Journal of Universal Computer Science, vol. 12, no. 6 (2006), 608-626
submitted: 31/5/06, accepted: 23/6/06, appeared: 28/6/06 © J.UCS

Sequential Data Assimilation : Information Fusion of a Numerical Simulation and Large Scale Observation Data

Kazuyuki Nakamura
(The Graduate University for Advanced Studies / JST CREST, Japan
nakakazu@ism.ac.jp)

Tomoyuki Higuchi
(The Institute of Statistical Mathematics / JST CREST, Japan
higuchi@ism.ac.jp)

Naoki Hirose
(Kyushu University, Research Institute for Applied Mechanics, Japan
hirose@riam.kyushu-u.ac.jp)

Abstract: Data assimilation is a method of combining an imperfect simulation model and a number of incomplete observation data. Sequential data assimilation is a data assimilation in which simulation variables are corrected at every time step of observation. The ensemble Kalman filter is developed for a sequential data assimilation and frequently used in geophysics. On the other hand, the particle filter developed and used in statistics is similar in view of ensemble-based method, but it has different properties. In this paper, these two ensemble based filters are compared and characterized through matrix representation. An application of sequential data assimilation to tsunami simulation model with a numerical experiment is also shown. The particle filter is employed for this application. An erroneous bottom topography is corrected in the numerical experiment, which demonstrates that the particle filter is useful tool as the sequential data assimilation method.

Key Words: particle filter, simulation science, data fusion
Category: G.1.0, I.6.0, I.6.4, J.2

1 Introduction

The Indian Ocean tsunami has called attention to the need for further studies on tsunamis. Past studies have been based on numerical simulation models. The simulations and their validation based on obtained data have been conducted separately. In these studies, a sea bottom topography is fixed in the simulation model. However, uncertainty in the sea bottom topography and inaccuracies in the numerical model exist in fact. In general, a numerical simulation model has limit in approximation of physical processes and initial and boundary conditions, which result in unpredictability often referred to as “butterfly effect” [Lorenz (1963)]. On the other hand, observable physical variables are incomplete data because technical and budgetary limitations exist. A natural development

Table 1: Difference of characteristics between DA and other fields.

	DA	other fields
State vector dim.	10^3–10^6	$10^0 - 10^2$
Observation vector dim.	10–10^4	$10^0 - 10^2$
Evolutional model	physical	statistical
System representation	source code	analytic form
Computational cost	high	low

to compensate for insufficient information obtained by numerical simulations or observations alone is to combine observations with the numerical models.

Data assimilation (DA, [Wunsch (1996)]) is a concept used in geophysics that combines observations with numerical models. It can be formulated as a state estimation problem by a nonlinear state space model (SSM). The SSM has given a platform in non-stationary time series ([Kitagawa and Gersch (1996)], [Higuchi (2001)]) and control studies for three decades after [Kalman (1960)]. In this formulation, all physical variables are included in a state vector and observed data are included in an observation vector in SSM. A simulation model is embedded into SSM as evolutional model which makes up system model in SSM. The DA problem is an inverse problem in that there is less information about the observation data than the estimated variables. However, many differences exist and make problem hard. [Tab. 1] represents some typical differences between the DA and other fields. Difficulties with DA are that the scale of the system model is extremely large compared with other fields using an SSM, the simulation model which is based on physical model is often given only by source code and computational cost to solve the simulation model is often high.

There are two types of DA, batch-type DA (4DVAR, [Courtier et al. (1994)]) and sequential DA. We concentrate on sequential DA in this paper. Sequential DA estimates unobserved variables at each observation time. In the context of sequential DA, the Kalman filter (KF, [Kalman (1960)]) and the extended Kalman filter (EKF, [Anderson and Moore (1979)]) were applied to weakly nonlinear problems until the middle of 1990s. However, there are problems in applying them to strongly nonlinear problems, because the state vector cannot be efficiently estimated or the covariance matrices and estimated state variables are liable to be unstable. [Evensen (1994)] proposed an assimilation method for strongly nonlinear problems, called the ensemble Kalman filter (EnKF). It should be noted that the EnKF is different from the EKF. In the EnKF, predictive probability density functions (PDFs) of the state vector are constructed by Monte Carlo simulation. However, the EnKF uses only the first and second moments to construct the filter PDFs and a nonlinear observation model cannot be dealt with directly.

The particle filter (PF, [Kitagawa (1996)], [Gordon et al. (1993)]) has been developed in the field of statistics, which has the same structure in terms of ensemble-based and sequential filtering, but it has some advantages over the EnKF. The PF does not need any assumptions for the PDFs. In addition, it can deal with nonlinear observation models. In spite of these advantages, real applications of the PF to DA have been limited [Manda et al. (2003)], [van Leeuwen (2003)]. Additionally, works to date have assumed linear observations. The motivations of this work are to develop and to apply nonlinear filtering to DA method and to make clear the important points in formulating sequential DA.

In Section 2 we give the framework and characteristics of the sequential DA. Relationship between the EnKF and the PF is also given. An application of sequential DA are presented in section 3 by dealing with a tsunami simulation model. A numerical experiment for validating the framework is described in section 4. Conclusions are given in section 5.

2 Sequential data assimilation

2.1 Embedding of a simulation model

In this subsection, we demonstrate the formulation of a sequential DA problem with an SSM. Additionally, we explain the distinctions between DA and other estimation problems.

[Fig. 1] shows the procedure of design of DA problems. Partial differential equations (PDE) are usually employed to approximate a real physical system in geophysics. First, the PDE are discretized spatially and temporally for calculation on computer. This discretization gives finite difference equations (FDE) which can be represented as

$$\boldsymbol{x}_t = \tilde{f}_t(\boldsymbol{x}_{t-1}), \tag{1}$$

where $\boldsymbol{x}_t$ denotes all the variables included in the FDE at time step t. Many numerical simulation studies in geophysics and other areas are conducted with solving these equations. Such sets of equations are called simulation models. In the next step, the uncertainties of the simulation model, *i.e.*, the boundary conditions, unknown parameters and some kind of model uncertainties, are modeled by introducing system noise $\boldsymbol{v}_t$. Finally, the simulation model and system noise are combined into an equation,

$$\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t), \tag{2}$$

which can be identified with the system model of SSM. f_t is determined from the design how to combine $\tilde{f}_t(\boldsymbol{x}_{t-1})$ and $\boldsymbol{v}_t$. Because the system noise $\boldsymbol{v}_t$ represents

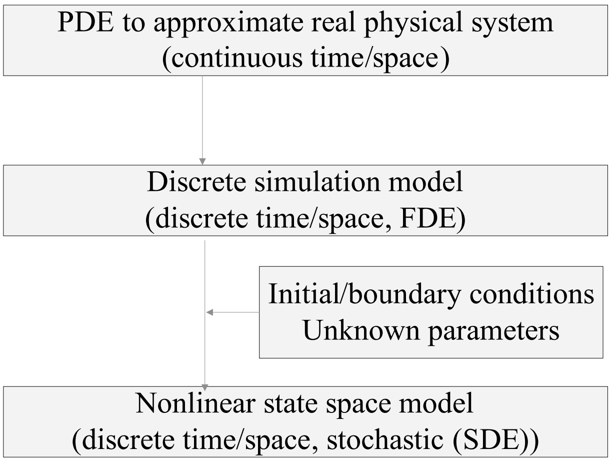

Figure 1: Procedure for constructing system model.

"uncertainties", it can be regarded as random variable whose distribution is usually assumed to be normal:

$$\boldsymbol{v}_t \sim N(\mathbf{0}, Q_t),$$

where Q_t is pre-determined covariance matrix. In some problems, the initial state $\boldsymbol{x}_0$ also has uncertainties and then can be regarded as random variable.

Compared with usual state estimation problems using SSM, there are several points of difference in a DA problem. One of them is that the simulation model is very large and complicated, often given only by computer source code because we should rely on reservoirs of simulation science in geophysics. Consequently, it is hard to change the simulation model drastically. Another point is the large dimension of the state vector. Each grid point has associated physical variables, which means that the dimension of the state vector is the product of the number of grid points and the dimension of the physical variables. For example, if there are 10 physical variables at each point of a two-dimensional 100×100 grid, the dimension of $\boldsymbol{x}_t$ is 10^5. All the differences including the other different points omitted here make the problems hard. They can be overcome however by clever design of the uncertainties, including system noise, and accurate calculation in the prediction steps.

2.2 Observation model

Observations are obtained by the partial measurements of the state vector or their (non)linear transformations with measurement errors. Therefore, observation equation is written by

$$\boldsymbol{y}_t = h_t(\boldsymbol{x}_t, \boldsymbol{w}_t) \quad \text{(nonlinear observation)}, \tag{3}$$

or

$$\boldsymbol{y}_t = H_t \boldsymbol{x}_t + \boldsymbol{w}_t \quad \text{(linear observation)}, \tag{4}$$

where y_t is composed of all the measurements, $\boldsymbol{w}_t$ represents measurement error, h_t is nonlinear operator representing observation and H_t represents observation matrix. The PDF of measurement error is usually normal distribution with average $\mathbf{0}$,

$$\boldsymbol{w}_t \sim N(\mathbf{0}, R_t),$$

where R_t is pre-determined covariance matrix of measurement errors at time t. Equation (3) (or (4)) is called observation model of SSM. When a measurement is made, a measurement error is included naturally. Therefore, if the transformation and measurement error statistics are known, construction of an observation system is straightforward.

2.3 State space model

The nonlinear SSM is summarized as follows:

$$\begin{aligned}
\boldsymbol{x}_t &= f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t), \\
\boldsymbol{y}_t &= h_t(\boldsymbol{x}_t, \boldsymbol{w}_t), \\
\boldsymbol{v}_t &\sim N(\mathbf{0}, Q_t), \;\; \boldsymbol{w}_t \sim N(\mathbf{0}, R_t),
\end{aligned}$$

where Q_t and R_t are pre-determined or estimated covariance matrices. In the following, the dimension of the state vector $\boldsymbol{x}_t$ is denoted by n_x and the dimension of the observation (measurement) vector $\boldsymbol{y}_t$ is n_y. We use $\boldsymbol{y}_{1:t}$ to represent the set $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_t\}$.

[Fig. 2] shows a schematic representation of the data assimilation concept for the special case, i.e. linear simulation model such that

$$\begin{aligned}
x_t &= F_t x_{t-1} + v_t, \\
y_t &= x_t + w_t, \\
&(n_x = n_y = 1).
\end{aligned}$$

At each time the state vector is updated according to this scheme (called the filtering step in the Kalman filter). $\boldsymbol{x}_{t|t-1}$ is a state estimation at time t only via simulations given $\boldsymbol{y}_{1:t-1}$. $\boldsymbol{e}_{t|t-1}$ is a prediction error between an actual observation and predictive value of the observation based on the result of simulations. K_t is a trade off parameter to control how the simulation model accommodates an actual observation. K_t is called the Kalman gain. When $K_t = 0$, an actual

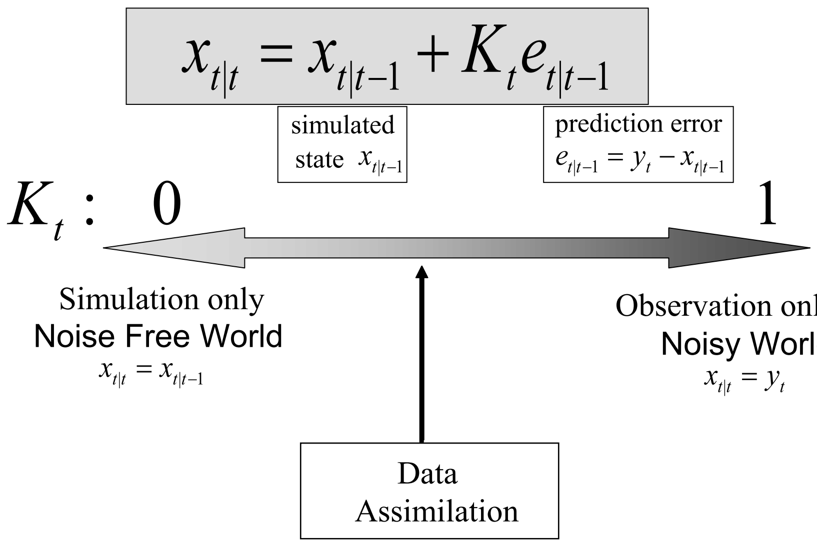

Figure 2: Schematic Representation

observation has no effect on a simulation process. In this case we totally rely on the simulation result. On the other hand, when $K_t = 1$, any discrepancy between the predictive and real values of observations is perfectly adjusted. In this case, it is difficult to identify a dynamics inherent to a simulation model from an estimation of the state vector, because a state vector is highly sensitive to the observation errors.

In general, a filtering procedure estimates $\boldsymbol{x}_t$ from the observed data set $\boldsymbol{y}_{1:t}$ at time t. That is, the procedure estimates the most likely value of the vector $\boldsymbol{x}_t$, or the PDF, as the time series data $\boldsymbol{y}_t$ are observed. If both (2) and (3) are linear equations, the KF is efficient and accurate. However, it is almost impossible to estimate $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ accurately if one of the equations or the noise term is nonlinear.

3 Ensemble based filters

If the DA problem can be formulated in the context of an SSM, it becomes a problem of state vector estimation and then the PF or the EnKF presented is applicable. The PF can deal with nonlinear state space model (2), (3) directly. On the other hand, though the system model in the EnKF can be nonlinear form (2), observation system in the EnKF should be linear equation (4). It should be remarked that nonlinear observation (3) can be dealt with through state vector extension [Evensen (2003)].

3.1 Ensemble approximation

In the EnKF and the PF, the estimated mean and variance which appear in the Kalman filter are replaced by obtaining the set of realizations sampled from PDFs. This realization set is called ensemble and each realization is called ensemble member. Ensemble represents approximation of PDFs:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) \cong \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1}^{(i)}),$$
$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) \cong \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x}_t - \boldsymbol{x}_{t|t}^{(i)}),$$

where $\{\boldsymbol{x}_{t|\cdot}^{(i)}\}_{i=1}^{N}$ is ensemble of $\boldsymbol{x}_{t|\cdot}^{(i)}$, $\boldsymbol{y}_{1:t} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_t\}$ and N is the number of realizations. $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$ is called predictive PDF, $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ is called filtered PDF and corresponding sets are called predictive and filtered ensemble respectively. The filtering problem is how to estimate and update the ensemble set at each time step t using y_t. The filtering procedure consists of two step, the prediction step and the filtering step. These two steps are calculated in turn.

3.2 Prediction and filtering steps

In the prediction step, $\{\boldsymbol{x}_{t|t-1}^{(i)}\}_{i=1}^{N}$ is obtained from $\{\boldsymbol{x}_{t-1|t-1}^{(i)}\}_{i=1}^{N}$ and system model (2). It is common in the EnKF and the PF. The ensemble $\{\boldsymbol{x}_{t|t-1}^{(i)}\}_{i=1}^{N}$ is given by the following Monte Carlo simulation:

$$\boldsymbol{x}_{t|t-1}^{(i)} = f_t(\boldsymbol{x}_{t-1|t-1}^{(i)}, \boldsymbol{v}_t^{(i)}), \quad \boldsymbol{v}_t^{(i)} \sim N(\boldsymbol{0}, Q_t).$$

In the DA, $\boldsymbol{v}_t^{(i)}$ corresponds to undetermined boundary conditions and unmodeled dynamics of simulation model as noted above.

In the filtering step $\{\boldsymbol{x}_{t|t}^{(i)}\}_{i=1}^{N}$ is calculated from $\{\boldsymbol{x}_{t|t-1}^{(i)}\}_{i=1}^{N}$ and y_t. The difference between the EnKF and the PF exists in this step. The filtering step of the EnKF procedure is as follows. At first, sample mean $\hat{\boldsymbol{x}}_{t|t-1}$ and sample covariance matrix $\hat{V}_{t|t-1}$ are calculated:

$$\hat{\boldsymbol{x}}_{t|t-1} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_{t|t-1}^{(i)},$$
$$\hat{V}_{t|t-1} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_{t|t-1}^{(i)} - \hat{\boldsymbol{x}}_{t|t-1})(\boldsymbol{x}_{t|t-1}^{(i)} - \hat{\boldsymbol{x}}_{t|t-1})^{\mathrm{T}},$$

where $\cdot^{\mathrm{T}}$ denotes transpose of matrix. Filtered ensemble $\{\boldsymbol{x}_{t|t}^{(i)}\}_{i=1}^{N}$ is calculated through the update equation of the Kalman filtering

$$K_t = \hat{V}_{t|t-1} H_n^{\mathrm{T}} (H_n \hat{V}_{t|t-1} H_n^{\mathrm{T}} + \hat{R}_t)^{-1},$$
$$\boldsymbol{x}_{t|t}^{(i)} = \boldsymbol{x}_{t|t-1}^{(i)} + K_t(\boldsymbol{y}_t + \boldsymbol{w}_t^{(i)} - H_t \boldsymbol{x}_{t|t-1}^{(i)}),$$

where $\boldsymbol{w}_t^{(i)}$ denotes sample from $N(\mathbf{0}, R_t)$ and $\hat{R}_t$ denotes the sample covariance matrix of $\boldsymbol{w}_t^{(i)}$. It is important to note that this procedure expects Gaussianity of $\boldsymbol{x}_{t|t-1}$ and linearity in the observation model.

The filtering step of the PF is as follows. Weight $q_t^{(i)}$ of each ensemble member is calculated from the observation system and the observation y_t:

$$q_t^{(i)} = p(\boldsymbol{y}_t | \boldsymbol{x}_t = \boldsymbol{x}_{t|t-1}^{(i)}). \tag{5}$$

If the observation system is linear form (4), calculated weight is the following form:

$$q_t^{(i)} = |(2\pi)^{n_t} R_t|^{-\frac{1}{2}} \exp\left(-2(\boldsymbol{y}_t - H_n \boldsymbol{x}_{t|t-1}^{(i)})^{\mathrm{T}} R_t^{-1} (\boldsymbol{y}_t - H_t \boldsymbol{x}_{t|t-1}^{(i)})\right).$$

After this calculation, filtered ensemble $\{\boldsymbol{x}_{t|t}^{(i)}\}_{i=1}^{N}$ is made by sampling with replacement from predictive ensemble $\{\boldsymbol{x}_{t|t-1}^{(i)}\}_{i=1}^{N}$ in proportion to the weight $q_t^{(i)}$.

3.3 Matrix representation of the filters

In the context of the EnKF, the filtering step of the EnKF is written by matrix representation [Evensen (2003)]. All the ensemble members are included in the matrix as column vectors. For example, the ensemble $\{\boldsymbol{x}_{t|t-1}^{(i)}\}_{i=1}^{N}$ is written by

$$X_{t|t-1} = [\boldsymbol{x}_{t|t-1}^{(1)}, \boldsymbol{x}_{t|t-1}^{(2)}, \cdots, \boldsymbol{x}_{t|t-1}^{(N)}].$$

The update rule of the EnKF can be written through this representation. Evensen showed that the update rule can be written by

$$X_{t|t} = X_{t|t-1} Z_{t|t-1},$$

where $Z_{t|t-1}$ is calculated from $\{\boldsymbol{x}_{t|t-1}^{(i)}\}_{i=1}^{N}$, $\{\boldsymbol{w}_t^{(i)}\}_{i=1}^{N}$, H_t and $\boldsymbol{y}_t$ [Evensen (2003)]. It is also shown that the sum of each column of $Z_{t|t-1}$ is one. These things show that the filtered ensemble members are weighted linear combination of the predictive ensemble members.

Once the matrix representation of the ensemble is introduced, the PF can be written by the same way. Resample from $\{\boldsymbol{x}_{t|t-1}^{(i)}\}_{i=1}^{N}$ is written by

$$X_{t|t} = X_{t|t-1} Z_{t|t-1},$$

where each element of $Z_{t|t-1}$ takes the value of one or zero. In this formulation, each column of $Z_{t|t-1}$ consists of one and $N-1$ zeros. Additionally, the number of ones which appear in the ith row of $Z_{t|t-1}$ is proportional to the weight of ith member of prediction ensemble $\boldsymbol{x}_{t|t-1}^{(i)}$.

This representation of filters gives us the consistent view of the EnKF and the PF. The rank of $Z_{t|t-1}$ shows the non-degeneracy of the ensemble, which means decay of variation of the ensemble members. Degeneracy may cause severe estimation bias in the sequential DA because predictive ensemble members which are generated from the same filtered ensemble member may have almost the same value and physical characteristics. Because $Z_{t|t-1}$ of the EnKF is usually full rank, degeneration of the ensemble $\{\boldsymbol{x}_{t|t}^{(i)}\}_{i=1}^{N}$ rarely occurs. In addition, predictive ensemble members usually span the same space as filtered ensemble members do. This property is desirable if nonlinearity of system model (2) is weak from the viewpoint of physical simulation model. However, the filtered ensemble of the EnKF can assure accuracy of moment only up to second order, nonlinearity of SSM may cause estimation errors. On the other hand, $Z_{t|t-1}$ of the PF is rarely full rank because the PF is resample based method and the number of ensemble members are finite. This rank deficiency is cause of degeneration and estimation bias in the PF though the PF can assure accuracy of higher order statistics.

As a consequence of the properties, if the $Z_{t|t-1}$ is efficiently calculated without loss of rank and statistical bias, estimated ensembles will be more efficient and accurate estimation.

3.4 Fixed lag smoother

One of the important and suggestive points of the matrix representation is the case of fixed lag smoothers. The EnKF and the PF have fixed lag smoother, the ensemble Kalman Smoother (EnKS) [Evensen and van Leeuwen (2000)] and the Particle Smoother (PS) [Kitagawa (1996)] respectively. Using matrix representation, we can obtain these smoothers by the same way. The fixed L-lag smoother can be obtained only to replace $X_{t|t-1}$ in the filtering equation (3.3) with L-lag stored matrix:

$$\Xi_{t|t-1} = \begin{pmatrix} \boldsymbol{x}_{t|t-1}^{(1)} & \boldsymbol{x}_{t|t-1}^{(2)} & \cdots & \boldsymbol{x}_{t|t-1}^{(N)} \\ \boldsymbol{x}_{t-1|t-1}^{(1)} & \boldsymbol{x}_{t-1|t-1}^{(2)} & \cdots & \boldsymbol{x}_{t-1|t-1}^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{t-L+1|t-1}^{(1)} & \boldsymbol{x}_{t-L+1|t-1}^{(2)} & \cdots & \boldsymbol{x}_{t-L+1|t-1}^{(N)} \end{pmatrix},$$

where $\boldsymbol{x}_{t'|t^*}^{(i)} (t' < t^*)$ denotes smoothed ensemble members at time t' using $y_{1:t^*}$. Then, we can obtain smoothed ensemble through the equation

$$\Xi_{t|t} = \Xi_{t|t-1} Z_{t|t-1},$$

where $\Xi_{t|t}$ denotes

$$\Xi_{t|t} = \begin{pmatrix} \boldsymbol{x}_{t|t}^{(1)} & \boldsymbol{x}_{t|t}^{(2)} & \cdots & \boldsymbol{x}_{t|t}^{(N)} \\ \boldsymbol{x}_{t-1|t}^{(1)} & \boldsymbol{x}_{t-1|t}^{(2)} & \cdots & \boldsymbol{x}_{t-1|t}^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{t-L+1|t}^{(1)} & \boldsymbol{x}_{t-L+1|t}^{(2)} & \cdots & \boldsymbol{x}_{t-L+1|t}^{(N)} \end{pmatrix}.$$

As we can see, the degeneration problem is more critical for the L-lag PS because $Z_{\cdot|\cdot}$ is multiplied L times after the prediction step whereas the PF is multiplied only once.

4 Data assimilation for tsunami model

Several assimilation methods [Titov et al. (2005)], [Abe (2006)] have been proposed for the application of DA to the study of tsunamis. They concentrate on tsunami source estimation or correction of tsunami height including the run-up height near the shore and the bottom topography is fixed in these studies. However, it is well known that the bottom topography used is erroneous, which generates inaccurate simulation results critical in forecasting tsunami propagation. Therefore, correction of the bottom topography is needed for accurate forecasting. In this section, we explain the formulation scheme of DA for a tsunami model. As described in the following, the PF is used for estimation because the tsunami simulation model has nonlinear part,

4.1 Simulation model

The tsunami simulation model is based on the shallow-water equations model [Choi and Hong (2001)]. This model is a standard model for tsunami simulation studies in geophysics. For the tsunami simulation model, the continuous shallow-water equations are discretized spatially and temporally. The leap-frog scheme is used in the discretization step. Discretization produces a two-dimensional lattice and each point m of the lattice has four physical scalar variables. The variables are depth d_m, sea surface height η_m, which is measured from the average sea surface, and the two components of the two-dimensional water flow vector (u_m, v_m) ([see Fig. 3]). The set of all depth variables d_m represents the bottom topography. There are two types of boundary conditions in the simulation model. Non-reflecting boundary conditions are imposed on the edge points of the lattice. The behavior of the water around the points at the border between land and sea are determined by another set of boundary conditions. The initial conditions are the initial values of d_m, η_m, u_m and v_m. The initial values of η_m are determined from information on land slip, which is estimated from observed data on an

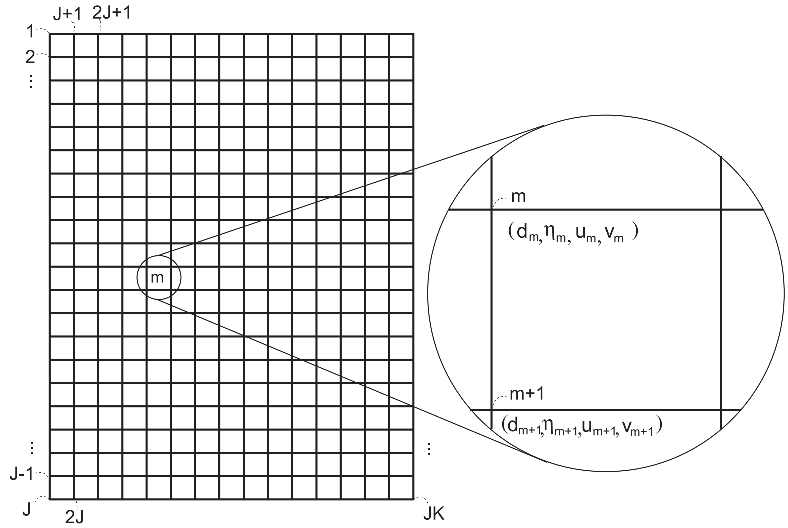

Figure 3: Grid of simulation model and physical variables.

earthquake. The initial values of u_m and v_m are set to zero. The depth d_m of each grid point m is taken from the bottom topography data set; however, this data set is known to be erroneous and is to be re-determined.

4.2 Errors in simulation model

In this paper, the system model is constructed by introducing uncertainty in the bottom topography. There are two ways to introduce this uncertainty; one is to introduce uncertainty into the initial water depth d_m in the form of PDFs, and the other is to add uncertainty as system noise to d_m. It is also possible to introduce both. Which method is appropriate depends on the problem. If the initial conditions are distributed and system noise is not introduced, the fixed parameters can be determined or the model which is suitable for observations can be identified [Nakamura et al. (2005)]. However, applying the PF to such a system model can give rise to degeneration, which causes estimation bias. Hence this system model should be used with care. On the other hand, introducing system noise can give a more robust state estimation with regard to degeneration, though it can exhibit some inconsistency in estimations. Also, it is difficult to justify the system noise in geophysics.

We adopt the former approach in this paper. The uncertainty in the model and in the bottom topography is time invariant over the time scale of tsunami propagation and hence it is more natural to introduce and fix the uncertainty at time step 0 than to introduce the uncertainty as system noise at every time step. The uncertainty in the bottom topography can be regarded as the partial uncertainty in the model and therefore the tsunami DA problem can be regarded as the model identification problem.

Correspondence between each component of [Fig. 1] and tsunami simulation model is as follows. PDE corresponds to shallow-water equations, FDE corresponds to discrete shallow-water equations model which has four variables per grid and initial conditions is uncertainty of bottom topography.

4.3 Data and DA procedure

The data set used is derived from tide gauge records. Each tide gauge station records a one-dimensional time series of sea surface height (SSH) near the installation point. The observation vector $\boldsymbol{y}_t$ consists of measurements at each time step. Hence each component of $\boldsymbol{y}_t$ corresponds to the tide gauge time series of each station. Measurement errors are determined from the observed series.

[Fig. 4] illustrates the progress of the DA at each time step t. In the prediction step, the tsunami propagates by the system model. The SSH is updated and the bottom topography is not changed in this step. In the filtering step, the bottom topography and SSH are modified by the filtering. As a result, the bottom topography and the tsunami height are corrected at every input of the observation.

5 Numerical experiment

5.1 Identical twin experiment

To check the correction ability of the assimilation methods, we conducted a numerical experiment of identification using test data. [Fig. 5] shows the procedure of the identical twin experiment. At first, a simulation is run and the results are recorded. This is called the model-run stage in this paper. Next, the observation data set is constructed using the observation model and simulation results. These simulation results are considered “true”, therefore, the observation data obtained here are regarded as coming from the “true” model. Consequently, we estimate the state vector and parameters from the observation data set by sequential DA, referred to as the assimilation stage. Finally, we check whether these estimated results are identical with the “true” simulation results.

This experimental method for validation is known as the identical twin experiment in the DA field. To use real tide gauge data, an observation matrix H_t or nonlinear observation h_t must be designed.

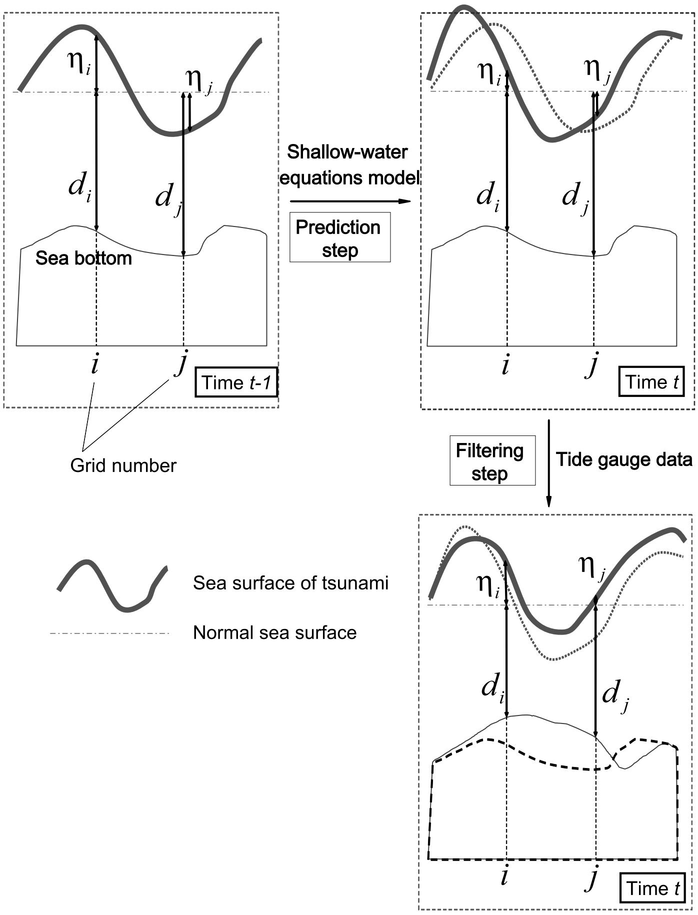

Figure 4: Illustration of each assimilation step.

5.2 Identical twin experiment in tsunami simulation model

We used for the numerical experiment the Okushiri tsunami which occurred in the Japan Sea and killed about 200 people in 1993. This area is discretized longitudinally and latitudinally in a 192(longitude) $\times$ 240(latitude) grid. About half of the grid points are on the sea and the number of the state of each grid point on the sea is four; therefore, the dimension of the state vector is about 9×10^4. The dimension of the observation vector is also four. The four observation points are depicted in [Fig. 6]. The initial conditions of the SSH are determined from data on the earthquake.

In the model-run stage, we fix the bottom topography using a data set and run the Okushiri tsunami simulation. The observation data are then generated

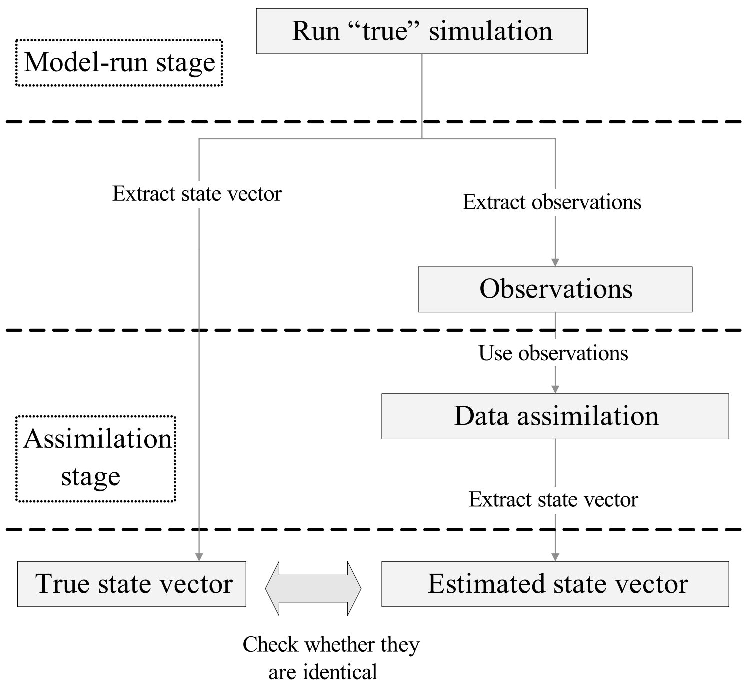

Figure 5: Procedure of identical twin experiment.

using the set of SSH time series at each observation point through the simulation. This procedure specifies the observation model by the following linear model:

$$\boldsymbol{y}_t = H\boldsymbol{x}_t + \boldsymbol{w}_t,$$

where H is a $4 \times n_x$ zero-one matrix and $\boldsymbol{w}_t \equiv 0$. Though this assumes an idealized situation, it is sufficient to show the validity of this method. We discuss the problem of real tide gauge data later in the paper.

In the framework of this study, the most important subject is the bottom topography. Correction of the bottom topography is executed through correction of the water depth at each grid point, which is included in the state vector. As we noted earlier, the bottom topography is time invariant. Therefore, allowing only distribution of the initial conditions is more natural for its parameterization. Hence, we distribute each d_m part of $\boldsymbol{x}_0$ and set $\boldsymbol{v}_t \equiv \mathbf{0}$ in the assimilation stage. To check the effectiveness of the error correction, we set an initial bottom topography estimation biased from that of the model-run. More precisely, d_m is approximated using the ensemble set $\{d_m^{(i)}\}_{i=1}^N$. This ensemble is generated by

$$d_m^{(i)} = c^{(i)}\hat{d}_m,$$

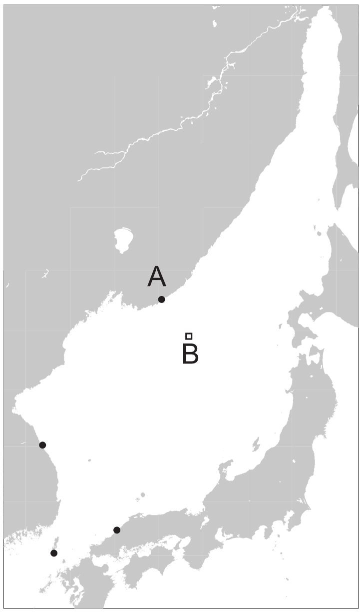

Figure 6: Area of identical twin experiment. The installation points of the tide gauge are shown by dots. The check point for water depth is shown by the square marked B.

where $c^{(i)} \sim N(1.1, 1.5^2)$ and $\hat{d}_m$ is the "true" depth at m. This setting means that the initial estimation depth is deeper than the "true" depth and the degree of bias from the "true" one is uniform regardless of grid point. This is equivalent to generating topography ensemble members whose average is biased from the "true" values and whose distribution is modified at every filtering step. The number of ensemble members is set to 100.

5.3 DA Result

[Fig. 7] shows the result of bottom topography correction. The left side of each image shows the state of the tsunami and the right side shows the estimated bottom topography at that time step. The number above each image shows the number of six-minute intervals passed. The lines drawn in the graphs on the right side of each image show the sea surface height, the bottom topography of the shallowest ensemble member, the "true" bottom topography, the estimated bottom topography and the bottom topography of the deepest ensemble member along the white line of the left side of the image.

The results of an observed SSH at an observation point on the Russian coast (point A in [Fig. 6]) are shown in [Fig. 8]. The estimated and true water depth

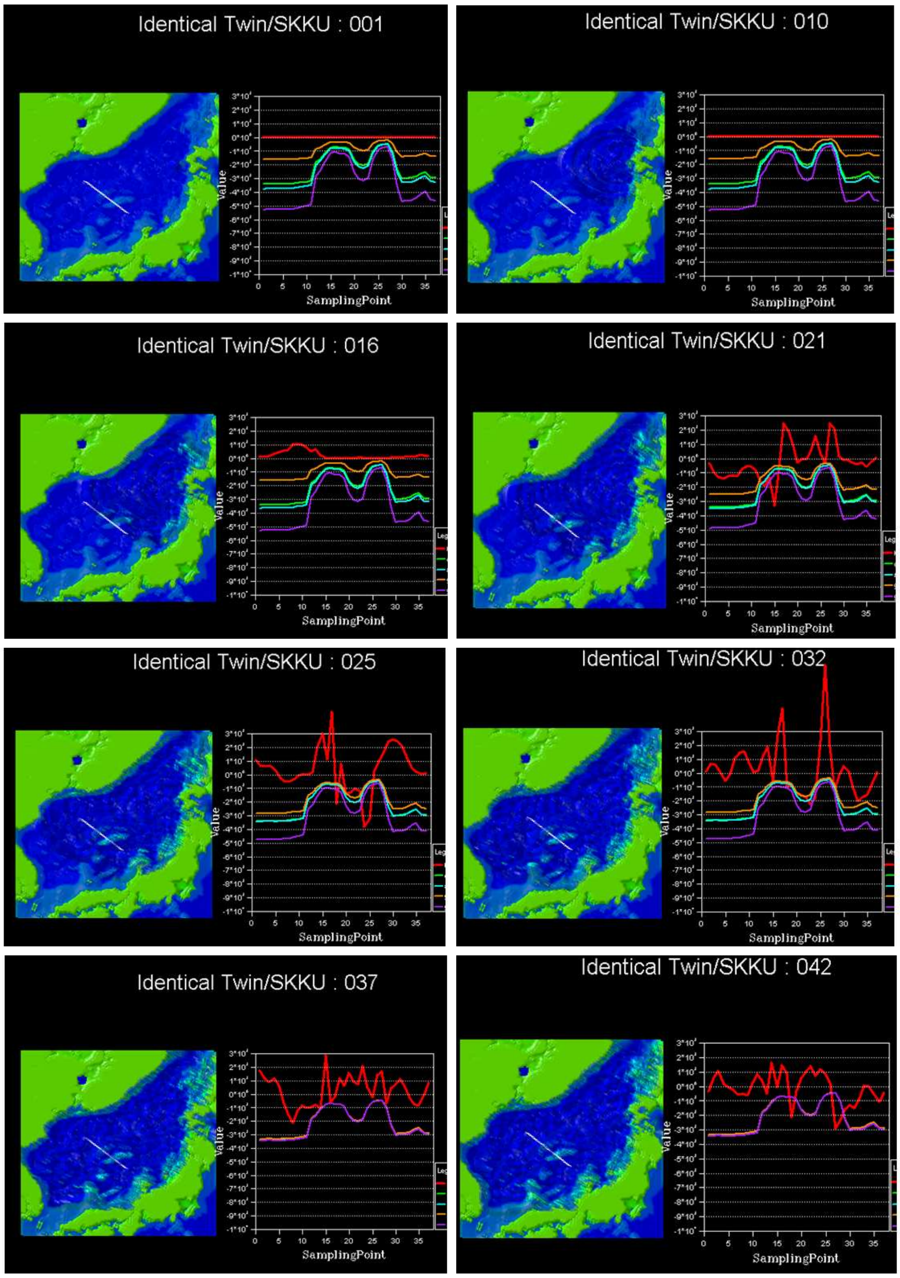

Figure 7: Result of identical twin experiment.

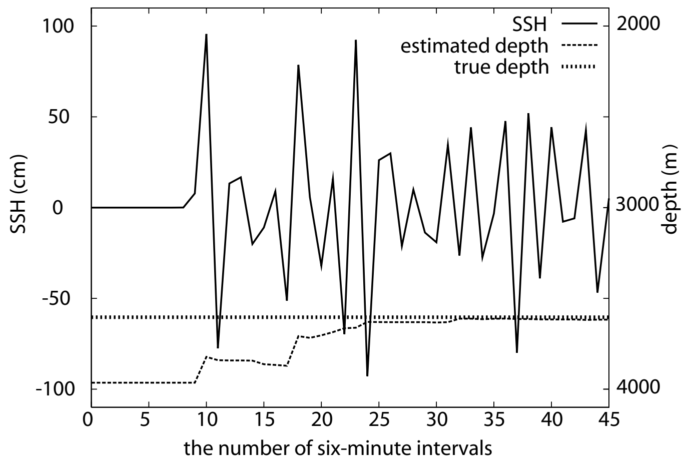

Figure 8: Time series of observed SSH at point A and estimation of water depth at point B in [Fig. 6].

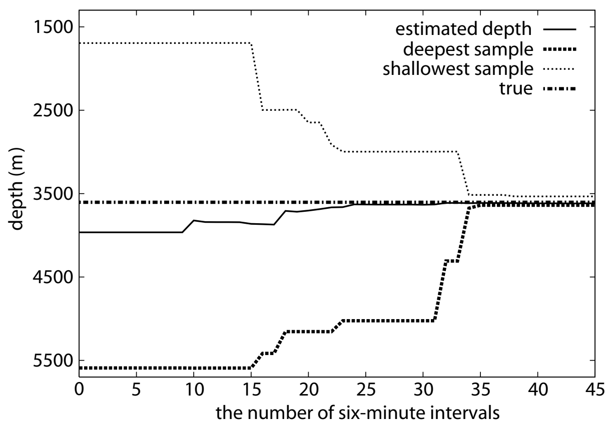

Figure 9: Estimation of water depth at point B in [Fig. 6]. The deepest and shallowest ensemble members of water depths are also shown.

at each time step in the middle of the Japan Sea (point B in [Fig. 6]) are also shown. Point A is the first arrival point of the tsunami amongst the observation points. Immediately after the arrival of the tsunami at point A, the estimated water depth starts to converge to the true water depth.

The time evolution of the deepest and shallowest values of $d_m^{(i)}$ at point B is plotted in [Fig. 9]. The range between the largest and smallest values shrinks and converges to true water depth, indicating that the reliability of the bottom topography estimation improves over time. This result demonstrates that the bottom topography can be effectively corrected by using tide gauge data and hence the method has potential for real data.

6 Conclusion

We have presented the sequential DA framework for a tsunami simulation model and conducted an identical twin experiment. In this framework, the PF is used for estimation. The following two relevant findings are obtained. First, the identical twin experiment shows the potential of the framework for real data. That is to say, sequential DA for a tsunami simulation model and tide gauge data can be used to correct bottom topography and estimate tsunami height. Secondly, the state estimation works well for a small number of ensemble members, in spite of the large dimensionality of the simulation model and the sparseness of the observations. This is allowed by the simple parameterization of uncertainties. Applying this framework for more general conditions, would require a more flexible representation for uncertainties, that is, satisfying both the increase of the degrees of freedom parameter and the avoidance of degeneration problem.

Assimilation experiments for real tide gauge data are currently in progress, analyzing two tsunamis that occurred in the Japan Sea in 1983 and 1993. These tsunamis were selected because reliable tide gauge data set is available for them. The real tide gauge data set is not ideal because the SSH are transformed by many factors, such as the local geography. This problem must be resolved for real DA problems.

References

[Abe (2006)] Abe, K.: "Dominant periods of the 2004 Sumatra tsunami and the estimated source size"; Earth Planets and Space, 58, 2 (2006), 217–221.

[Anderson and Moore (1979)] Anderson, B., Moore, J.: "Optimal Filtering"; Prentice-Hall, New Jersey (1979).

[Choi and Hong (2001)] Choi, B., Hong, S.: "Simulation of prognostic tsunamis on the Korean coast"; Geophysical Research Letters, 28, 10 (2001), 2013–2016.

[Courtier et al. (1994)] Courtier, P., Thepaut, T., Hollingsworth, A.: "A strategy for operational implementation of 4DVAR using an incremental approach"; Quarterly Journal of the Royal Meteorological Society, 120, 519 (1994), 1367–1387.

[Evensen (1994)] Evensen, G.: "Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics"; Journal of Geophysical Research, 99, C5 (1994), 10143–10162.

[Evensen (2003)] Evensen, G.: "The ensemble Kalman filter: Theoretical formulation and practical implementation"; Ocean Dynamics, 53, 4 (2003), 343–367.

[Evensen and van Leeuwen (2000)] Evensen, G., van Leeuwen, P.: "An ensemble Kalman smoother for nonlinear dynamics"; Monthly Weather Review, 128, 6 (2000), 1852–1867.

[Gordon et al. (1993)] Gordon, N. J., Salmond, D. J., Smith, A. F. M.: "Novel approach to nonlinear/non-Gaussian Bayesian state estimation"; IEE Proceedings-F, 140, 2 (1993), 107–113.

[Higuchi (2001)] Higuchi, T.: "Self-organizing time series model"; Sequential Monte Carlo methods in practice, Doucet, A., et al. ed., Springer, New York (2001), 429–44.

[Kalman (1960)] Kalman, R. E.: "A new approach to linear filtering and prediction problems"; Journal of Basic Engineering, 82, (1960), 35–45.

[Kitagawa (1996)] Kitagawa, G.: "Monte Carlo filter and smoother for non-Gaussian nonlinear state space model"; Journal of Computational and Graphical Statistics, 5, 1 (1996), 1–25.

[Kitagawa and Gersch (1996)] Kitagawa, G., Gersch, W.: "Smoothness Priors Analysis of Time Series"; Springer-Verlag, New York (1996).

[Lorenz (1963)] Lorenz, E. N.: "Deterministic nonperiodic flow"; Journal of Atmospheric Science, 20, 2 (1963), 130–141.

[Manda et al. (2003)] Manda, A., Hirose, N., Yanagi, T.: "Application of a nonlinear and non-Gaussian sequential estimation method for an ocean mixed layer model"; Engineering Sciences Reports, Kyushu University, 25, (2003), 285–289.

[Nakamura et al. (2005)] Nakamura, K., Ueno, G., Higuchi, T.: "Data assimilation : Concept and algorithm"; Proceedings of the Institute of Statistical Mathematics, 53, 2 (2005), 201–219, (in Japanese with English abstract).

[Titov et al. (2005)] Titov, V., González, F., Bernard, E., Eble, M., Mofjeld, H. O., Newman, J., Venturato, A.: "Real-time tsunami forecasting: Challenges and solutions"; Natural Hazards, 35, 1 (2005), 41–58.

[van Leeuwen (2003)] van Leeuwen, P.: "A variance-minimizing filter for large-scale applications"; Monthly Weather Review, 131, 9 (2003), 2071–2084.

[Wunsch (1996)] Wunsch, C.: "The Ocean Circulation Inverse Problem"; Cambridge University Press, Cambridge (1996).