Bayesian inference of dynamics from partial and noisy observations using data assimilation and machine learning

Marc Bocquet¹, Julien Brajard²,³, Alberto Carrassi⁴,⁵, and Laurent Bertino²

¹CEREA joint laboratory École des Ponts ParisTech and EdF R&D, Université Paris-Est, France
²NERSC, data assimilation, Bergen, Norway (julien.brajard@nersc.no)
³Sorbonne University, CNRS-IRD-MNHN, LOCEAN, Paris, France
⁴Dept. of Meteorology, University of Reading and National Centre for Earth Observation, United Kingdom
⁵Mathematical Institute, University of Utrecht, The Netherlands

The reconstruction from observations of the dynamics of high-dimensional chaotic models such as geophysical fluids is hampered by (i) the inevitably partial and noisy observations that can realistically be obtained, (ii) the need and difficulty to learn from long time series of data, and (iii) the unstable nature of the dynamics. To achieve such inference from the observations over long time series, it has recently been suggested to combine data assimilation and machine learning in several ways. We first rigorously show how to unify these approaches from a Bayesian perspective, yielding a non-trivial loss function.

Existing techniques to optimize the loss function (or simplified variants thereof) are re-interpreted here as coordinate descent schemes. The expectation-maximization (EM) method is used to estimate jointly the most likely model and model error statistics. The main algorithm alternates two steps: first, a posterior ensemble is derived using a traditional data assimilation step using an ensemble Kalman smoother (EnKS); second, both the surrogate model and the model error are updated using machine learning tools, a quasi-Newton optimizer, and analytical formula. In our case, the spatially extended surrogate model is formalized as a neural network with convolutional layers leveraging on the locality of the dynamics.

This scheme has been successfully tested on two low-order chaotic models with distinct identifiability, namely the 40-variable and the two-scale Lorenz models. Additionally, an approximate algorithm is tested to mitigate the numerical cost, yielding similar performances. Using indicators that probe short-term and asymptotic properties of the surrogate model, we investigate the sensitivity of the inference to the length of the training window, to the observation error magnitude, to the density of the monitoring network, and to the lag of the EnKS. In these iterative schemes, model error statistics are automatically adjusted to the improvement of the surrogate model dynamics. The outcome of the minimization is not only a deterministic surrogate model but also its associated stochastic correction, representative of the uncertainty attached to the deterministic part and which accounts for residual model errors.