LSE Research Online

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Hailiang Du and Leonard A. Smith

Pseudo-orbit data assimilation. Part I: the perfect model scenario

Article (Published version) (Refereed)

Original citation:
Du, Hailiang and Smith, Leonard A. (2014) *Pseudo-orbit data assimilation. part I: the perfect model scenario.* Journal of the Atmospheric Sciences, 71 (2). pp. 469-482. ISSN 0022-4928

DOI: 10.1175/JAS-D-13-032.1

© 2014 American Meteorological Society

This version available at: http://eprints.lse.ac.uk/55849/

Available in LSE Research Online: July 2014

LSE has developed LSE Research Online so that users may access research output of the School. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LSE Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain. You may freely distribute the URL (http://eprints.lse.ac.uk) of the LSE Research Online website.

http://eprints.lse.ac.uk

Pseudo-Orbit Data Assimilation. Part I: The Perfect Model Scenario

HAILIANG DU AND LEONARD A. SMITH

Centre for the Analysis of Time Series, London School of Economics and Political Science, London, United Kingdom

(Manuscript received 26 January 2013, in final form 21 October 2013)

ABSTRACT

State estimation lies at the heart of many meteorological tasks. Pseudo-orbit-based data assimilation provides an attractive alternative approach to data assimilation in nonlinear systems such as weather forecasting models. In the perfect model scenario, noisy observations prevent a precise estimate of the current state. In this setting, ensemble Kalman filter approaches are hampered by their foundational assumptions of dynamical linearity, while variational approaches may fail in practice owing to local minima in their cost function. The pseudo-orbit data assimilation approach improves state estimation by enhancing the balance between the information derived from the dynamic equations and that derived from the observations. The potential use of this approach for numerical weather prediction is explored in the perfect model scenario within two deterministic chaotic systems: the two-dimensional Ikeda map and 18-dimensional Lorenz96 flow. Empirical results demonstrate improved performance over that of the two most common traditional approaches of data assimilation (ensemble Kalman filter and four-dimensional variational assimilation).

1. Introduction

The quality of forecasts from dynamical nonlinear models depends both on the model and on the quality of the initial conditions. Even under the idealized conditions of a perfect model of a deterministic chaotic nonlinear system and with infinite past observations, uncertainty in the observations can make identification of the exact state impossible (Berliner 1991; Lalley 2000; Judd and Smith 2001). Such limitations make a single ''best guess'' prediction a suboptimal approach to state estimation—an approach that would be frustrated even in ideal cases. Alternatively, an ensemble of initial conditions can better reflect the inescapable uncertainty in the observations by capturing the sensitivity of each particular forecast. A major application of state estimation is in the production of analyses of the state of the atmosphere in order to initialize a numerical weather prediction (NWP) model (e.g., Lorenc 1986; Keppenne and Rienecker 2002; Kalnay 2003; Anderson et al. 2005; Houtekamer et al. 2005). This paper is concerned with the identification of the current state of a nonlinear chaotic system given a sequence of observations in the perfect model scenario (PMS). The use of imperfect models is discussed elsewhere (Du and Smith 2014, hereafter Part II).

In PMS, there are states that are consistent with the model's long-term dynamics and states that are not; in dissipative systems the consistent states are often said to ''lie on the model's attractor.'' Intuitively, it makes sense to distinguish those states that are not only consistent with the observations but also consistent with the model's long-term dynamics in state estimation. The problem of state estimation in PMS is addressed by applying the pseudo-orbit data assimilation (PDA) approach (Judd and Smith 2001; Judd et al. 2008; Stemler and Judd 2009) to locate a reference trajectory (a model trajectory consistent with the observation sequence; Gilmour 1998; Smith et al. 2010) and constructing an initial condition ensemble by using the model dynamics to sample the local state space. The PDA approach is shown to be more efficient[1] and robust in finding a reference trajectory than four-dimensional variational assimilation (4DVAR). The differences between PDA and 4DVAR are discussed. Ensemble Kalman filter (EnKF) approaches provide an

Corresponding author address: Hailiang Du, Centre for the Analysis of Time Series, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom. E-mail: h.l.du@lse.ac.uk

[1] In high-dimensional space, sampling on or near the relevant low-dimensional manifold would be much more efficient than sampling a sphere in the entire space.

DOI: 10.1175/JAS-D-13-032.1

© 2014 American Meteorological Society

alternative to 4DVAR. Illustrated here both on a low-dimensional map and on a higher-dimensional flow, PDA is demonstrated to outperform one of the many EnKF approaches—that is, the ensemble adjustment Kalman filter (Anderson 2001, 2003). It is suggested that this is a general result.

The data assimilation problem of interest is defined and alternative approaches are reviewed in section 2. A full description of the methodology of the PDA approach is presented in section 3. In section 4, the 4DVAR approach and the PDA approach are contrasted using the two-dimensional Ikeda map and the 18-dimensional Lorenz96 flow. Comparisons between the EnKF approach and the PDA approach for the same two systems are made in section 5. Section 6 provides a brief summary and conclusions.

2. Problem description

The problem of state estimation is addressed within the perfect model scenario focusing only on nonlinear deterministic dynamical systems.[2] Let $\tilde{\mathbf{x}}_t \in \mathbb{R}^{\tilde{m}}$ be the state[3] of a deterministic dynamical system at time $t \in \mathbb{Z}$. The evolution of the system is given by $F(\tilde{\mathbf{x}}_t, \tilde{\mathbf{a}}): \mathbb{R}^m \to \mathbb{R}^m$ with $\tilde{\mathbf{x}}_{t+1} = F(\tilde{\mathbf{x}}_t, \tilde{\mathbf{a}})$, where F denotes the system dynamics that evolve the state forward in time in the system state space $\mathbb{R}^m$ and the system's parameters are contained in the vector $\tilde{\mathbf{a}} \in \mathbb{R}^l$. In PMS, assume the mathematical structure of F is known while uncertainty in the values of $\tilde{\mathbf{a}}$ may remain. Note that PMS does not require knowing the parameters of the system. Robust approaches have been proposed to identify the parameter values (e.g., Schittkowski 1994; McSharry and Smith 1999; Pisarenko and Sornette 2004; Tarantola 2004; Smith et al. 2010; Du and Smith 2012). This paper, however, considers only the strong case where both the model class (i.e., the mathematical structure of F) and the model parameter values $\tilde{\mathbf{a}}$ are identical to those of the system. In PMS, the system state and the model state share the same state space and are thus "subtractable" (Smith 2006). Define the observation at time t to be $\mathbf{s}_t = h(\tilde{\mathbf{x}}_t) + \boldsymbol{\eta}_t$, where $h(\cdot)$ is the observation operator that projects the true system state $\tilde{x}$ into observation space. For simplicity, $h(\cdot)$ is taken to be the identity operator below. Full observations are made; that is, observations are available for all state variables (all components of $\mathbf{x}$) at every observation time.[4] The $\boldsymbol{\eta}_t \in \mathbb{R}^m$ represent observational noise (or "measurement error"). The statistical characteristics of the observational noise (i.e., the noise model) for $\boldsymbol{\eta}_t$ are known exactly.

The problem of state estimation in PMS consists of forming an ensemble estimate of the current state $\tilde{x}_0$ given (i) the history of observations $\mathbf{s}_t$, for $t = -n + 1, \ldots, 0$; (ii) a perfect model class with perfect parameter values; and (iii) knowledge of the observational noise model. The various approaches that have been developed to address this problem divide into two major categories: the sequential approaches and the variational approaches. Sequential approaches can be built on the foundation of Bayesian theory, which conceptually generates a posterior distribution of the state variables by updating the prior distribution with new observations (Cohn 1997; Anderson and Anderson 1999). Unfortunately, application of the complete approach is computationally prohibitive in state estimation (Hamill 2006). An approximation to the Bayesian approach, called the Kalman filter, introduced by Kalman (1960), is optimal only for linear models and a Gaussian observational noise. To better address the state estimation problem in nonlinear cases, the extended Kalman filter (Jazwinski 1970; Gelb 1974; Ghil and Malanotte-Rizzoli 1991; Bouttier 1994) uses tangent linear dynamics to estimate the error covariances while assuming linear growth and normality of errors. The extended Kalman filter performs poorly where nonlinearity is pronounced; more accurate state estimates are available from ensemble Kalman filter approaches (Evensen 1994; Houtekamer and Mitchell 1998; Burgers et al. 1998; Lermusiaux and Robinson 1999; Anderson 2001; Bishop et al. 2001; Hamill et al. 2001), which have been developed using Monte Carlo[5] techniques. While these filters admit the non-Gaussian probability density function, only the mean and covariance are updated in these sequential approaches; information beyond the second

[2] For linear systems, see Wiener (1949), Kalman (1960), and references therein.

[3] In the perfect model scenario, the true state $\tilde{\mathbf{x}}$ is in the same state space as the model state $\mathbf{x}$ and $\tilde{m} = m$. Motivation for the use of tildes in this context can be found in Smith (2002).

[4] As shown elsewhere (Judd et al. 2008; Du 2009; Smith et al. 2010), various generalization to partial observations can be made. The approach could be applied in operational weather forecasting following Judd et al. (2008), using available 3DVAR analysis. The general case of partial observations will be considered elsewhere. In short, one could take a two pass approach to PDA, first using background information (e.g., the climatology distribution) of the unobserved state variables with the observations frozen to obtain initial estimates of unobserved state variables, and then applying full PDA as discussed below with those estimates of unobserved state variables and the original observed state variables. While interesting, discussion of this case is omitted here. Note there is some loss of generality in assuming full observations.

[5] As stressed by a reviewer, not all ensemble Kalman filter approaches need be considered as Monte Carlo approaches.

moment is discarded. Stressing the impact of these linear assumptions, Lawson and Hansen (2004) demonstrated that inaccuracies in state estimation are to be expected if the dynamics are nonlinear owing to higher moments of the distribution of the state variables. Another well-known sequential approach, the particle filter (Metropolis and Ulam 1944), is fully nonlinear in both model evolution and analysis steps. It, however, suffers from the so-called curse of dimensionality, which prevents the particles from staying "close" to the observations in a large-dimensional system (Snyder et al. 2008). Variational approaches [e.g., four-dimensional variational assimilation (Dimet and Talagrand 1986; Lorenc 1986; Talagrand and Courtier 1987; Courtier et al. 1994)], based upon maximum likelihood estimation, consist of shooting techniques that seek a set of initial conditions corresponding to a system trajectory that is consistent with a sequence of system observations; this is a strong constraint (Courtier et al. 1994; Bennett et al. 1996). In PMS, the variational approaches both are computationally expensive and are known to suffer from local minima owing to chaotic likelihoods (where the likelihood function of initial conditions is extremely jagged for chaotic systems). This was pointed out by Berliner (1991); see also Miller et al. (1994) and Pires et al. (1996). The approach presented in this paper applies the PDA approach to locate a reference trajectory and construct an initial condition ensemble by sampling the local state space. Khare and Smith (2011) applied a similar approach with a target of indistinguishable states (Q density) to form an initial-condition ensemble. More details about PDA and indistinguishable states can be found in Du (2009). State estimation outside PMS is discussed in Part II.

3. Pseudo-orbit data assimilation in sequence space

The analytic intractability of the relevant probability distributions and the dimension of the model state space suggest the adoption of a (Monte Carlo) ensemble scheme (Lorenz 1965; see also Smith 1996 and Leutbecher and Palmer 2008) to account for the uncertainties of observations in state estimation approach. The ensemble approach is by far the most common for quantifying uncertainty in operational weather forecasting (e.g., Toth and Kalnay 1993; Leutbecher and Palmer 2008). An algorithm may generate an ensemble directly, as with the particle filter and ensemble Kalman filters, or an ensemble may be generated from perturbations of a reference trajectory. The approach presented in this paper belongs in the second category. Of course, the quality of the state estimates will vary strongly with the quality of the reference trajectory(s). The pseudo-orbit data assimilation approach (Judd and Smith 2001; Ridout and Judd 2002; Judd and Smith 2004; Stemler and Judd 2009) provides a reference model trajectory given a sequence of observations. A brief introduction to the PDA approach is given in the following paragraph.

Let the dimension of the model state space be m and the number of observation times in the window be n. The sequence space is an $m \times n$ dimensional space in which a single point can be thought of as a particular series of n states $\mathbf{u}_t$, $t = -n + 1, \ldots, 0$. Here each $\mathbf{u}_t$ is an m-dimensional vector. Some points in the sequence space are trajectories of the model and some are not. Define a pseudo orbit, $\mathbf{U} \equiv \{\mathbf{u}_{-n+1}, \ldots, \mathbf{u}_{-1}, \mathbf{u}_0\}$, to be a point in the $m \times n$ dimensional sequence space for which $\mathbf{u}_{t+1} \neq F(\mathbf{u}_t)$ for any[6] component of $\mathbf{U}$. This implies that $\mathbf{U}$ corresponds to a sequence of model states that is not a trajectory of the model. Define the mismatch as an $m \times (n-1)$ dimensional vector $(\mathbf{e}_{-n+1}, \ldots, \mathbf{e}_{-1})$, where the component of the mismatch at time t is $\mathbf{e}_t = |F(\mathbf{u}_t) - \mathbf{u}_{t+1}|$, $t = -\mathrm{n} + 1, \ldots, -1$.

By construction, model trajectories have a mismatch of zero. A gradient descent (GD) algorithm (details in the following paragraph) can be used to minimize the sum of the squared mismatch errors. Define the mismatch cost function to be

$$C(\mathbf{U}) = \sum \mathbf{e}_t^2 . \tag{1}$$

The pseudo-orbit data assimilation minimizes the mismatch cost function for $\mathbf{U}$ in the $m \times n$ dimensional sequence space. The minimum of the mismatch cost function can be obtained by solving the ordinary differential equation:

$$\frac{d\mathbf{U}}{d\tau} = -\nabla C(\mathbf{U}), \tag{2}$$

where τ denotes algorithmic time. An important advantage of PDA is that the minimization is done in the sequence space: information from across the assimilation window is used simultaneously. Let the elements of $\mathbf{U}$ corresponding to the model state at a given time be called a component of the pseudo orbit. PDA optimizes all components simultaneously.

The observations themselves projected into the model state space define a pseudo orbit—call this pseudo orbit the observation-based pseudo-orbit $\mathbf{S} \equiv \{\mathbf{s}_{-n+1}, \ldots, \mathbf{s}_{-1}, \mathbf{s}_0\}$; with probability 1 that it will not be a trajectory.

[6] Technically, the inequality need hold only for one pair of consecutive components in the sequence space vector. Alternatively, one could define pseudo orbits so as to include trajectories; in this paper, this is not done.

In practice, the minimization is initialized with the observation-based pseudo orbit; that is, ${}^{0}\mathbf{U} = \mathbf{S}$, where the presuperscript 0 on $\mathbf{U}$ denotes the initial stage of the GD. After every iteration of the GD minimization, $\mathbf{U}$ will be updated. Recall that the pseudo orbit is a point in the sequence space; it is updated in the sense that under the GD algorithm it moves toward the manifold of trajectories. All points on the trajectory manifold have zero mismatch (each is a trajectory) and only points on the trajectory manifold have zero mismatch. To iterate the algorithm, one needs to differentiate the mismatch cost function given by

$$\frac{\partial C(\mathbf{U})}{\partial \mathbf{u}_t} = 2 \times \begin{cases} -[\mathbf{u}_{t+1} - F(\mathbf{u}_t)]d_tF(\mathbf{u}_t) & t = -n+1 \\ [\mathbf{u}_t - F(\mathbf{u}_{t-1})] - [\mathbf{u}_{t+1} - F(\mathbf{u}_t)]d_tF(\mathbf{u}_t) & -n+1 < t < 0, \\ \mathbf{u}_t - F(\mathbf{u}_{t-1}) & t = 0 \end{cases} \tag{3}$$

where $d_tF(\mathbf{u}_t)$ is the Jacobian of the model F at $\mathbf{u}_t$. The ordinary differential equation [Eq. (2)] is solved using the Euler approximation.

The mismatch cost function has no local minima other than on the manifold[7] for which $C(\mathbf{U}) = 0$ (Judd and Smith 2001). Let the result of the GD minimization be ${}^{\alpha}\mathbf{U}$, where α indicates discrete algorithmic time in GD (i.e., the number of iterations of the GD minimization). As $\alpha \to \infty$, ${}^{\alpha}\mathbf{U} \equiv \{{}^{\alpha}\mathbf{u}_{-n+1}, \ldots, {}^{\alpha}\mathbf{u}_0\}$ asymptotically approaches a trajectory of the model. In other words, PDA takes the observation-based pseudo orbit toward a model trajectory (i.e., toward the trajectory manifold). That trajectory need not shadow the observations in any sense (Smith et al. 2010); its merits for state estimation must be demonstrated empirically.[8] In practice, the GD minimization is run for a finite time and thus a pseudo orbit is obtained[9] rather than a trajectory. (In the experiments presented in the paper, the minimization is terminated after 1024 minimization iterations.) Each component of a pseudo orbit defines a model trajectory. To keep the notation clear, denote the pseudo orbit obtained from some large finite GD runs as $\mathbf{y}_{-n+1}, \ldots, \mathbf{y}_0$ and iterate the middle component[10] $\mathbf{y}_{-n/2}$ forward to create a segment of model trajectory $\mathbf{z}_{-n/2}, \ldots, \mathbf{z}_0$ [$\mathbf{y}_{-n/2} \equiv \mathbf{z}_{-n/2}$ and $\mathbf{z}_{t+1} = F(\mathbf{z}_t)$]. Such a model trajectory defines a reference trajectory. The middle component is expected to provide a better estimate of model state at that time than the end component at its time, as the middle component has information from both its past and its future, while the end component only has information from its past. It is important to note that although the PDA approach can be applied to any length of observation window, given finite computational resource a reference trajectory corresponding to the midcomponent will almost certainly diverge from the pseudo orbit when n is large, simply as a consequence of sensitivity to initial conditions. The midcomponent need not always be used.

To form an ensemble of initial conditions,[11] first generate a large number of model trajectories, called candidate trajectories, from which ensemble members can be selected. Ensemble members are drawn from the candidate trajectories according to their relative likelihood given the segment of observations. There are many ways to produce candidate trajectories; three methods of producing candidate trajectories are listed here: (i) Sample the local space around the reference trajectory. One can perturb the starting component of the reference trajectory and iterate the perturbed component forward to create candidate trajectories. (ii) Perturb the whole segment of observations $\mathbf{s}_t$, $t = -n+1, \ldots, 0$ and apply PDA onto the perturbed orbit to produce the candidate trajectories—that is, the same way that the reference trajectory is produced. (iii) Similar to method (ii), perturb the reference trajectory and repeat PDA. Although methods (ii) and (iii) may produce more informative candidates, they are obviously more expensive than method (i) since the GD minimization must be repeated. To make the computational cost between PDA and other state-estimation approaches more comparable, method (i) is used to generate candidate trajectories to produce the results presented in the paper. (Details about computational

[7] Back substitution of the solution of $\mathbf{u}_0 - F(\mathbf{u}_{-1}) = 0$ into Eq. (3) shows that the only critical points for $C(\mathbf{U})$ have $\mathbf{u}_t - F(\mathbf{u}_{t-1}) = 0$ for all t in $-n+1 \le t \le 0$.

[8] In fact, the trajectory need not be near the observations at all. The authors conjecture that the manifolds of interest have large reach in the high-dimensional sequence space, and thus the "curse of dimensionality" comes to the aid of PDA.

[9] Such a single pseudo orbit itself may provide a good estimation of the trajectory over the window.

[10] If n is odd, take $\mathbf{y}_{(-n+1)/2}$.

[11] Ideally, one forms a perfect ensemble under the model by sampling the states that define model trajectories that are consistent with past observations. This is, however, prohibitively expensive computationally.

costs can be found in appendix C.) The perturbations to the starting component $\mathbf{z}_{-n/2}$ are generated using a random variable $\boldsymbol{\zeta}$, $\boldsymbol{\zeta}$ is Gaussian with zero mean and standard deviation $\boldsymbol{\sigma}_p \in \mathbb{R}^m$, and $\boldsymbol{\sigma}_p$ is a constant diagonal matrix estimated by standard deviation of the difference between $\tilde{\mathbf{x}}_{-n/2}$ (the truth) and $\mathbf{z}_{-n/2}$. In practice, the value of $\boldsymbol{\sigma}_p$ would have to be determined empirically. (Values of $\boldsymbol{\sigma}_p$ for the experiments below are given in Table B1.)

Given N^{cand} candidate trajectories, the relative likelihood of each is then used to select ensemble members of the assimilation. To form an N^{ens}-member ensemble estimate of the current state, randomly draw N^{ens} trajectories from N^{cand} candidate trajectories according to their log-likelihood function over the time interval $(-n/2, 0)$. Specifically,

$$L(\mathbf{z}^*) = \frac{1}{2} \sum_{t=-n/2}^{0} [h(\mathbf{z}_t^*) - \mathrm{s}_t]^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} [h(\mathbf{z}_t^*) - \mathrm{s}_t], \quad (4)$$

where $\boldsymbol{\Gamma}^{-1}$ is the inverse of the covariance matrix of the observational noise and $\mathbf{z}_t^*$ denotes a candidate trajectory. The end component (at $t = 0$) of each selected candidate trajectory $\mathbf{z}_0^*$ is treated as an ensemble member. In this case, each ensemble member is then of equal weight avoiding any confusion regarding how to interpret the Q density in Judd and Smith (2001).

4. Contrasting 4DVAR with PDA

4DVAR is a popular approach to noise reduction in data assimilation (Dimet and Talagrand 1986; Lorenc 1986; Talagrand and Courtier 1987; Courtier et al. 1994). 4DVAR is a shooting technique that seeks initial conditions of system trajectories consistent with a sequence of system observations. It aims to find the initial state of a model trajectory that minimizes a cost function reflecting the misfit between the trajectory and the observations. The 4DVAR cost function is

$$C_{\text{4DVAR}} = \frac{1}{2}(\mathbf{x}_{-n+1} - \mathbf{x}_{-n+1}^b)^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}_{-n+1} - \mathbf{x}_{-n+1}^b) + \frac{1}{2} \sum_{t=-n+1}^{0} [h(\mathbf{x}_t) - \mathrm{s}_t]^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} [h(\mathbf{x}_t) - \mathrm{s}_t], \quad (5)$$

where $\mathbf{x}_{-n+1}$ is the initial state and $\mathbf{x}_{t+1} = F(\mathbf{x}_t)$, $\mathbf{x}_{-n+1}^b$ is the background (or first guess) at $t = -n + 1$, $\mathbf{B}$ is the background error covariance matrix, $\mathbf{s}_t$ reflects the observations at time t, and $\boldsymbol{\Gamma}$ is the covariance matrix of the observational noise. The second term of the cost function can be easily derived from the maximum likelihood estimate under the assumption that the observational noise model is independent and identically distributed (IID) Gaussian. The first term (background term) of the cost function aims to take account of the information from previous estimates (and any other available prior distribution of the initial state). The background $\mathbf{x}_{-n+1}^b$ is typically provided by the 4DVAR analyses from previous cycles. An estimation of $\mathbf{B}$ is required to minimize the cost function. For 4DVAR experiments conducted here, $\mathbf{B}$ is obtained iteratively [following Fertig et al. (2007)] as follows. Initially run 4DVAR using an arbitrary background covariance matrix $\mathbf{B}_0$, then compute the covariance $\mathbf{B}_1$ of the difference between the true state[12] and the background at all of the assimilation times. Next, run 4DVAR using $\mathbf{B}_1$ as the background error covariance matrix and again compute the covariance $\mathbf{B}_2$ of the difference between the true state and the background. Repeat this process until the estimated background covariance matrix does not change significantly.

As in PDA, the 4DVAR analysis provides a reference trajectory for use in building an initial condition ensemble. Although both PDA and 4DVAR use the information of both the model dynamics and the observations to produce the model trajectories, there are fundamental differences between them.

The PDA cost function itself does not constrain the result to stay close to the observation-based pseudo orbit [Eq. (1)]. The GD minimization is, however, initialized with the observation-based pseudo orbit.[13] Unlike the 4DVAR approach, the PDA approach does not penalize ${}^{\alpha}\mathbf{U}$ if it strays far from the observation-based pseudo orbit; in fact, the PDA approach is almost certain to force ${}^{\alpha}\mathbf{U}$ to move away, on average, from the observation-based pseudo orbit as the minimization goes further and further (see Part II). The 4DVAR approach is derived from the maximum likelihood estimate in the case of additive IID Gaussian observational noise. For other noise models, including those non-Gaussian in distribution or with either spatial or temporal correlations (red noises), 4DVAR is expected to converge to an incorrect solution (Lu and Browning 1998). That is, the true state is not the minimum of the 4DVAR cost function even in expectation. The PDA approach itself does not impose any significant assumptions on the noise model.

Another important difference is that the 4DVAR approach considers only model trajectories, adjusting the initial condition of each model trajectory only to minimize its cost function in the m-dimensional state space [Eq. (5)]. It starts with a model trajectory and ends

[12] Using knowledge of the true states in this way confers some advantage to 4DVAR—an advantage not given to PDA.

[13] One may initialize the GD minimization with a better series of analyses if it is available.

with a model trajectory. The PDA approach converges to a model trajectory by minimizing the mismatch cost function in the $n \times m$ dimensional sequence space. It starts from a pseudo orbit and, if run to completion, approaches at a model trajectory. In practice, given only finite computational power, only a pseudo orbit is reached and, of course, each component of a pseudo orbit will (with probability 1) define a unique trajectory.

The behavior of the 4DVAR cost function can sometimes vary so strongly with the assimilation window that Berliner (1991) dubbed it a "chaotic likelihood." The number of local minima[14] in the 4DVAR cost function increases with the length of the data assimilation window (Miller et al. 1994; Pires et al. 1996). The results trapped in the local minima are likely to be inconsistent with the observations. Gauthier (1992), Stensrud and Bao (1992), and Miller et al. (1994) have performed 4DVAR experiments with the Lorenz63 system (Lorenz 1963). They each found that performance of 4DVAR varies significantly depending on the length of the assimilation window, and difficulties arise with the extension of the assimilation window owing to the occurrence of multiple minima in the cost function. Applying the 4DVAR approach, one faces the dilemma between the impacts of local minima with a long assimilation window and the loss of information from the model dynamics given only a short window. The mismatch cost function in PDA avoids this dilemma. Although the cost function itself does not have a unique minimum, all minima of the mismatch cost function are model trajectories. The major limitation of longer assimilation windows in PDA is merely computational cost. And, as a longer assimilation window allows more information from the model dynamics and observations, the quality of the assimilation improves.

To contrast the model trajectory produced in practice by 4DVAR with that generated by PDA, both approaches are applied to the Ikeda map (Ikeda 1979; Haramel et al. 1985) and to the 18-dimensional Lorenz96 system (Lorenz 1995). Details of the systems are given in appendix A. For each system, five different length assimilation windows are tested. For the Ikeda map, assimilation windows with lengths between 4 and 16 steps are considered—for Lorenz96, lengths between 12 and 60 h. In the case of Lorenz96, 6 h indicates 0.05 time unit of the Lorenz96 system; assuming that 1 time unit is equal to 5 days, the doubling time of the Lorenz96 system roughly matches the characteristic time scale of dissipation in the atmosphere [see Lorenz (1995) for details]. PDA uses a GD minimization algorithm; the minimization terminates after 1024 GD iterations for each assimilation. 4DVAR uses a nonlinear conjugate gradient descent algorithm[15] (using the Fletcher–Reeves formula; Shewchuk 1994) to minimize its cost function; the minimization terminates when the derivative of the cost function is small. (Details are given in appendix B.)

TABLE 1. Distance between the observation-based pseudo orbits and the model trajectory generated by 4DVAR and PDA for Ikeda map, and distance between the true states and the model trajectory generated by 4DVAR and PDA for the Ikeda map. The columns show the average distance (average) and the 90% bootstrap resampling bounds (lower and upper). The noise model is $N(0, 0.05^2)$. The statistics are calculated based on 8192 assimilations and 4096 bootstrap resamples are used to calculate the resampling bounds.

Window length (steps)	Distance from observations					
	Average		Lower		Upper	
	4DVAR	PDA	4DVAR	PDA	4DVAR	PDA
4	1.81	1.65	1.74	1.59	1.88	1.72
6	11.5	1.77	8.80	1.72	14.5	1.83
8	45.2	1.85	38.9	1.81	51.9	1.90
12	105.5	1.91	97.7	1.87	113.0	1.95
16	174.0	1.95	165.3	1.91	182.9	1.98
	Distance from truth					
4	0.24	0.60	0.22	0.55	0.27	0.65
6	9.83	0.38	7.01	0.35	12.6	0.41
8	43.1	0.27	37.6	0.25	49.0	0.30
12	103.6	0.17	96.5	0.16	111.9	0.18
16	172.4	0.13	164.0	0.12	182.4	0.14

The second term of the 4DVAR cost function in Eq. (5) [specifically $(1/n)\sum_{t=-n+1}^{0}(\mathbf{x}_t - \mathrm{s}_t)^{\mathrm{T}}\boldsymbol{\Gamma}_t^{-1}(\mathbf{x}_t - \mathrm{s}_t)$, which is the distance between the observation-based pseudo orbit and the model trajectory] and the distance between the true states and model trajectory $[(1/n)\sum_{t=-n+1}^{0}(\mathbf{x}_t - \tilde{\mathbf{x}}_t)^{\mathrm{T}}\boldsymbol{\Gamma}_t^{-1}(\mathbf{x}_t - \tilde{\mathbf{x}}_t)]$ are interpreted as diagnostic tools to evaluate the quality of the model trajectories generated. Results are presented in Tables 1 and 2. When the assimilation window is relatively short, both 4DVAR and PDA tend to generate model trajectories that are closer to the true states than to the observation-based pseudo orbit.[16] This happens both in the Ikeda and in Lorenz96 experiments, where each approach provides

[14] As stressed by a reviewer, it is conceivable that local minima are not a problem in all types of models as the window length increases.

[15] 4DVAR shows similar results under GD, conjugate gradient descent was employed to relieve any concerns about convergence in local minima. GD is retained for PDA inasmuch as it is adequate for our purposes and has advantages outside PMS (see Part II).

[16] Although the trajectories are slightly farther away from the observation-based pseudo orbit, they are still consistent with the observational noise.

TABLE 2. Distance between the observation-based pseudo orbits and the model trajectory generated by 4DVAR and PDA for the Lorenz96 system, and distance between the true states and the model trajectory generated by 4DVAR and PDA for the Lorenz96 system. The columns show the average distance (average) and the 90% bootstrap resampling bounds (lower and upper). The noise model is $N(0, 0.05^2)$. The statistics are calculated based on 8192 assimilations and 4096 bootstrap resamples are used to calculate the resampling bounds.

Window length (h)	Distance from observations					
	Average		Lower		Upper	
	4DVAR	PDA	4DVAR	PDA	4DVAR	PDA
12	16.3	15.58	16.2	15.51	16.4	15.65
24	17.1	15.91	17.0	15.86	17.1	15.97
36	24.7	16.06	24.1	16.01	25.2	16.10
48	39.7	16.13	38.7	16.10	40.8	16.17
60	56.0	16.17	53.8	16.14	58.2	58.0
	Distance from truth					
12	1.62	1.73	1.59	1.70	1.65	1.77
24	1.07	1.09	1.04	1.07	1.09	1.11
36	8.22	0.93	7.69	0.92	8.77	0.94
48	23.0	0.86	22.0	0.85	24.0	0.87
60	39.2	0.81	36.8	0.80	41.4	0.82

effective noise reduction. Note 4DVAR outperforms[17] PDA for very short windows. For the larger window lengths, however, PDA yields consistently the best results. Given the information available from the observations in a longer-window, PDA provides better state estimation than short-window 4DVAR. When the reference trajectory is ''closer'' to the true state of the system, a better ensemble is expected.

Various ''fixes'' have been proposed to allow the application of 4DVAR with a long window length avoiding local minima (see Pires et al. 1996 and references therein). Voss et al. (2004) also applied a multiple-shooting approach to address the local minima problem—an initial-value approach to short windows, resembling a similar spinup procedure applied to 4DVAR using multiple short windows. The approach remains expensive and Voss's examples show varying success. Abarbanel et al. (2009) successfully applied synchronization to smooth the (cost function) surfaces in the space of parameters and initial conditions. Abarbanel's approach also requires extensive computations and may prove more applicable to parameter estimation. In practice, application of 4DVAR has been restricted to relatively short windows. PDA can exploit information available in longer windows (see also Judd et al. 2004). Within PMS, there is valuable information both in the observations and in the model dynamics in longer windows of observations. When the model is imperfect, this case is less easy to make; focusing on pseudo orbits, however, still holds an additional advantage over 4DVAR: the ability to diagnose model error (this point is discussed in Part II).

5. Ensemble Kalman filter versus PDA

Another well-established approach to state estimation is sequential estimation (Kalman 1960; Anderson and Moore 1979; Kaipio and Somersalo 2005). With sequential approaches, one integrates the model forward until the time that observations are available; the state provided by the model at that time is usually called the first guess. The first guess is then modified using the new observations. Sequential approaches encode all knowledge gleaned from the past in the current state information. Alternatively, when windows over time are considered, an observation inconsistent with the dynamics of any trajectory over that window can be identified as such. Sequential approaches cannot do this. This is not a question of assigning an appropriate prior for the observational noise distribution, but rather one of seeing the dynamical inconsistency of a given observation within a particular region of state space. More generally, in high-dimensional nonlinear dissipative systems, the quest for a general encoding of such information analytically is misguided:[18] a given procedure must demonstrate its superiority in each case. Ensemble Kalman filter approaches (Evensen 1994; Burgers et al. 1998; Houtekamer and Mitchell 1998; Lermusiaux and Robinson 1999; Anderson 2001; Bishop et al. 2001; Hamill et al. 2001) can explore some nonlinearity. There are many different ensemble Kalman filters; the approach used here is the ensemble adjustment Kalman filter (Anderson 2001, 2003). Large ensemble sizes (512 members) have been considered in this case so as to avoid some of the complications required in operational implementations (i.e., ensemble covariance localization). Covariance inflation is adopted to improve the EnKF data assimilation results.[19] For each experiment the inflation parameter value is tuned to optimize the ignorance score. Values of inflation parameter for each

[17] Closer to the true state of the system.

[18] Arguably, Kalman foresaw this in footnote 4 of his original paper (Kalman 1960).

[19] To ensure the EnKF implementations are high-quality benchmarks, experiments paralleling those of Sakov and Oke (2008) were performed. Specifically, the EnKF approach was applied to Lorenz96 with 40 variables; the RMS result reflects the RMS results of various versions of EnKF presented in Fig. 4 of Sakov and Oke (2008) to within 5%.

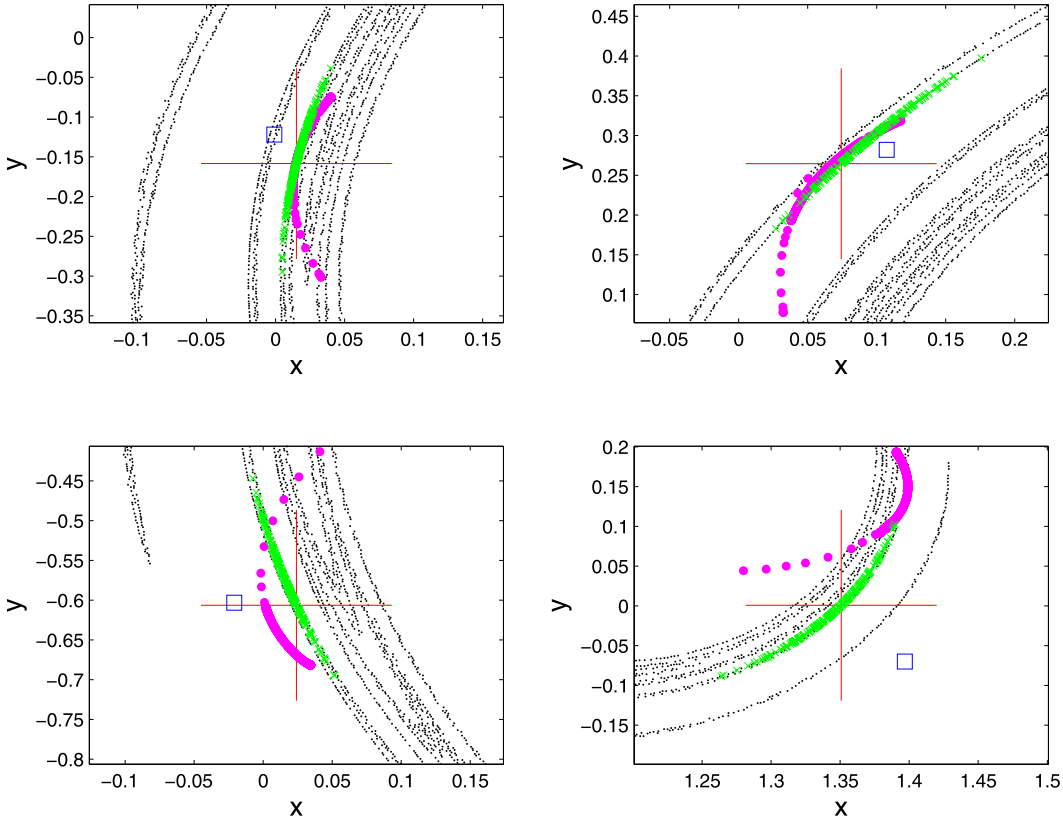

FIG. 1. Ensemble results from both EnKF and PDA for the Ikeda map. The true state is centered in each panel (large cross), the square is the corresponding observation, and the background dots indicate samples from the Ikeda map attractor. The EnKF ensemble is depicted by 512 magenta dots. PDA ensemble is depicted by 512 green crosses. Each panel is an example of one case of state estimation.

experiment are given in appendix B. Even after these adjustments to EnKF, its performance is inferior to PDA.

The comparison is first made in the lower-dimensional case in order to provide easily visualized evidence. Both PDA and the EnKF are applied in the two-dimensional Ikeda map and the resulting ensemble is plotted in the state space (the details of the experiments are given in appendix B). Four examples are shown in Fig. 1; in all panels the ensemble produced by the PDA approach is not only closer to the true state but also reflects the dynamical manifolds as the ensemble members lie near the system attractor. While the EnKF ensemble has its own distinctive structure, the ensemble members do not lie along the system attractor. This is expected in general, inasmuch as the EnKF approach assumes a second-moment closure, the distributions are assumed to be fully described by means and covariances (Anderson and Anderson 1999; Lawson and Hansen 2004; Hamill 2006). This may also lead to filter divergence or even catastrophic filter divergence, as reported by Harlim and Majda (2010). In the top panels of Fig. 1, the EnKF ensemble is distributed about the true state and fairly close to the model's attractor, while in the bottom panels, the ensemble members are systematically off the attractor and not well distributed about the true state.

To assess the difference between these two approaches quantitatively, the initial-condition ensemble is translated into a predictive distribution function by standard kernel dressing (Brocker and Smith 2008). Each ensemble member is replaced by a Gaussian kernel centered on that member, providing a continuous distribution (a non-Gaussian sum of Gaussian kernels). The width of each kernel (the standard deviation of the Gaussian, called the "kernel width") is determined by optimizing the ignorance score, introduced below.

The performance of a state-estimation technique can be evaluated with the "log p score" (ignorance score; Good 1952; Roulston and Smith 2002). The ignorance score is the only proper local score for continuous variables (Bernardo 1979; Raftery et al. 2005; Brocker and Smith 2007). Although there are other nonlocal proper scores, the authors prefer using ignorance as it has a clear interpretation in terms of information theory and can be easily communicated in terms of effective interest

TABLE 3. Ignorance score and optimized kernel width of initial condition ensemble for the Ikeda map and Lorenz96 system for various noise models. The 512-member ensembles generated by the PDA approach and the EnKF approach are compared. Lower and upper are the 90% bootstrap resampling bounds of the ignorance score, the statistics are calculated based on 8192 assimilations, and 4096 bootstrap resamples are used to calculate the resampling bounds.

		Ignorance		Lower		Upper		Kernel width	
Systems	$\boldsymbol{\sigma}_{\mathrm{ns}}$	EnKF	PDA	EnKF	PDA	EnKF	PDA	EnKF	PDA
Ikeda	0.01	−5.04	−6.86	−5.09	−6.91	−5.00	−6.81	0.005	0.0002
	0.02	−4.22	−5.77	−4.26	−5.79	−4.17	−5.75	0.016	0.0005
	0.05	−2.72	−4.11	−2.77	−4.15	−2.68	−4.06	0.021	0.001
Lorenz96	0.1	−7.42	−7.93	−7.46	−7.98	−7.37	−7.89	0.01	0.009
	0.5	−4.79	−5.36	−4.85	−5.41	−4.74	−5.30	0.08	0.06
	1	−3.62	−4.43	−3.69	−4.47	−3.56	−4.38	0.41	0.1

returns (Good 1952; Roulston and Smith 2002; Hagedorn and Smith 2009), not to mention the lack of any compelling example in favor of the general use of nonlocal scores. The ignorance score is defined by

$$S[p(y), Y] = -\log_2[p(Y)], \tag{6}$$

where Y is the outcome and $p(Y)$ is the probability of event Y. In practice, given K forecast–outcome pairs (p_i, Y_i, $i = 1, \ldots, K$), the empirical average ignorance skill score is

$$S_{\mathrm{Emp}}[p(y), Y] = \frac{1}{K}\sum_{i=1}^{K} -\log_2[p_i(Y_i)]. \tag{7}$$

The PDA approach is compared with the EnKF approach in both the lower-dimensional Ikeda map and higher-dimensional Lorenz96 flow (the details of the systems are given in appendix A). Noise models of the form $\boldsymbol{\eta} \sim N(0, \boldsymbol{\sigma}_{\mathrm{ns}}^2)$ are considered where each element of $\boldsymbol{\eta}$ is identically distributed. In both cases the empirical ignorance score is computed using 8192 assimilations. Table 3 shows the comparison between EnKF and PDA using the ignorance score (the optimized kernel width is also presented). To quantify the robustness of the result, 90% bootstrap resampling bounds of the ignorance score are also presented. From the table, it is clear that in both experiments the ensemble generated by PDA significantly outperforms the one generated by EnKF. In the Ikeda experiment, the relative ignorance between the two approaches is found to be around 1.4 bits for the noise model with $\boldsymbol{\sigma}_{\mathrm{ns}} = 0.05$ and 1.8 bits for $\boldsymbol{\sigma}_{\mathrm{ns}} = 0.01$. This can be interpreted as PDA placing, on average, 160% (and 250%) more probability mass on the outcome than EnKF. In the Lorenz96 experiment, relative ignorance between the two approaches is found to be around 0.8 bits for the noise model with $\boldsymbol{\sigma}_{\mathrm{ns}} = 1$ and 0.5 bits for $\boldsymbol{\sigma}_{\mathrm{ns}} = 0.1$, which can be interpreted as PDA placing, on average, 75% (and 40%) more probability mass on the outcome than EnKF. The smaller kernel width for the PDA ensemble also indicates that the PDA ensemble members are more concentrated around the true state than the EnKF ensemble. Note that although PDA seems to outperform EnKF, the computational cost for PDA is much greater in order to obtain pseudo orbit with small mismatch errors. (Computational cost is discussed in appendix C.) This cost is substantially reduced outside PMS, as with imperfect models PDA need not approach a trajectory. This issue is discussed in Part II.

6. Conclusions

A new methodology for state estimation in the perfect model scenario is introduced. This pseudo-orbit data assimilation (PDA) approach aims to identify a reference trajectory about which an ensemble can be assembled.

The well-established 4DVAR approach is contrasted with the PDA approach. Results for the 4DVAR approach based on the two-dimensional Ikeda map and the 18-dimensional Lorenz96 flow are limited by the occurrence of local minima. It has been noted by Miller et al. (1994) and Pires et al. (1996) that 4DVAR suffers from the multiple local minima when applied to long windows. Long windows, on the other hand, allow benefits from having more dynamical information; PDA can exploit these benefits. The 4DVAR approach is expected to fail in practice in cases of chaotic likelihood (Berliner 1991). PDA can solve this problem posed by Berliner (H. Du and L. A. Smith 2014, unpublished manuscript).

Comparisons between the PDA approach and the EnKF approach have been made in the lower-dimensional Ikeda map and a higher-dimensional Lorenz96 model. By looking at the ensembles generated in the state space of the Ikeda map, the structure of the ensemble obtained by the PDA approach seems to be more consistent with

the model dynamics (closer to the attractor) than that of the ensemble produced by the EnKF approach. By evaluating initial-condition ensembles using ignorance, in both the Ikeda map and Lorenz96 model experiments, it is demonstrated that the PDA approach systematically outperforms the EnKF approach considered (Anderson 2001, 2003). The failure of EnKF is due in large part to the loss of information beyond the second moment.

One might ask why any approach might be expected to provide better state estimation than the statistically "most likely" state given the observations. In dynamical systems with attractors, the statistically most likely state given only the observations will not lie on the attractor (with probability 1); similarly, the trajectory that generated the observations will not provide the most likely state at $t = 0$ (with probability 1). By allowing the use of long windows of observations, PDA gains access to more information in the dynamics; this allows more "balanced" states in the sense that relationships between components of the state vector are preserved. This includes relationships that reflect dynamically realized states (speaking loosely, states "on the attractor" and assuming one exists). If the system admits coherent structures, longer trajectories near realized states will reflect more realistic structures and their evolution, as observed by Judd et al. (2008). The key here is reducing the role of statistical distance (which does not respect such structures) and increasing the attention to the geometry of the realized flow (which does); longer windows are an advantage here (Judd et al. 2004, 2008). Within PMS, PDA might be applied to determine initial states for 4DVAR, thereby extending the window length accessible to 4DVAR. Outside the perfect model scenario, PDA finds states more consistent with the model dynamics at the cost of optimizing the statistical fit in the examples considered; this is arguably a generalization of balance to include time and coherent structure (see Part II). The aim is for coherent structures of the system to be reproduced to the extent that the model can reproduce them, and then not to be perverted to improve an inappropriate statistical fit to the observations. A description of PDA outside PMS is presented in Part II. Data assimilation for deterministic nonlinear models will always be a challenging task. PDA provides a step forward by allowing an enhanced balance between extracting information from the dynamic equations and information in the observations.

Acknowledgments. This research was supported by the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re. L. A. Smith gratefully acknowledges support from Pembroke College, Oxford.

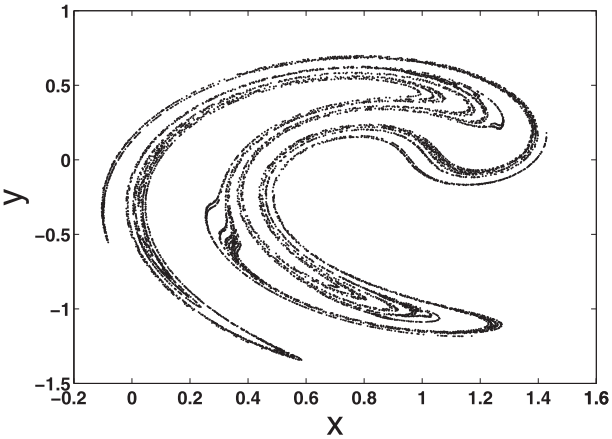

FIG. A1. The attractor of the Ikeda map.

APPENDIX A

Dynamical Systems

The Ikeda map was introduced by Haramel et al. (1985) based on Ikeda's model of laser pulses in an optical cavity (Ikeda 1979). With real variables it has the form

$$X_{n+1} = \gamma + u(X_n \cos\phi - Y_n \sin\phi), \tag{A1}$$

$$Y_{n+1} = u(X_n \sin\phi + Y_n \cos\phi), \tag{A2}$$

where $\phi = \beta - \alpha/(1 + X_n^2 + Y_n^2)$. With the parameters $\alpha = 6, \beta = 0.4, \gamma = 1$, and $u = 0.83$, the system is believed to be chaotic. Figure A1 shows the attractor of the Ikeda map in the state space.

A system of nonlinear ordinary differential equations (Lorenz96 system) was introduced by Lorenz (1995). The variables involved in the system are intended to resemble an atmospheric variable regionally distributed around the Earth. For the system containing m variables $x_1, \ldots, x_m$ with cyclic boundary conditions (where $x_{m+1} = x_1$), the equations are

$$\frac{dx_i}{dt} = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + F, \tag{A3}$$

where the parameter F is set to be 10 in all of the experiments considered following Smith (2000) and Orrell (1999). The model is simulated using a standard fourth-order Runge–Kutta scheme. The simulation time step is 0.01 time unit and the model time step is 0.05; that is, each model time step is conducted by five steps of the fourth-order Runge–Kutta integrator.

TABLE B1. Details of the PDA implementation. Note that for results comparing PDA with 4DVAR, window length varies as stated in Tables 1 and 2.

	Ikeda			Lorenz96		
Observational noise	$N(0, 0.01^2)$	$N(0, 0.02^2)$	$N(0, 0.05^2)$	$N(0, 0.1^2)$	$N(0, 0.5^2)$	$N(0, 1^2)$
Window length	$16\Delta t$	$16\Delta t$	$16\Delta t$	$32\Delta t$	$32\Delta t$	$32\Delta t$
No. of GD iterations	1024	1024	1024	1024	1024	1024
GD step size	0.1	0.2	0.4	0.2	0.4	0.8
Perturbation $\boldsymbol{\sigma}_p$	0.002	0.005	0.01	0.02	0.08	0.12
No. of candidates	4096	4096	4096	4096	4096	4096
No. of ensemble members	512	512	512	512	512	512

APPENDIX B

Experiments' Details

Details of the experiments discussed in the paper are given here.

- PDA
  - (i) Generate pseudo-orbit ${}^{\alpha}\mathbf{U}$ by minimizing the mismatch cost function using GD. A fixed GD minimization step size is used. The minimization stops after 1024 GD iterations (i.e., $\alpha = 1024$ here).
  - (ii) A reference trajectory is obtained by iterating the middle state of ${}^{\alpha}\mathbf{U}$ forward in time until $t = 0$.
  - (iii) Generate a number of candidate trajectories by perturbing the starting component of the reference trajectory and iterate forward in time until $t = 0$.
  - (iv) Select ensemble members from candidate trajectories by random draw according to the log-likelihood function of the candidate trajectories.

  Table B1 provides specific experimental details of the PDA implementation conducted in the paper.
- 4DVAR

  A model trajectory is generated by minimizing the 4DVAR cost function [Eq. (5)] using a nonlinear conjugate gradient descent algorithm (using the Fletcher–Reeves formula; Shewchuk 1994). The minimization step size is calculated by using the secant (Allen and Isaacson 1998) method to approximate the second derivative. The minimization terminates when the derivative of the cost function is small—specifically, when the ratio of the length of the derivative vector in the updated model state to that of the initial state is smaller than 10^{-4}.
- EnKF

  The ensemble adjustment Kalman filter (Anderson 2001, 2003) is applied to produce an ensemble of initial conditions. Large ensemble sizes (512 members) have been considered in this case so as to avoid some of the complications required in operational implementations (i.e., ensemble covariance localization). Covariance inflation is adopted to improve the EnKF data assimilation results. For each experiment the inflation parameter value is properly tuned in order to achieve a better ignorance score. Table B2 provides specific experimental details of the EnKF implementation conducted in the paper.

APPENDIX C

Computational Costs

Information concerning the computational costs is provided here.

- PDA versus 4DVAR

  For each minimization step, 4DVAR requires running the initial state $l_w - 1$ steps forward (l_w is the assimilation window length), PDA requires running $l_w - 1$ states one step forward. Both approaches therefore require running the model $l_w - 1$ times. When calculating the gradient of the cost function, PDA requires the adjoint of the model for each state vector within the assimilation window; 4DVAR requires

TABLE B2. Details of the EnKF implementation. Note that the initial ensemble is generated by perturbing the observation with the inverse of observational noise; the first 1000 assimilations (as transient) are not considered in the evaluations.

	Ikeda			Lorenz96		
Observational noise	$N(0, 0.01^2)$	$N(0, 0.02^2)$	$N(0, 0.05^2)$	$N(0, 0.1^2)$	$N(0, 0.5^2)$	$N(0, 1^2)$
Inflation parameter	1.0	1.01	1.02	1	1.02	1.04
No. of ensemble members	512	512	512	512	512	512

not only the adjoint of the each model state along the trajectory within the assimilation window but also multiplying those adjoints together, which makes 4DVAR slightly more expensive. The computational cost for the other parts of each approach is known to scale at a lower order than the cost of the model integrations. In general, the computational cost per step for PDA and 4DVAR is rather similar. Computational costs decrease when an algorithm converges more quickly; while 4DVAR often converges in fewer iterations than PDA, it also converges significantly further from a desired target than PDA in long windows; in short windows it is comparable.

- PDA versus EnKF

For EnKF, each ensemble member requires one model integration; the computational cost for updating the analysis ensemble is $O(n_{\mathrm{ens}}^2 p + n_{\mathrm{ens}}^3 + n_{\mathrm{ens}}^2 m)$ (Tippett et al. 2003), where n_{ens} is the ensemble size, m is the dimension of model state space, and p is the number of observations (m in our case). For PDA (implemented according to the experiments presented in the paper) to generate the reference trajectory requires $1024 \times (l_w - 1)$ model runs. To generate the candidate trajectories and for selecting ensemble members, it requires $4 \times (l_w/2 - 1)$ model runs for each ensemble member. The rest of the computational cost for PDA is known to scale at lower order than the cost of the model integrations. Obviously, the computational cost for PDA is significantly more expensive than for EnKF. As shown in Part II, the cost of PDA is substantially reduced when it is applied outside PMS; this makes PDA feasible in practice. A more efficient minimization algorithm would further reduce the cost.

In short, PDA is shown to provide significantly improved state estimation at a higher cost than EnKF and a comparable cost to 4DVAR. The extent to which the improved state estimation justifies the additional cost will vary with the details of the application. A central aim of this paper is merely to establish that results from PDA are in fact distinct and, at times, can be superior.

REFERENCES

Abarbanel, H. D. I., D. R. Creveling, R. Farsian, and M. Kostuk, 2009: Dynamical state and parameter estimation. *SIAM J. Appl. Dyn. Syst.,* **8,** 1341–1381.

Allen, M. B., and E. L. Isaacson, 1998: The secant method. *Numerical Analysis for Applied Science,* M. B. Allen et al., Eds., John Wiley and Sons, 188–195.

Anderson, B. D. O., and J. B. Moore, 1979: *Optimal Filtering.* Prentice-Hall, 357 pp.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.,* **129,** 2884–2903.

——, 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.,* **131,** 634–642.

——, and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.,* **127,** 2741–2758.

——, B. Wyman, S. Q. Zhang, and T. Hoar, 2005: Assimilation of surface pressure observations using an ensemble filter in an idealized global atmospheric prediction system. *J. Atmos. Sci.,* **62,** 2925–2938.

Bennett, A. F., B. S. Chua, and L. M. Leslie, 1996: Generalized inversion of a global numerical weather prediction model. *Meteor. Atmos. Phys.,* **60,** 165–178.

Berliner, M. L., 1991: Likelihood and Bayesian prediction for chaotic systems. *J. Amer. Stat. Assoc.,* **86,** 938–952.

Bernardo, J. M., 1979: Expected information as expected utility. *Ann. Stat.,* **7,** 686–690.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.,* **129,** 420–436.

Bouttier, F., 1994: A dynamical estimation of forecast error covariances in an assimilation system. *Mon. Wea. Rev.,* **122,** 2376–2390.

Brocker, J., and L. A. Smith, 2007: Scoring probabilistic forecasts: On the importance of being proper. *Wea. Forecasting,* **22,** 382–388.

——, and ——, 2008: From ensemble forecasts to predictive distribution functions. *Tellus,* **60,** 663–678.

Burgers, G., P. J. V. Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.,* **126,** 1719–1724.

Cohn, S. E., 1997: An introduction to estimation theory. *J. Meteor. Soc. Japan,* **75,** 257–288.

Courtier, P., J. N. Thepaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.,* **120,** 1367–1387.

Dimet, F.-X. L., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus,* **38A,** 97–110.

Du, H., 2009: Combining statistical methods with dynamical insight to improve nonlinear estimation. Ph.D. dissertation, London School of Economics and Political Science, 190 pp.

——, and L. A. Smith, 2012: Parameter estimation through ignorance. *Phys. Rev. E,* **86,** 016213, doi:10.1103/PhysRevE.86.016213.

——, and ——, 2014: Pseudo-orbit data assimilation. Part II: Assimilation with imperfect models. *J. Atmos. Sci.,* **71,** 483–495.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99** (C5), 10 143–10 162.

Fertig, E. J., J. Harlim, and B. R. Hunt, 2007: A comparative study of 4D-VAR and a 4D ensemble Kalman filter: Perfect model simulations with Lorenz-96. *Tellus,* **59A,** 96–100.

Gauthier, P., 1992: Chaos and quadri-dimensional data assimilation: A study based on the Lorenz model. *Tellus,* **44A,** 2–17.

Gelb, A., 1974: *Applied Optimal Estimation.* MIT Press, 382 pp.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics,* Vol. 33, Academic Press, 141–266.

Gilmour, I., 1998: Nonlinear model evaluation: ι-shadowing, probabilistic prediction and weather forecasting. Ph.D. dissertation, University of Oxford, 184 pp.

Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc.,* **14A,** 107–114.

Hagedorn, R., and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.,* **16,** 143–155.

Hamill, T. M., 2006: Ensemble-based atmospheric data assimilation. *Predictability of Weather and Climate,* T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 124–156.

——, J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background-error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.,* **129,** 2776–2790.

Haramel, S. M., C. K. R. T. Jones, and J. V. Moloney, 1985: Global dynamical behavior of the optical field in a ring cavity. *J. Opt. Soc. Amer.,* **2B,** 552–564.

Harlim, J., and A. J. Majda, 2010: Catastrophic filter divergence in filtering nonlinear dissipative systems. *Commun. Math. Sci.,* **8,** 27–43.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

——, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with the ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.,* **133,** 604–620.

Ikeda, K., 1979: Multiple valued stationarity state and its instability of the transmitted light by a ring cavity system. *Opt. Commun.,* **30,** 257–261.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, 376 pp.

Judd, K., and L. A. Smith, 2001: Indistinguishable states I: The perfect model scenario. *Physica D,* **151,** 125–141.

——, and ——, 2004: Indistinguishable states II: The imperfect model scenario. *Physica D,* **196,** 224–242.

——, ——, and A. Weisheimer, 2004: Gradient free descent: Shadowing and state estimation with limited derivative information. *Physica D,* **190,** 153–166.

——, C. A. Reynolds, T. E. Rosmond, and L. A. Smith, 2008: The geometry of model error. *J. Atmos. Sci.,* **65,** 1749–1772.

Kaipio, J., and E. Somersalo, 2005: *Statistical and Computational Inverse Problems.* Applied Mathematical Sciences, Vol. 160, Springer, 344 pp.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.,* **82,** 35–45.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, 364 pp.

Keppenne, C. L., and M. Rienecker, 2002: Initial testing of a massively parallel ensemble Kalman filter with the Poseidon isopycnal ocean general circulation model. *Mon. Wea. Rev.,* **130,** 2951–2965.

Khare, S., and L. Smith, 2011: Data assimilation: A fully nonlinear approach to ensemble formation using indistinguishable states. *Mon. Wea. Rev.,* **139,** 2080–2097.

Lalley, S. P., 2000: Removing the noise from chaos plus noise. *Nonlinear Dynamics and Statistics,* A. I. Mees, Ed., Birkhäuser, 233–244.

Lawson, G., and J. A. Hansen, 2004: Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth. *Mon. Wea. Rev.,* **132,** 1966–1981.

Lermusiaux, P. F. J., and A. R. Robinson, 1999: Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Mon. Wea. Rev.,* **127,** 1385–1407.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.,* **227,** 3515–3539.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.,* **112,** 1177–1194.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1965: On the possible reasons for long-period fluctuations. WMO-IUGG symposium on research and development aspects of long-range forecasting, Boulder, Colorado, 1964, World Meteorological Organization Tech. Note 66, 203–211.

——, 1995: Predictability: A problem partly solved. *Proc. Seminar on Predictability,* Shinfield Park, United Kingdom, ECMWF, 40–58.

Lu, C. G., and G. L. Browning, 1998: The impact of observational and model errors on four-dimensional variational data assimilation. *J. Atmos. Sci.,* **55,** 995–1011.

McSharry, P. E., and L. A. Smith, 1999: Better nonlinear models from noisy data: Attractors with maximum likelihood. *Phys. Rev. Lett.,* **83,** 4285–4288.

Metropolis, N., and S. Ulam, 1944: The Monte Carlo method. *J. Amer. Stat. Assoc.,* **44,** 335–341.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.,* **51,** 1037–1056.

Orrell, D., 1999: A shadow of a doubt: Model error, uncertainty, and shadowing in nonlinear dynamical systems. Ph.D. dissertation, University of Oxford, 199 pp.

Pires, C., R. Vautard, and O. Talagrand, 1996: On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus,* **48A,** 96–121.

Pisarenko, V. F., and D. Sornette, 2004: Statistical methods of parameter estimation for deterministically chaotic time series. *Phys. Rev. E,* **69,** 036122, doi:10.1103/PhysRevE.69.036122.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Ridout, D., and K. Judd, 2002: Convergence properties of gradient descent noise reduction. *Physica D,* **165,** 27–48.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

Sakov, P., and P. R. Oke, 2008: Implications of the form of the ensemble transformations in the ensemble square root filters. *Mon. Wea. Rev.,* **136,** 1042–1053.

Schittkowski, K., 1994: Parameter estimation in systems of nonlinear equations. *Numer. Math.,* **68,** 129–142.

Shewchuk, J. R., 1994: An introduction to the conjugate gradient method without the agonizing pain. Carnegie Mellon University Tech. Rep., 58 pp.

Smith, L. A., 1996: Accountability and error in ensemble forecasting. *Proc. Seminar on Predictability,* Shinfield Park, United Kingdom, ECMWF, 351–368.

——, 2000: Disentangling uncertainty and error: On the predictability of nonlinear systems. *Nonlinear Dynamics and Statistics,* A. I. Mees, Ed., Birkhäuser, 31–64.

——, 2002: What might we learn from climate forecasts? *Proc. Natl. Acad. Sci. USA,* **4,** 2487–2492.

——, 2006: Predictability past predictability present. *Predictability of Weather and Climate,* T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 217–250.

——, M. C. Cuellar, H. Du, and K. Judd, 2010: Exploiting dynamical coherence: A geometric approach to parameter estimation in nonlinear models. *Phys. Lett.,* **374A,** 2618–2623.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to high-dimensional particle filtering. *Mon. Wea. Rev.,* **136,** 4629–4640.

Stemler, T., and K. Judd, 2009: A guide to using shadowing filters for forecasting and state estimation. *Physica D,* **238,** 1260–1273.

Stensrud, D. J., and J. W. Bao, 1992: Behaviors of variational and nudging assimilation techniques with a chaotic low-order model. *Mon. Wea. Rev.,* **120,** 3016–3028.

Talagrand, O., and P. Courtier, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quart. J. Roy. Meteor. Soc.,* **113,** 1311–1328.

Tarantola, A., 2004: *Inverse Problem Theory and Methods for Model Parameter Estimation.* SIAM, 352 pp.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea. Rev.,* **131,** 1485–1490.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

Voss, H. U., J. Timmer, and J. Kurths, 2004: Nonlinear dynamical system identification from uncertain and indirect measurements. *Int. J. Bifurcation Chaos,* **14,** 1905–1933.

Wiener, N., 1949: *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications.* MIT Press, 163 pp.