Journal of Marine Systems 90 (2012) 13–22

Contents lists available at SciVerse ScienceDirect

Journal of Marine Systems

journal homepage: www.elsevier.com/locate/jmarsys

Sequential data assimilation in fine-resolution models using error-subspace emulators: Theory and preliminary evaluation

N. Margvelashvili [a,*], E.P. Campbell [b]

[a] CSIRO Marine and Atmospheric Research, GPO Box 1538, Hobart, Tasmania 7001, Australia
[b] CSIRO Mathematics, Informatics and Statistics, Private Bag 5, PO Wembley, WA 6913, Australia

ARTICLE INFO

Article history:
Received 20 January 2011
Received in revised form 20 June 2011
Accepted 12 August 2011
Available online 22 August 2011

Keywords:
Data assimilation
Particle Filter
Gaussian Process Modelling
Emulator
Singular Value Decomposition

ABSTRACT

A novel technique for nonlinear sequential data assimilation in computationally expensive fine-resolution models is introduced. The technique involves basis function approximation for dimension reduction and Gaussian Process Modelling for simulation speedup. The basis function approximation is carried out via the Singular Value Decomposition (SVD) of the model ensemble. The Gaussian Process Models propagate the model solution in the error-subspace defined by a finite set of basis functions. The developed technique can also be considered approximate Particle Filtering with two classes of particles: model-particles representing an ensemble of computationally expensive model solutions, and emulator-particles representing an ensemble of fast and cheap model approximations. The algorithm was tested by assimilating synthetic data into a two-dimensional (one spatial dimension plus time) sediment transport model in an idealised vertically-resolved benthic–pelagic system. The assimilation algorithm updates 2 spatially varying state variables and 3 unknown parameters. Numerical experiments illustrate robust performance of the technique for a wide range of the assimilation settings. The capabilities and limitations of the approach are discussed, and further developments are outlined.

1. Introduction

Uncertainty of complex environmental models is often high or even unknown. Data assimilation techniques are employed to improve these models and reduce their uncertainty. Integral to statistical data assimilation is the evaluation of the quality of such improved model simulations. In the context of sequential assimilation it is essential also to propagate the error statistics in time.

For linear problems a Kalman Filter (KF) provides a variance minimising solution (Jazwinski, 1970). The Extended KF (EKF) designed for nonlinear problems can only handle weakly nonlinear models because the assumption is made that the error statistics evolves according to a tangent linear model. The Ensemble Kalman Filter (EnKF) is a popular Monte Carlo that uses stochastic dynamic prediction, but relies on EKF theory for data assimilation and thus assumes a Gaussian distribution for the error statistics (Evensen, 2003, 2009; Tippet et al., 2003). Particle Filters (PF—Doucet et al., 2001) don't make specific distributional assumptions, and so provide a more general assimilation scheme. PF has become well established in the statistical and signal processing literature with a number of pilot applications in physical oceanography and biogeochemistry (Dowd, 2006, 2007; Jones et al., 2009; Loza et al., 2003; van Leeuwen, 2003; Zhou et al., 2006).

The theory underlying PF is well understood, but its practical implementation is moderated by a number of issues (Bengtsson et al., 2008; Berliner and Wikle, 2007). One of the key problems, known as particle degeneracy, is a loss of diversity amongst the particle ensemble. The problem is particularly acute in high-dimensional systems because the number of particles required grows exponentially with dimension. In oceanographic applications this problem is further exacerbated by typically high computational expenses of forward modelling which limit the ensemble size to a few tens or hundreds of particles. Given the high dimensionality of ocean models (~10e5–10e7), the challenge is to track the evolution of high-dimensional probability density functions (PDFs) with a limited number of particles (~100).

This paper describes a novel technique for data assimilation in computationally expensive fine-resolution models. It employs Singular Value Decomposition (SVD) for dimension reduction and model emulation for simulation speedup. The sampling strategy is based on Particle Filtering. The algorithm is tested by assimilating synthetic data into a 2-d (one spatial dimension plus time) model of sediment transport in a coupled vertically-resolved benthic-pelagic system.

2. Methodology

2.1. Background theory

Consider a physical system represented in discrete form by its state vector $\tilde{x}$ of dimension $N_{\tilde{x}}$. Augment the vector $\tilde{x}$ with the vector

* Corresponding author. Tel.: +61 3 62325142; fax: +61 3 62325123.
E-mail address: Nugzar.margvelashvili@csiro.au (N. Margvelashvili).

0924-7963/$ – see front matter.
doi:10.1016/j.jmarsys.2011.08.004

of model parameters $\boldsymbol{\theta}$ (of dimension N_θ) and define the extended state vector $x = (\tilde{x}, \theta)$ of dimension $N_x = N_{\tilde{x}} + N_\theta$. Assume that x evolves according to the following model:

$$\mathbf{x}(t_i) \equiv \mathbf{x}_i = M_{i-1}[\mathbf{x}(t_{i-1}), \eta(t_{i-1})] \tag{1}$$

where M is the system of the model governing equations, and $\eta_i \equiv \eta(t_i)$ is a zero-mean, white noise sequence independent of past and current states. At time t_i, observations are available as a vector $y^o(t_i)$ of dimension N_y. The true state x^t at time t_i is assumed to be related to the observation vector via the observation equation

$$\mathbf{y}^o(t_i) \equiv \mathbf{y}_i^o = H_i\left[\mathbf{x}^t(t_i), \varepsilon(t_i)\right] \tag{2}$$

where $H_i(\cdot,\cdot)$ is the measurement operator and $\varepsilon_i \equiv \varepsilon(t_i)$ is another zero-mean, white-noise sequence of known distribution.

It is assumed that at time t_i a set of measurements $Y_i^o = \{y_j^o; j = 1,\ldots, i\}$ is available and $p(x_{i-1}|Y_{i-1}^o)$ is known. The requirement is to construct the PDF of the current state x_i, given all the available data: $p(x_i|Y_i^o)$, the analysis distribution.

The posterior distribution $p(x_i|Y_i^o)$ can be obtained in two stages: forecast and analysis (Gordon et al., 1993; Wikle and Berliner, 2007). During the forecast step, we obtain the prior PDF of the state at time t_i by propagating $p(x_{i-1}|Y_{i-1}^o)$ forward in time:

$$p(x_i|Y_{i-1}^o) = \int p(x_i|x_{i-1}) p(x_{i-1}|Y_{i-1}^o) dx_{i-1}. \tag{3}$$

During the analysis step, a measurement y_i^o is used to update the prior via Bayes rule

$$p(x_i|Y_i^o) = \frac{p(y_i^o|x_i) p(x_i|Y_{i-1}^o)}{p(y_i^o|Y_{i-1}^o)} \tag{4}$$

where the normalising denominator is given by

$$p\left(y_i^o|Y_{i-1}^o\right) = \int p(y_i^o|x_i) p(x_i|Y_{i-1}^o) dx_i. \tag{5}$$

Under the assumption that the model and measurement functions are linear and η_i, $\boldsymbol{\varepsilon}_i$ are additive Gaussian variables, the analytical solution to Eqs. (3)–(5) gives the Kalman Filter. The first order extension of the KF to nonlinear models (EKF) is obtained by linearising the model and measurement operators Eqs. (1) and (2) around the most recent state estimate. Monte Carlo approximation to the mean and variance of the posterior distribution Eq. (4) gives EnKF. The analysis step of all these techniques relies on KF theory and assumes Gaussian distribution of the forecast error statistics.

Particle Filter provides an alternative, fully non-linear data assimilation method, which relies on the assumption that the particle ensemble is an appropriate representation of a large-scale nonlinear system. The key idea is to approximate the required PDF by a finite number of model realisations (particles), with discrete weights assigned to each particle, and propagate the particles forward in time. The forecast step is based on the model Eq. (1). The analysis step draws samples from the posterior Eq. (4) where one may think of $p(x_i|Y_{i-1}^o)$ as a "prior" density, which is combined with the likelihood $p(y_i^o|x_i)$. A number of statistical techniques are available to draw samples from $p(x_i|Y_i^o)$. A comprehensive review can be found in (Andrieu et al., 2003; Arulampalam et al., 2002; Doucet et al., 2001). Here we outline the Metropolis–Hastings (MH) algorithm, which is a Markov Chain Monte Carlo (MCMC) sampling technique (Chen, 2003; Gelman et al., 2004) underlying our sampling strategy. We abbreviate this as MH–MCMC.

The idea of MCMC is to structure a random walk through the space we want to sample from, with the probability density governing the walk structured in such a way that the limiting distribution of the sampled points is the distribution of interest. In our case we would generate a sample of size n $\{x_i^j : j = 1,\ldots,n\}$ from $p(x_i|Y_i^o)$. The MH–MCMC algorithm proceeds as follows:

1. Draw a proposal from the density $q(x_i^{j+1}|x_i^j)$;
2. Accept this proposal as the next state in the random walk with probability

$$p(x_i^{j+1}|x_i^j) = \min\left[\frac{p(x_i^{j+1}|Y_i^o)/q\left(x_i^{j+1}\middle|x_i^j\right)}{p\left(x_i^j|Y_i^o\right)/q\left(x_i^j\middle|x_i^{j+1}\right)}, \quad 1\right] \tag{6}$$

$$= \min\left[\frac{p\left(y_i^o\middle|x_i^{j+1}\right) p\left(x_i^{j+1}|Y_{i-1}^o\right) q\left(x_i^j\middle|x_i^{j+1}\right)}{p\left(y_i^o\middle|x_i^j\right) p\left(x_i^j|Y_{i-1}^o\right) q\left(x_i^{j+1}\middle|x_i^j\right)}, \quad 1\right],$$

3. If the move is rejected, remain in the same state, so set $x_i^{j+1} = x_i^j$.

The logic for the algorithm is clear from the form of the acceptance probability in Eq. (6). The algorithm is driven into areas of high probability, with a correction for asymmetry of the proposal density. Note that the ratio of proposal density term cancels if the proposal is symmetric. A very readable introduction to MCMC methods is provided by Smith and Roberts (1993).

After the burn-in period, samples of x_i^j converge to the posterior distribution Eq. (4). In practice, typically only samples of the "prior" $p(x_i|Y_{i-1}^o)$ are available but the probability density itself is not known which makes evaluation of Eq. (6) not trivial. However, taking the proposal distribution equal to the "prior"

$$q(x_i|x_i^*) = p(x_i|Y_{i-1}^o) \tag{7}$$

reduces Eq. (6) to the ratio of the likelihoods

$$p(x_i^{j+1}|x_i^j) = \min\left[\frac{p\left(y_i^o\middle|x_i^{j+1}\right)}{p\left(y_i^o\middle|x_i^j\right)}, \quad 1\right] \tag{8}$$

which is relatively easy to evaluate.

Since the proposal density now is taken equal to the "prior", in what follows, unless otherwise specified, we refer to $p(x_i|Y_{i-1}^o)$ as a proposal density.

According to Eq. (7), during MCMC sampling a random sample must be drawn from the proposal density $p(x_i|Y_{i-1}^o)$ represented by a discrete set of particles. In high-dimensional systems an insufficiently large set of these particles may give a poor approximation of the continuous distribution, which in turn may lead to degeneracy of the ensemble. In practice, the number of particles one can afford to propagate forward in time is often limited by computational expenses of forward modelling. In the case of complex ocean models this number can be as low as a few tens.

In the following sections we develop fast and cheap approximation of the complex model called an emulator. An emulator is typically orders of magnitude faster than the model, and allows one for quick evaluation of many thousands of approximate model trajectories. Instead of a few tens of discrete samples, continuous distributions may in principle be approximated by many thousands of emulator particles.

In what follows, the model is considered deterministic (i.e., for a given set of parameters and initial conditions, it always predicts the same solution). Observations and emulators are stochastic, and emulators are used as a model substitute during the MCMC sampling. The stochasticity of the emulator is only due to errors of approximating complex deterministic model.