Embedding reality in a numerical simulation with data assimilation

Tomoyuki Higuchi
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
Email: higuchi@ism.ac.jp

Abstract—Data assimilation (DA) is a synthesis technique based on the Bayesian filtering method by embedding observation/experiment data in a numerical simulation. It yields an accommodation ability to make a simulation real, and the better initial and boundary conditions can be automatically obtained. In statistical methodology, DA can be formulated in the state space model that draws much interest of the researchers in various domains such as the time series analysis, signal processing, and control theory. There are two types of DA in terms of a methodology; sequential DA and variational (non-sequential) DA. An ensemble-based sequential DA (EnSDA) has an advantage in terms of less human resources which is achieved by plugging into the existing "forward" simulation codes. We briefly explain a recent advancement in EnSDA, and give a simple description on the relationship among the nonlinear non-Gaussian filters. Keywords: Nonlinear state space model, particle filter, Tsunami simulation, Systems biology, peta-scale computing.

I. INTRODUCTION

Scientists are faced with mountains of data that stem from the flood of data from new scientific instruments and from simulation results. The ability to economically store these data on line and Internet service boost a rate of data exploration as time goes. To accommodate with these trends, data-centric science emerges as the fourth generation in the history of science, namely starting from empirical (experimental) science (E), theoretical science (T), simulation science (S), to data-centric science (D) [11]. In the data-centric science, to build the cyber-enabled discovery system is one of the ambitions and challenging research target. A schematic illustration of these four methodologies to conduct a scientific research is shown in Figure 1. The data-centric science is depicted as the massive data analysis in this figure.

Data Assimilation (DA) is a technique for a synthesis of information from a dynamic numerical model and observation data, and is stimulated by recent improvements in computational and modeling capabilities and the increase in the amount of available observations [7], [13], [21]. DA plays an important role as an axle to connect the two front-wheels in Figure 1; simulation (S) and massive data analysis (D). A simulation model cannot represent real phenomena accurately. It is easily understood simply if we notice that there are too many uncertainties in the model such as the boundary condition, initial condition, unknown parameters, and unknown dynamics. On the other hand, we are allowed to conduct a direct observation on only a small part of the physical variables in the numerical model due to the physical and/or budgetary restrictions. It is therefore reasonable to combine these two different information sources: simulation result and observation data.

DA is an emerging area in earth sciences, particularly oceanography. In statistical methodology, DA can be formulated in the state space model that draws much interest of the researchers in various domains such as the time series analysis, signal processing, and control theory [5], [9], [19], [20]. As a result, DA can be applicable to any scientific domains involving numerical simulation models. Major objects of DA can be categorized in the following five aspects.

1) To produce the best (better) initial condition for forecasting. It is actually realized in the real weather forecast.
2) To find the best (better) boundary condition in constructing a simulation model. This procedure includes a setting of appropriate boundary conditions necessary for dealing with the coupled phenomena.
3) To attain an optimal parameter vector that appears in an empirical law (scheme) employed for describing complicated phenomena which possesses the different time and spatial scales. A validation of the empirically given values is regarded as this problem.
4) To inter/extrapolate (estimate) physical quantity at times and locations without observations based on a numerical simulation model. This procedure is called “a generation of re-analysis dataset (product)”. This dataset is used to
discover a new scientific finding by general geophysical researchers.

5) To conduct an experiment with a virtual observation network and perform a sensitivity analysis in an attempt to construct an effective observation network system with less budgetary cost and less consuming time.

These objects in DA can be realized by a state estimation problem within a framework of the state space model. However, to realize DA is hampered by a large dimension of the state vector, which is much larger than that of the observation vector. To deal with this difficulty, there are two approaches of DA: four-dimensional variational method (4Dvar) and ensemble-based sequential DA (EnSDA). 4Dvar is regarded as an off-line estimation procedure because it is applied to the fixed data set. Meanwhile, EnSDA is an online estimation procedure because an estimator is improved every time we have a new observation. We briefly explain a recent advancement in EnSDA and demonstrate a part of applications carried out by our DA research group. Finally, we discuss on what and how we develop the EnSDA methods for realizing DA on high performance computing (HPC).

II. MATHEMATICAL FORMULATION OF DATA ASSIMILATION

A. Numerical simulation model

A numerical simulation model is usually carried out with the discrete version of a partial differential equations (PDE) which is assumed to represent a real system. The discretization in time and space for PED produces the finite difference equations (FDE) given by

$$ x_t = f_t(x_{t-1}, v_t) $$

(1)

where $x_t$ is called the state vector and defined by setting all variables at discrete time $t$ in a numerical simulation to an element of $x_t$. For example, suppose a case where we conduct a two-dimensional simulation experiment for understanding the flow of shallow water such as tides, storm surges, river flow, and tsunami. A space discretization generates a discretization grid with $M$ grid points on water surface in a limited region. All physical variables on the $m$th grid point at time $t$ is set to be an element of vector $\xi_{m,t}$. As for a shallow water simulation, $\xi_{m,t}$ consists of depth $d_m(t)$, sea surface height $\eta_m(t)$, which is measured from the average sea surface, and the two components of the two-dimensional water flow vector $(U_m(t), V_m(t))$:

$$ \xi_{m,t}^p = [d_m(t), \eta_m(t), U_m(t), V_m(t)] . $$

(2)

These variables at the $m$th grid point are depicted in Figure 2. The depth $d_m(t)$ is taken from the bottom topography data set. In fact, this data set is known to be erroneous and then treated as a state variable for adjustment. A detailed explanation is shown in [23], [24] for this example.

$x_t$ is formed by concatenating all $\xi_{m,t}$ into a state vector;

$$ x_t^p = [\xi_{1,t}^p, \ldots, \xi_{m,t}^p, \ldots, \xi_{M,t}^p] . $$

(3)

Here $t^p$ denotes a transposition. A dimension of a state vector $x_t$ is denoted by $n_x$. When a dimension of $\xi_{m,t}$ is $n_{\xi}, n_x$ is $n_x = M \cdot n_{\xi}$. Then, for a case with $M$ being very large, a dimension of the state vector becomes huge size, leading to a difficulty in making a statistical inference. In Eq. (1), $v_t$ called a system noise is introduced to allow for flexible representation of the uncertainties involved in a simulation model. For instance, a time-dependent change in the boundary condition can be attributed to a stochastic perturbation induced by $v_t$. In usually, DA employs a normal distribution with the prescribed variance-covariance matrix $Q$ for a distribution function of $v_t$ because of a simple computation for drawing a random vector of $v_t$. A dimension of a state vector $v_t$ is denoted by $n_v$.

B. State space representation

A usual setting of DA assumes that observations are obtained by a partial measurement of the state vector or a weighted linear combination of the elements with a measurement error. Thus an observation vector at time $t$, $y_t$, is represented with a prescribed matrix of $H_t$ as follows;

$$ y_t = H_t x_{t-1} + w_t, $$

(4)

A dimension of $y_t$ is denoted by $n_y$. Here $w_t$ is assumed to follow a multivariate normal distribution of the prescribed variance-covariance matrix of $R$. Eq. (4) is called the observation model. Accordingly, $H_t$ becomes a $n_y \times n_x$ matrix. A combination of Eqs. (1) and (4) yields the state space model (SSM)

$$ \left\{ \begin{array}{l} x_t = f_t(x_{t-1}, v_t), \quad v_t \sim N(0, Q) \quad \text{System model} \\ y_t = H_t x_t + w_t, \quad w_t \sim N(0, R) \quad \text{Observation model} \end{array} \right. $$

(5)

where $N(0, Q)$ and $N(0, R)$ denote a normal distribution with a mean 0 and variance-covariance matrix of $Q$ and $R$, respectively [5], [9], [19], [20]. An initial distribution of $x_0$, $p_{\text{init}}(x_0)$ has to be specified beforehand. When $p_{\text{init}}(x_0)$ is given by a delta function such that $p_{\text{init}}(x_0) = \delta(x_0 - x_{\text{init}})$, $x_{\text{init}}$ corresponds to an initial condition in a usual simulation.
In data assimilation, $n_x$ is huge such as $n_x \approx 10^5 \sim 10^6$. In contrast, $n_y$ is much smaller than $n_x$ ($n_x \gg n_y$), and then to get an optimal solution for a sequence of $x_t$ becomes a time-dependent inversion problem. It should be noticed that $n_y$ is extremely larger than those in the conventional state space models. For example, $n_y$ exceeds a few thousands.

III. TWO APPROACHES IN DATA ASSIMILATION

A. MAP estimate and four-dimensional variational method

The information from the time series obtained by time $j$ is denoted by $y_{1:j}$. The number of data points of time series is specified by $T$, and then we have $y_{1:T}$. The set of the realizations of the state vector $x_t$ up to time $j$ is denoted by $x_{1:j}$ similarly. Under an assumption of Eq. (5), the joint probability of $x_{1:T}$ and $y_{1:T}$ is given by

$$p(x_{1:T}, y_{1:T}) = p_{\text{init}}(x_0) \prod_{t=1}^{T} p(y_t|x_t)p(x_t|x_{t-1}). \quad (6)$$

For a given data set $y_{1:T}$, the posterior distribution of a sequence of the state vector, $p(x_{1:T}|y_{1:T})$ is proportional to the joint distribution of Eq. (6). So that one of ways to determine the sequence of the state vector is to maximizes the joint distribution to attain a maximum a posterior (MAP) solution:

$$x_{1:T}^{\text{MAP}} = \arg\max_{x_{1:T}} p(x_{1:T}, y_{1:T}). \quad (7)$$

The most popular method to obtain the MAP solution in DA is the four-dimensional variational analysis (4DVar) which is based on a variational method. In particular, the adjoint method based on a method of steepest descent, is usually employed in an oceanography and meteorology.

B. Marginal distribution and ensemble-based filter

As we might see, the 4DVAR is an optimization technique, and then its notion of making a state estimation can be easily accepted by geophysical researchers. However, there remains difficulty in applying the adjoint method to a wide variety of problems in any scientific domain, because the adjoint method requires much human resources to develop an original code for each numerical simulation. On the other hand, an ensemble-based sequential DA (EnSDA) is not the optimization technique but a procedure for making a state estimation through a marginal distribution, $p(x_t|y_{1:t})$, where $0 \leq t \leq T$. A representation with SSM brings in the various benefits in computations for obtaining the marginal distribution. In particular, any marginal distribution can be obtained with the recursive formulas: non-Gaussian filtering and non-Gaussian smoothing algorithms [4], [12], [19]. These algorithms can be realized with a numerical approximation to any conditional distribution, and result in a recursive numerical computation. Unfortunately, an enormous dimension of the state vector makes the recursive numerical computation impossible to apply. To overcome this intractableness, the particle filter (PF) has proposed in a statistical community [8], [17]. A basic part in PF’s algorithm is the same as that in the Condensation algorithm in a computer vision community [16].

IV. ENSEMBLE-BASED SEQUENTIAL DATA ASSIMILATION (ENSDA)

A. Basic algorithm of EnSDA

The EnSDA has an advantage in terms of less human resources which is achieved by plugging into the existing “forward” simulation codes, and now widely used in a wide variety of scientific domains instead of the 4DVar. In particular, the ensemble Kalman filter (EnKF) [6], [7] becomes very popular in the EnSDA because of having less difficulties in a numerical computation rather than other methods in EnSDA. Although the PF has a drawback in terms of an ability to approximate the conditional distribution, both simple idea and less computational burden are very attractive. Here we give a brief explanation for the EnSDA with the PF. Each ensemble member is called the particle in the PF. A basic scheme of the EnSDA is as follows. In this method, two conditional density functions at time $t$, i.e., one ahead predictive density and filter density are approximated by many (say $m = 10^2 \sim 10^3$) particles that can be considered as independent realizations from that distribution. Namely,

$$X_{t|t-1} = \{x_{t|t-1}^{(i)}\}_{i=1}^{m} \approx p(x_t|y_{1:t-1}) \quad (8)$$

$$X_{t|t} = \{x_{t|t}^{(i)}\}_{i=1}^{m} \approx p(x_t|y_t). \quad (9)$$

It can be shown that these particles can be generated recursively by the following algorithm.

1) For $i = 1, \ldots, m$, generate $n_x$ dimensional random number $x_{0|0}^{(i)} \sim p_{\text{init}}(x_0)$

2) Repeat (a)-(c) the following steps for $t = 1, \ldots, T$.

   a) Repeat (i)-(iii) for $i = 1, \ldots, m$.
      i) Generate $n_v$ dimensional random number $v_t^{(i)} \sim N(0, Q)$ for system noise.
      ii) Run a simulation and obtain the predictive particle by $x_{t|t-1}^{(i)} = f_t(x_{t-1|t-1}^{(i)}, v_t^{(i)})$.
      iii) Compute $w_t^{(i)} = p(y_t|x_{t|t-1}^{(i)}) = N((y_t - H_t x_{t|t-1}^{(i)}), Q)$.

   b) Calculate the total weight $W_t = \sum_{i=1}^{m} w_t^{(i)}$

   c) Obtain the filter particle by the sampling with replacement from $\{x_{t|t-1}^{(i)}\}_{i=1}^{m}$ with sampling probabilities $\bar{w}_t = w_t^{(i)}/W_t$.

Here $y_{1:0}$ means an empty set: $y_{1:0} = \phi$.

We give an illustration for explaining the algorithm in Figure 3. The top panel shows a filter distribution at time $t-1$, $p(x_{t-1|t-1})$. Only ten particles are used to approximate $p(x_{t-1|t-1})$ for understanding the algorithm visually. Namely, $m = 10$. At time $t$, a system noise ball $v_t^{(i)}$ for each particle is drawn from $N(0, Q)$, and hit to the filter particle $x_{t-1|t-1}^{(i)}$. Then, the predictive particle, $x_{t|t-1}^{(i)}$ is obtained through performing a numerical simulation with $x_{t-1|t-1}^{(i)}$ and $v_{t}^{(i)}$. The second top panel shows a predictive distribution at
time $t$. It is seen that a location of ten particles is changed. Once we observe a new data $y_t$, the fitness of each particle to a new data is evaluated with $p(y_t|x_{t-1}^{(i)})$. This fitness is called a particle predictive likelihood. An area of the circle in the bottom panel indicates the fitness. Larger circles correspond to particles with higher fitness to data. The resampling procedure of the predictive particles produces a new ensemble each of which member is located on the bottom horizontal line in the figure. The cross indicates the vanished particle due to its lower fitness to data. On the other hand, the predictive particle with high fitness generates its clone and then, two or three identical particles have been produced as seen in the figure. It should be noticed that a total number of particles $m=10$ is unchanged.

**B. Ensemble Kalman filter (EnKF)**

What EnKF makes a difference from PF is in a filtering process. In other words, a prediction process (2.a) is common to PF and EnKF. The update process performed on the predictive particle in EnKF is replaced by the Kalman filter instead of the resampling procedure in PF. It should be noticed that the Kalman gain in EnKF is defined by a sample-based approximation of a similarity of distributions such as the Kullbuck-Leibler information criterion. So that, once the ensemble of the EnKF shows a different behavior from the realizations from the true distribution, even if they keep the same first and second moments, this deviation is enhanced by the following prediction and update processes. Therefore, the ensemble obtained by the EnKF fails to resemble a true distribution in a metric of similarity. This fact suggests that a likelihood value with Monte Carlo evaluation through the ensemble of EnKF tends to be biased.

Figure 4 demonstrates a relationship among the PF, EnKF, and GSF from a viewpoint of operations to deal with an element to approximate PDF.

c) Repeat (i) and (ii) the following step for $t = 1, \ldots, T$.

i) Draw a $n_w$ dimensional random number $w_{t}^{(i)} \sim N(0, R)$.

ii) Update the particle by $x_{t|i} = x_{t|t-1}^{(i)} + K_t(y_t + w_{t}^{(i)} - x_{t|t-1}^{(i)})$.

**C. Relationship among the non-Gaussian filters**

The EnKF has been proposed to track the first and second moments of a true conditional distributions. However, it is not designed to approximate the true distribution in terms of a similarity of distributions such as the Kullbuck-Leibler information criterion. So that, once the ensemble of the EnKF shows a different behavior from the realizations from the true distribution, even if they keep the same first and second moments, this deviation is enhanced by the following prediction and update processes. Therefore, the ensemble obtained by the EnKF tends to be biased.
Concerning to a family of the non-Gaussian filters with an element, there are the Gaussian particle filter (GPF) and Gaussian sum particle filter (GSPF) [3]. The GPF approximates the density function by a Gaussian distribution, but it differs from the extended Kalman filter (EKF) in terms of its update scheme for approximations. While the EKF employs a linearization of the nonlinear functions, the GPF obtains statistical information for an approximation with the particles. The GSPF is an natural extension of this methodology with the Gaussian mixtures in stead of a Gaussian in GPF. It should be noted that an usage of GSM and GSPF is limited to a case of the nonlinear but additive non-Gaussian noise models. The mixture Kalman filter (MKF) proposed by [2] relies on an assumption of Gaussian mixtures approximation to the statistical information for an approximation with the particles, are usually adopted. The maximum likelihood approaches, maximum likelihood method and state vector augmentation, are usually adopted. The maximum likelihood method maximizes the Monte Carlo-based log-likelihood value with respect to a parameter vector, which is defined by

$$ p(y_{1:T} | \theta) = \sum_{t=1}^{T} \log p(y_t | y_{1:t-1}, \theta) $$

$$ \cong \sum_{t=1}^{T} \log W_t - T \log m $$

(10)

where $W_t$ already appears in the explanation for a basic structure of the EnSDA algorithm. In a description of the EnSDA, a presence of $\theta$ is suppressed because of a simplicity. An optimal parameter vector can be achieved by maximizing the Eq. (10) with respect to $\theta$. An alternative way to make an estimate on $\theta$ is to augment the state vector $x_t$ with the parameter vector $\theta$

$$ \tilde{x}_t^{\theta} = [x_t^p, \theta^p]. $$

(11)

To give a prior distribution on $\theta$, $p_{\text{prior}}(\theta)$, enables us to apply the PF with $\tilde{x}_t$. As a result, we can obtain a posterior distribution of $\theta$ by marginalizing the filter distribution at time $T$, $p(\tilde{x}_T | y_{1:T})$:

$$ p(\theta | y_{1:T}) = \int p(\tilde{x}_T | y_{1:T}) d\tilde{x}_T. $$

(12)

This approach is called the self-organizing state space model [18].

VI. APPLICATIONS

We are studying the EnSDA methods such as the EnKF and PF, and conducting the DA experiments in five specific areas: atmosphere-ocean system, tsunami, ocean tide, outer space, and genome information. We demonstrate a part of applications carried out by our DA research group.

One is done with the EnKF that assimilates the TOPEX/Poseidon altimeter to the coupled ocean-atmosphere simulation model [29]. The second DA is done with a particle filter for Tsunami simulation model to correct bottom topography [24]. The third is the DA project to estimate distributions of ring current ions and electric potential in the inner magnetosphere by assimilating the series of the ENA data obtained by the HENA imager onboard of IMAGE satellite into a kinetic ring current model (CRCM) [26]. Ocean tide DA experiment is performed to correct the bathymetry data which is based on a estimation from satellite gravity observations [15]. The last project is to apply the DA methodology to combine a simulation model with observed data like microarray gene expression data for understanding biological pathways [22].

Here we demonstrate one of the DA experiments among five categories [15]. Figure 6 demonstrates a simulation result for an accurate ocean tide modeling in southeast Alaska where there are many coastal and inland seas with complicated coastlines and bathymetry. The numerical simulation model uses a regional, barotropic ocean model with a finite difference
scheme. Bathymetry, i.e., sea depth data is treated as an unknown parameter in the DA experiment, because there is much uncertainty on the bathymetry data which is estimated from satellite gravity observations. The bottom friction coefficient, which plays very important role in a fundamental rule (i.e., Navier-Stokes equation), is also treated as the unknown parameter. The tide gauge data observed at 11 coastal sites in this region is used for this DA to reduce uncertainty involved in the simulation. Corrected parameter vector provides us with much improved prediction as to the ocean tide in amplitude and phase [15].

Our recent research interests are also focused on a development of the EnSDA methods for realizing DA on a high performance computing (HPC) using GPGPU (General-Purpose computing on Graphics Processing Units) and massive parallel computers [10], [27]. In particular, we are studying the hierarchical particle filter embedded in the heterogeneous parallel computing platform for data assimilation.

VII. CONCLUSION

We would like to emphasize that a research on the DA can be regarded as the “Creation of meta-simulation model”. We summarize our research direction as follows.

- We automate a procedure searching for better simulation model to describe real phenomena.
- We develop a procedure to generate a new simulation model that has greater ability of predictive performance than existing ones.
- We give consistent view to assessment of simulation model that is said to be subsidiary problem in simulation science; Maximum Likelihood Principle.
- We give a platform to design a measurement system in an attempt to enhance a scientific return together with reducing a total budgetary cost.

Acknowledgment

This research is partly supported by the Japan Science Technology Agency for CREST (Core Research for Evolutinal Science and Technology) project 2005–2010. The team leader for this project is the author, and the members are Dr. Ueno (atmosphere-ocean system DA), Dr. Nakamura (tsunami DA), Dr. Inazu (ocean tide DA), Dr. Nakano (outer space DA), and Dr. Yoshida (genome information DA). The research on HPC in DA is in part supported by the members, Dr. Hayashi and Dr. Saitoh.

REFERENCES


