{l=1}^{\infty} l p_l < \infty$.

A3. Random vectors $\xi_1^n, \ldots \xi_l^n, \ldots$ have two plus δ for some δ moments for each *n*, i.e. $E\xi_i^n = \lambda_{i1}^n, E(\xi_i^n \xi_j'^n) = \gamma_{ij}^n$, $E \mid \xi_i^n \mid^{2+\delta} < \infty$, and these variables uniformly bounded with respect to *n*, i.e. $\overline{\lim} \mid \lambda_{i1}^n \mid < \infty, \overline{\lim} = \| \gamma_{ij}^n \| < \infty$, $\gamma_{i2}^n = \gamma_{ii}^n$, *i,j=1,2….*

A4. Operator $\Lambda(x,t)$ is a continuos function of its arguments.

A5. The set of weight coefficients $\alpha_{l,n}$ are bounded uniformly with respect to *n*, $\overline{\lim} \| \alpha_{l,n} \| < \infty$, *l=1,2…*

A6. The sequence of distributions of random variables θ_0^n converges to the distribution of some random variable θ_0, i.e. $P(\theta_0^n < x) \to P(\theta_0 < x), n \to \infty$ for each *x* as $n \to \infty$.

Without loss of generality, the limit values of the variables $\lambda_{i1}^n, \gamma_{ij}^n$ and $\alpha_{l,n}$ when *n* tends to infinity are supposed to exist and to be equal, respectively, to $\lambda_{i1}, \gamma_{ij}$ and α_l. Otherwise, the corresponding subsequence can be chosen to provide the convergence to the limit points, which exist as a consequence of the conditions A3 and A5.

3.1 Theorem 1

Let the conditions A1-A6 hold. Then the sequence of finite-dimensional distributions of random processes $\theta^n(t)$ converges to the stochastic process, which will be a solution of the stochastic differential equation

$$\theta(t) = \theta_0 + \int_0^t a(s, \theta(s)) ds + \int_0^t b^2(s, \theta(s)) dw \qquad (4)$$

where

$$a(t,x) = \sum_{l=1}^{\infty} p_l [\sum_{j=1}^{l} \alpha_j (\lambda_{j1} - x)] + \Lambda(x,t), \qquad (5)$$

and

$$b^2(t,x) = \sum_{m=1}^{\infty} p_m \{ \sum_{i,j=1}^{m(m+1)} \alpha_i [\gamma_{ij} - (x\lambda_{1,i}' + \lambda_{1,j} x') + xx'] \alpha_j' \} \qquad (6).$$

The Wiener process *w(t)* is defined on the interval (*0,T*) and is independent of random variable θ_0.

Remark. All parameters $p, \alpha, \lambda, \gamma$ depend, in general, on both *t* and *x*. To simplify the notation, this dependence is not explicitly shown. This theorem is easily generalised on the case when a random index ν is a vector with different distributions for each component, i.e. $p(\nu = l) = p(\nu_1 = l_1, \ldots \nu_r = l_r)$.

Corollary 1.

If no observations are assimilated, *a(t,x)=Λ(t,x) and b(t,x)=0.* Thus, the limiting diffusion process coincides with the model simulation, which has to be expected..

Corollary 2.

The probability distribution of the trajectory is determined by the Fokker-Planck equation (Kolmogorov's second equation)

$$\frac{\partial p(t,x)}{\partial t} = -\frac{\partial (a(t,x) p(t,x))}{\partial x} + \frac{1}{2} \frac{\partial^2 (b^2(t,x) p(t,x))}{\partial x^2} \qquad (7)$$

with the initial conditions $p(0,x) = p(\theta_0 = x)$ and boundary conditions $p(t, \pm\infty) = 0$.

In (7) the drift vector $a(t,x)$ and the diffusion matrix $b^2(t,x)$ are given by (5) and (6).

4 Optimisation problem

In the previous section, the convergence of a sequence of random variables was considered where all parameters were fixed. However, in practical applications, some parameters may not be known, but rather be sought according to a given criteria. The most relevant physical problem is the determination of optimal weight coefficients α_l. By Theorem 1, the limiting trajectories will be those of the diffusion process, with the drift vector $a(t,x)$ and the diffusion matrix $b^2(t,x)$, which determine its mean and variance. More precisely, the following equalities are valid:

$$\frac{\partial}{\partial t}E(\theta(t)/\theta_0)=\int_{-\infty}^{\infty}a(t,x)d_x p(t,x)$$

$$\frac{\partial}{\partial t}\mathrm{var}(\theta(t)/\theta_0)=\int_{-\infty}^{\infty}b^2(t,x)d_x p(t,x)$$

These relations lead to the investigation of the optimisation problem under the following constrains: Let the scheme of the previous paragraph be considered. The conditional average and conditional variance of each of the member of the Markov chain will be a function of weight coefficients $\alpha_{l,n}$. The problem is to define those coefficients such that the variance is minimised while the average remains unchanged. This is the well-known problem of constraining the unbiased estimator with minimum variance.

4.1 Theorem 2

Let the conditions of Theorem 1 A1-A6 hold. Also, let, $\alpha^*_{l,n}$ be the coefficients that provide the minimum norm of conditional variance $(\Delta t_{k,n})^{-1}\,\|\,E[(\theta_k^n)(\theta_k^n)'/\theta_0]\,\|$, while the conditional average $(\Delta t_{k,n})^{-1}E[(\theta_k^n)\ /\theta_0]$ remains fixed and equals *C* in each series. The numerical sequence $\alpha^*_{l,n}$ will then converge to the coefficients α^*_l, satisfying the system of equations with the unknown vector $\Phi=(\varphi_1,...\varphi_r)$.

$$\sum_{m=1}^{\infty}p_m\sum_{j=1}^{m}\sum_{s=1}^{r}[g_{(ij)qs}\alpha'_{(j)qs}+\varphi_s(\lambda_{(1,i)s}-x_s)]=0 \quad \text{for each } i=1,..m;\ q=1,..r$$

$$g_{ij}=\gamma_{ij}-(x\lambda'_{1,i}-\lambda_{1,j}x')+xx'$$

$$\sum_{l=1}^{\infty}p_l[\sum_{j=1}^{l}\alpha_j^*(\lambda_{j1}-x)]+\Lambda(x,t)=C \qquad (8)$$

where $a_{(i)s}$ denotes the *s*-th component of the vector a_i; $b_{(i)qs}$ denotes the *qs*-th element of matrix b_i.

Given any constant *C*, system (8) has a unique solution for α^*_l and Φ.

Remark. Statistically, the constant *C* measures the model bias with respect to the observations. This bias can be eliminated from the algorithm if it is known or if it is previously determined.

5 Numerical experiment

In this section, a numerical experiment is presented as illustration of the applicability of theorems 1 and 2. The heat equation is utilised as part of the hidro-dynamical numerical model AusCOM, a version of the Geophysical Fluid Dynamics Laboratory (GFDL/NOAA) model MOM4,

$$\frac{\partial\theta(t,\bar{x})}{\partial t}=-\nabla(\bar{u}\,\theta(t,\bar{x}))+k^2\nabla^2\theta(t,\bar{x}) \qquad (9).$$

Here, $\theta=\theta(t,\bar{x})$ is temperature; $t,\bar{x}$ denote temporal and spatial coordinates, $\bar{u}$ is a known velocity vector, ∇,∇^2 are the gradient and the Laplace operators, respectively, and k^2 is a given viscosity coefficient. The bar over *x* distinguishes a 3-dimensional spatial point from its coordinate. Considering equation (1)

$$\frac{\partial\theta}{\partial t}=\Lambda(\theta,t),$$

equation (9) takes the form

$$\Lambda(\theta,t)=-\nabla(\bar{u}\,\theta(t,\bar{x}))+k^2\nabla^2\theta(t,\bar{x}).$$

Starting from known initial conditions θ_0, equation (9) is solved numerically. Independently, three temperature profiles are chosen from the TOGA/TAO array dataset (downloaded from http://www.pmel.noaa.gov/tao). The scheme proposed above addresses infinitesimal characteristics. Hence, temperature variability on the time-frame of one day (24 hours) is considered.

Fig1 shows the temperature profiles of the differences between two consecutive days of observational data, taken at the coordinates: (156 °45'E, 0° 30'N); (156° 52'E, 1° 0'N); (157° 24'E, 0°36'N), together with the model profile. The model was started from the average over these 3 temperature profiles and the model results is considered at the point with coordinates (156,5° E, 0.3°N). On the AusCom grid, this is the nearest point to the

observations. Model profile is given by the red curve, while observations are in green, white and yellow.

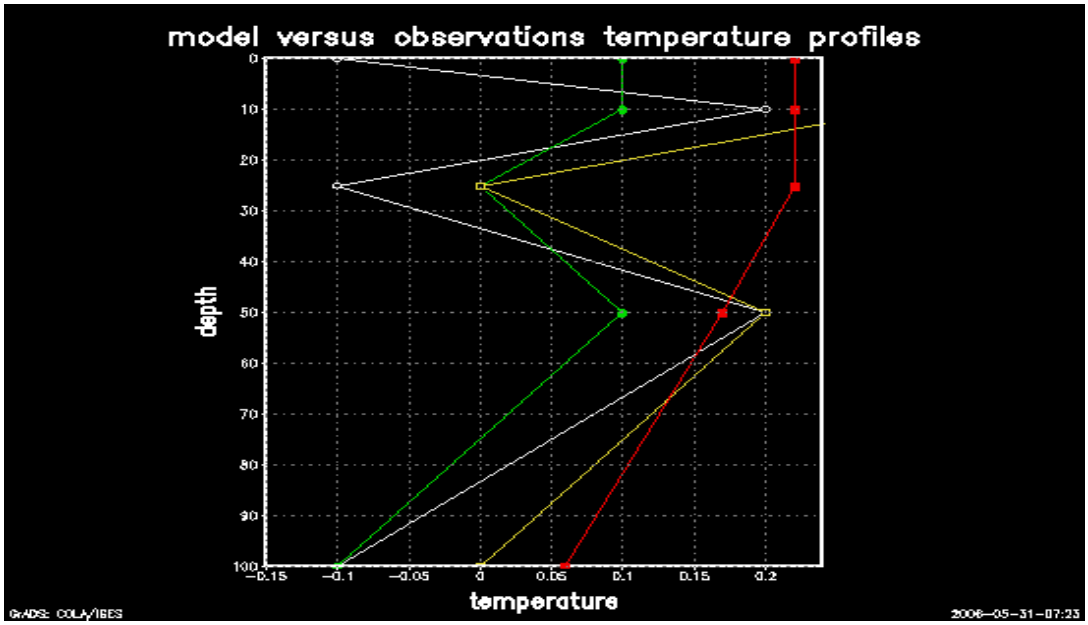

Fig.1. Vertical profiles of temperature differences given by observations at (156 °45'E, 0° 30'N); (156° 52'E, 1° 0'N); (157° 24'E, 0°36'N) in green, white and yellow; and by the model at (156,5° E, 0.3°N) in red.

The optimisation scheme has been applied to the vectors ξ_1, ξ_2, ξ_3, which are the temperature differences profiles with 5-components corresponding to temperature values at 0, 10, 25, 50 and 100 m. The probability parameters were $p_3 = 1, p_i = 0, i \neq 3$. All other necessary parameters required have been determined from observations and model. The calculations with respect to formulae (8), then (2) and (3) lead to the results presented in Fig. 2.

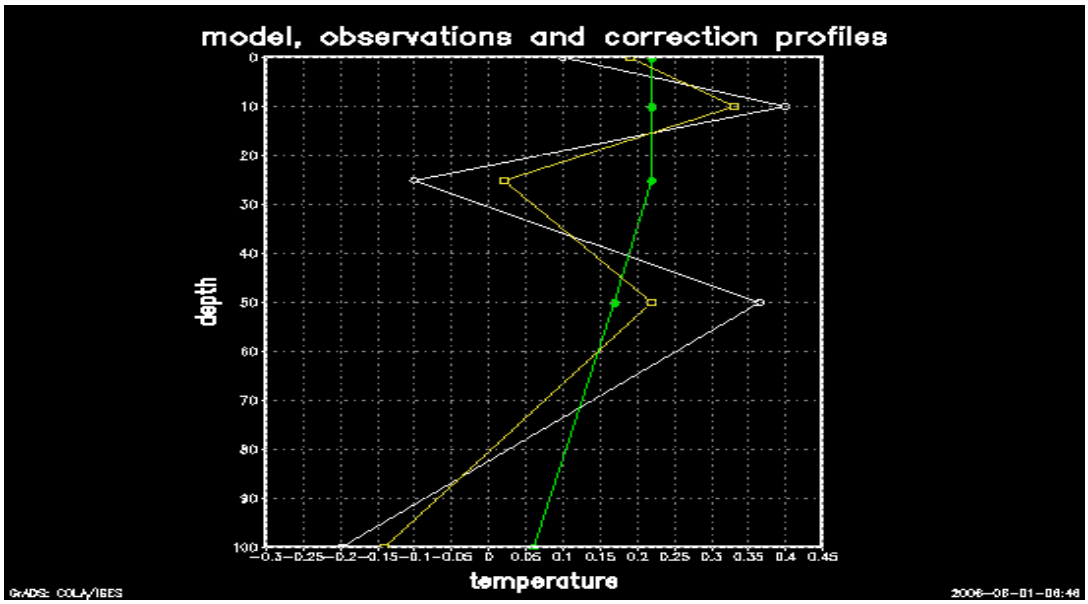

Fig.2. Vertical profiles of temperature given in white by the average of the three observed temperature profiles shown in Fig. 1; in green by the model; and in yellow by the data assimilation scheme.

Fig. 2 shows the average profile of the three temperatures shown in Fig. 1, marked by white cycles. The model profile is shown in green and the optimal correction followed from (8), (2) and (3), in yellow. It

is clearly seen that the optimal curve lies between the model and observation and it resembles the shape of observation curve. This is to be expected since, by construction, the optimal curve conserves the average. It is noted that one realisation among possible profiles is shown and its trajectory has a definite probability according to the Fokker-Planck equation (7).

6 Conclusions

The results presented here are of both theoretical and practical interest. From the theoretical point of view, it is important to know when and under which conditions the curve, given as a solution of differential equations (1) along with the correction (2), may be approximated by a continuos function, and how the probability of the limit function could be calculated. In practical applications, the scheme can be utilised to determine a variety of relevant parameters such as the probability of extremes, probability for crossing some levels and similar issues. Also, the optimisation scheme may open a new approach to the data assimilation problem, since the method merely requires the solution of linear system of equations with simple-form matrix. However, a number of non-trivial questions remain unsolved. For instance, how is the distribution of statistics to be calculated if all parameters are taken from the observed sample? Other topics refer to the applicability of the method. It is clear, for instance, that two consecutive assimilations should be fairly close in time in order to apply the limit theorem. This and other issues may be investigated in a later time and were not within the scope of the present work.

References

[1] M. Gill, P. Malanotte-Rizoli. Data assimilation in meteorology and oceanography. *Adv. Geophysics,* Vol. 75, 1991, pp. 1-435.

[2] I. Gikhman, A.Skorokhod, An introduction of a random process theory, *, M. Nauka,* 1966, 350 pp. (translation from Russian).

[3] D.Strook, S.R.S.Varadhan. Multidimensional random processes. *Springer-Verlag. Berlin,* 1995, pp. 233.