Physica D 360 (2017) 36–45

Contents lists available at ScienceDirect

Physica D

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

A coherent structure approach for parameter estimation in Lagrangian Data Assimilation

John Maclean [a,*], Naratip Santitissadeekorn [b], Christopher K.R.T. Jones [a]

[a] Department of Mathematics and RENCI, University of North Carolina at Chapel Hill, United States
[b] Department of Mathematics, University of Surrey, Guildford, United Kingdom

HIGHLIGHTS

- Coherent patterns can be used to form effective data assimilation schemes.
- A Pattern-based distance is used in a likelihood-free data assimilation.
- The pattern-based scheme is unaffected by chaotic advection.

ARTICLE INFO

Article history:
Received 14 September 2016
Received in revised form 29 May 2017
Accepted 20 August 2017
Available online 5 September 2017
Communicated by H.A. Dijkstra

Keywords:
Data assimilation
Lagrangian data
Coherent structures

ABSTRACT

We introduce a data assimilation method to estimate model parameters with observations of passive tracers by directly assimilating Lagrangian Coherent Structures. Our approach differs from the usual Lagrangian Data Assimilation approach, where parameters are estimated based on tracer trajectories. We employ the Approximate Bayesian Computation (ABC) framework to avoid computing the likelihood function of the coherent structure, which is usually unavailable. We solve the ABC by a Sequential Monte Carlo (SMC) method, and use Principal Component Analysis (PCA) to identify the coherent patterns from tracer trajectory data. Our new method shows remarkably improved results compared to the bootstrap particle filter when the physical model exhibits chaotic advection.

1. Introduction

The problem of assimilating the path travelled by ocean instruments such as drifters to estimate states and/or parameters of dynamical systems is commonly known as Lagrangian data assimilation (LaDA). Every method for LaDA depends on the knowledge of error statistics for the observed position of the drifter paths; typically, it assumes the normal likelihood function and uncorrelated errors of observed position among all drifters. In this paper, we explore a novel idea for LaDA that assimilates a "coherent (or persistent) structure" hidden in the Lagrangian path of the drifters instead of directly using the observed positions. We discuss the main advantage of this idea in the situation where the number of drifters is large, observed position has small variance and the flow is chaotic. We also demonstrate the improved accuracy of the parameter estimates of the new method when comparing with the particle filtering (PF) approach, which is apparent if the drifter paths are dominated by chaotic advection.

Modern satellite-tracked ocean drifters [1,2] have been recognized as an important source of data for oceanographic and climate research. Drifters travel near the surface and measure physical data such as surface temperature along their paths via drifter sensors [3]. In addition, they offer Lagrangian path data through the Doppler frequency shift on their satellite-based transmission. The accuracy of drifter locations can be less than 150 m [2], and there is a global ocean drifter set providing measurement data at a mean time interval of 1.2 h spanning over a decade [1].

Oceanographic flows exhibit patterns at scales from metres to kilometres, and correspondingly drifters are seen to move in eddies or gyres and follow currents, creating small and large-scale patterns which can be used to infer the sea-surface flows [4,5]. These patterns are associated with fundamental structures of the flow, including stable and unstable directions, centres and saddles. The goal of the current work is to infer the underlying flow state, or the parameter set governing the flow, *from the patterns* rather than directly from the drifter trajectories.

Several LaDA methods have been presented in the last decade for sequentially assimilating Lagrangian path data into dynamical system models of the ocean [6–14]. These methods append a Lagrangian model of the drifter locations to the Eulerian flow model

* Corresponding author.
E-mail address: jmaclean@ad.unc.edu (J. Maclean).

http://dx.doi.org/10.1016/j.physd.2017.08.007
0167-2789/

in order to make inferences about how drifter locations correlate with model states; this approach also avoids the difficulty in which Lagrangian path data does not correspond to the fixed Eulerian grid locations. The key challenge of Lagrangian Data Assimilation (LaDA) is that the Lagrangian path of the passive tracers or drifters can be strongly nonlinear, which makes it difficult to sequentially estimate model state variables or parameters by computationally efficient methods such as the Ensemble Kalman filter (EnKF) [10]. It has been emphasized in several pieces of work that a nonlinear, Bayesian filtering framework such as the Particle Filter (PF) is more reliable in quantifying the uncertainty of the estimates [9–11,13,15]. However, when the drifter locations are observed at a high accuracy (e.g. the class-three Argos), the observation is considered to be highly informative, corresponding to a likelihood probability with a small variance in the Bayesian framework. Assimilating highly-informative observations in a high-dimensional space is problematic since significant probability is contained in a region of extremely small "volume". The standard PF tends to perform very poorly in such a regime [16]. With a narrow prior, it is probable that no particles will lie in a region of significant probability after a short amount of time. In this situation the PF weights, which fall off exponentially, assign probability one to a single particle and probability 0 to the rest. Therefore, it would be difficult to apply PF or its variants to high-dimensional models of the ocean, or assimilate data from a large number of drifters, such as the approximately 2200 ocean drifters from the Global Ocean drifter Program (www.aoml.noaa.gov/envids/gld/). The problem of drawing inference from drifter trajectories is compounded by factors which tend to separate the numerically simulated drifters from the true trajectories. One such factor is model error, which is not considered in this paper. A second factor of particular importance to the current work is chaotic advection in the flow, which can readily create a situation where the simulated and observed drifter trajectories will appear unrelated, particularly if the observations are sparse.

A hybrid PF-EnKF method has been proposed to reduce the computational cost in high-dimensional problems [17,18]; in this approach it is assumed that the Eulerian flow model is high-dimensional and the Lagrangian model for the drifter locations is low-dimensional. A similar idea using a hybrid PF-EnKF to assimilate the Lagrangian path data has also been developed in [13,14] where EnKF is used to estimate drifter paths in tandem with PF for parameter estimation. However, to the best of our knowledge, the challenge of assimilating a high-dimensional drifter data set is still an open problem.

These issues motivate us to exploit the qualitative structure of complex nonlinear flows – such as coherent structures or persistent patterns, if they exist – to make inferences on the model parameters.

The idea of exploiting such structures is related at a high level of abstraction to other efforts to exploit the structure of nonlinear flows. We mention specifically the earlier work [19], that uses 'bred vectors' to construct the model ensemble. The goal of the prior work is to create an ensemble that correctly represents the extent and directions of model uncertainty; data assimilation then proceeds according to standard methods. By contrast in this work it is assumed that some informative structure is already present, but hidden, in the model forecast and observations. The goal is to design a data assimilation scheme that explicitly exposes and assimilates the structures in both model and observations.

In particular, we will exploit a large-scale "Lagrangian coherent pattern" or just "coherent pattern" hidden in the Lagrangian path data. The coherent pattern is imprecisely defined as a region in state space, for example a coherent vortex or nonlinear jet, that moves along with the flow without dispersing. Coherent patterns must also be robust under small diffusive perturbations, that is, the coherent object should still hold its geometric structure together under some diffusion process. When the flow is known either numerically or in a closed form, these coherent regions can be identified by several approaches. In the probabilistic approach, the almost invariant-set framework was developed for autonomous and periodic flows by using the transfer (or Perron–Frobenius) operator [20–23] or the infinitesimal generator [24]; the latter reduces the high computational cost of the former approach. For nonautonomous flows, the finite-time coherent set method via a transfer operator was introduced for the first time in [25–27] and it was applied to detect long-lived vortices such as the stratospheric polar vortex [26,27] and Agulhas rings [28], both as two and three dimensional coherent objects. This method has a strong connection with spectral clustering as described in Chapter of 4 of [29]. The coherent region as used in these works probabilistically minimizes the mass transport in and out of the region with respect to a reference measure (not necessarily the invariant measure), see [26,30] for full details.

Unfortunately, all of the above coherent set identification methods require a set of governing equations of the underlying dynamics and cannot be applied to identify the coherent pattern directly from the Lagrangian path data, which is an essential goal of using LaDA in the current work. Recently, new data-driven algorithms have been developed that can extract coherent patterns from possibly sparse Lagrangian path data without having to rely on governing equations. To name a few, these methods include the diffusion-map algorithm [31], fuzzy c-mean algorithm [32], bipartite spectral clustering [33] and dynamic mode decomposition (DMD) [34,35]. These methods can capture the time-dependent coherent pattern that represents a slowly-decaying mode of a complex flow field.

It is not a primary concern of this paper to select or consider the ideal algorithms to identify the spatio-temporal coherent patterns in a general context, but to advocate for a methodology in which identified coherent patterns are assimilated in a Bayesian framework. For this reason, numerical examples in this paper will be limited to the case of a stationary spatial pattern where an application of standard principal component analysis (PCA) is satisfactory for the coherent pattern identification. We will provide a brief review of PCA in Section 2.

In Bayesian data assimilation, a closed form of the likelihood function for the coherent pattern must be known. However, even if the likelihood of tracer positions is commonly known, it is still a difficult task to derive the likelihood function of any corresponding coherent patterns. We will address this issue by applying the so-called "Approximate" Bayesian computation (ABC) [36], which can be thought of as a "likelihood-free" Monte Carlo method. Originally, the ABC method was proposed to avoid evaluating a likelihood function that is computationally expensive by constructing a "distance function" together with a summary statistic in a rejection–acceptance algorithm. In our context, this summary statistic is the coherent pattern. It is typically impossible to construct the required distance function without knowledge of the likelihood function of the coherent pattern; we will argue that given the likelihood function is unknown, a distance function with minima in the most likely regions of parameter values is a useful surrogate. We will discuss our choice of the distance function in Section 3 and ABC methods in Section 5.

2. Coherent spatial patterns via PCA

As we will make inference about model parameters based on the coherent pattern of the flow, we use the dominant eigenvector (corresponding to the eigenvalue with the largest magnitude) of the covariance matrix as a representative of the coherent pattern, as conventionally performed by the Principal Component