Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica D 190 (2004) 153–166
PHYSICA D
www.elsevier.com/locate/physd

# Gradient free descent: shadowing, and state estimation using limited derivative information

Kevin Judd [a,b,*], Leonard Smith [b,c], Antje Weisheimer [c]

[a] *Department of Mathematics and Statistics, Centre for Applied Dynamics and Optimization, The University of Western Australia, Perth, Australia*
[b] *Department of Statistics, Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, UK*
[c] *Oxford Centre for Industrial and Applied Mathematics, Mathematics Institute, Oxford, UK*

Received 11 December 2002; received in revised form 25 September 2003; accepted 29 October 2003
Communicated by C.K.R.T. Jones

**Abstract**

Shadowing trajectories can play an important role in assessing the reliability of forecasting models, they can also play an important role in providing state estimates for ensemble forecasts. Gradient descent methods provide one approach for obtaining shadowing trajectories, which have been shown to have many useful properties. There remains the important question whether shadowing trajectories can be found in very high-dimensional systems, like weather and climate models. The principle impediment is the need to compute the derivative (or adjoint) of the system dynamics. In this paper we investigate gradient descent methods that use limited derivative information. We demonstrate the methods with an application to a moderately high-dimensional system using no derivative information at all.


*PACS:* 92.60.Wc; 05.45

*Keywords:* State estimation; Shadowing

## 1. Conceptual introduction

This research is motivated by a desire for finding solutions to weather and climate forecasting models that are consistent with atmospheric observations. These so-called *shadowing trajectories*[1] [2–4,9,11], are valuable for a number of reasons. For example, if one wants to make detailed, say, 4-day forecasts, one would hope that there are always trajectories of the forecast model that consistently shadow past observations over this period of time, because

* Corresponding author. Tel.: +61-8-9380-1357; fax: +61-8-9380-1028.
*E-mail address:* kevin@maths.uwa.edu.au (K. Judd).

[1] We use of the term *shadowing* to compare the trajectory of a mathematical model and a set of observations, following Smith [15]. This differs from the more common usage when contrasting two mathematical models. The proximity of an orbit to a set of target states derived from observations justifies the name shadowing only when it is consistent with respect to the observational noise; yet the observational noise (in the model state space) is only defined with respect to shadowing orbits of a perfect model. This dilemma will be discussed elsewhere. In the perfect model scenario, of course, infinitely long shadows always exist.

0167-2789/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2003.10.011

if this were not so, then the forecasts could not be reliable. The period of time over which shadowing is reliable, gives an indication of the accuracy of the model, and the atmospheric states when shadowing times are short, may reveal flaws in the model, or reveal states that are particularly unpredictable. Shadowing trajectories also play an important role in state estimation and ensemble forecasting, as demonstrated by the theory of indistinguishable states [9,10] and comparisons of shadowing methods to extended Kalman filters [8]. It should be noted that the sense of shadowing used here is distinct from, the original strong notion shadowing associated with questions of whether a numerically computed trajectory shadows a true trajectory [6,7]; our sense of shadowing is more closely related to the ideas of nonlinear noise reduction [1,5,13].

This paper addresses the question of how might one find trajectories that shadow observations in high-dimensional dynamical systems. There are a number of established algorithms for finding shadowing trajectories in low dimension systems, also referred to as nonlinear noise reduction [1,5,11,13]. These methods cannot be applied in high-dimensional systems because they require either explicit, or implicit, knowledge of the derivative of the model dynamics. For low-dimensional systems, derivatives are easily computed, or approximated, as required, but not so for high-dimensional systems. The principle content of this paper is an investigation of how one might proceed with only limited information about model derivatives, and a demonstration that useful shadowing can be obtained in a moderately high-dimensional system without any derivative information. These results suggest that methods like those described might be adapted for high-dimensional systems and encourage further investigation.

## 2. Mathematical introduction

We make a distinction between a system and a model: the system is reality, for example, the actual atmosphere, and the model is a mathematical or computer representation of the system. A model is never perfect, and neither are observations of the system. Even if the model were perfect it would be a difficult task to find a trajectory of the model consistent with observations over long periods of time [8,16]. When a chaotic model is imperfect this is impossible, but the period of time over which consistency can be maintained is a measure of the reliability of the model.

Observations of the system are by various means interpolated into states of the model, a process often referred to as *data assimilation*, and result in model states often referred to as the *analysis*. Typically an analysis is derived from the observations by an interpolation technique like 3D variational assimilation. The shadowing techniques we discuss are not necessarily intended to be a replacement for such interpolation techniques, they may just augment them so as to provide better analyses. However, when observations are relatively complete, which is not currently the situation for weather systems, they may eliminate the need for variational assimilation.

Suppose then we have a sequence of states $s_0, \ldots, s_n$ for a model $f$ of a dynamical system, where the states have been derived from observations by some means. Ideally this sequence of states is a trajectory of the model, that is, it is a sequence $x_i$ such that $x_{i+1} = f(x_i)$, but this is highly unlikely, because the observations are effected by measurement error, or because the observations were a sparse sample of the state variables, or because the model is an imperfect model of the system. The important question is whether there exists a trajectory $x_0, \ldots, x_n$ of the model consistent with the original observations and the derived analysis sequence $s_0, \ldots, s_n$, by always remaining close, $\|s_i - x_i\| < \epsilon$, where $\epsilon$ is some bound on the acceptable error. Such a trajectory, if one exists, will be called a *shadowing* trajectory.

One much studied collection of methods for finding shadowing trajectories are the *gradient descent methods*. These methods start with the initial sequence of states $s_0, \ldots, s_n$ and make continual adjustments to all the states so that asymptotically the adjusted sequence of states approaches a shadowing trajectory. The adjustments are made so as to move down the gradient of a specified "cost function" that measures the "distance" a sequence of states are from being a suitable shadowing trajectory. Since gradient descent methods require computing the gradient

of the cost function, this implies having to compute the derivative of the system dynamics. For low-dimensional systems this is not usually a problem, but for high-dimensional systems, like weather forecasting models, this is either difficult or impossible.

The main idea of this note is that in order to minimize the cost function to attain a shadowing trajectory, one does not necessarily have to take the direction of *steepest* descent; there are other descent directions that attain shadowing trajectories. These descent directions may require limited, or no, information about the derivative of the system dynamics. It should be made clear from the outset that different descent directions (or methodologies) will obtain different shadowing trajectories, just as different cost functions would, and the resulting trajectories may not be optimal by some other criteria, but are none-the-less useful.

This note proposes very simple descent methods that need not use any derivative information at all. These methods are probably not of any great value for low-dimensional systems, but might be of considerable value for high-dimensional systems such as those encountered in operational weather forecasting or climate modeling. We also discuss whether one need necessarily push convergence to a shadowing trajectory; in some circumstances a shadowing pseudo-orbit might be sufficient, that is, a sequence of states that are almost, but not quite, a trajectory. The states of a long shadowing pseudo-orbit can often be iterated to obtain short shadowing trajectories.

## 3. Gradient descent of indeterminism

Consider a dynamical system for which there is a $d$-dimensional model of the dynamics. The system is observed over a period of time and from these observations are obtained a sequence of estimates $s_i \in \mathrm{R}^d$, $i = 1, \ldots, n$, for the state of the model; call these the *analysis*. Just how the analysis is obtained is irrelevant to the following discussion, for example, the analysis could be just interpolated raw observations.

The aim is to find a trajectory $x_i \in \mathrm{R}^d$, $i = 1, \ldots, n$ of a model, $x_{i+1} = f(x_i)$, such that the trajectory shadows the observations, that is, the trajectory remains close to, or consistent with, the observations.[2] The cost function we chose to use here is *indeterminism* relative to the model $f$. For any sequence of states $x = (x_1, \ldots, x_n)$, $x_i \in \mathrm{R}^d$, $i = 1, \ldots, n$, define its *indeterminism*,[3] by

$$L(x) = \frac{1}{2} \sum_{i=1}^{n-1} \|x_{i+1} - f(x_i)\|^2, \tag{1}$$

where we make the convenient identification of the sequence of $n$ states in $\mathrm{R}^d$ with a single point $x \in \mathrm{R}^{nd}$. Clearly, $L(x) = 0$ if $x$ is a trajectory of the model $f$. When the analysis $s = (s_1, s_2, \ldots, s_n) \in \mathrm{R}^{nd}$ is not a trajectory the *gradient descent algorithm* starts at the initial point $x = s$, then follows the gradient of $L$ in $\mathrm{R}^{nd}$ down the steepest descent path to a minimum where $L(x) = 0$. This is equivalent to solving the differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = -\nabla L(x(\tau)) \tag{2}$$

with $x(0) = s$ and finding the limit of $x(\tau)$ as $\tau \to \infty$. A discussion of the properties of this gradient descent method can be found in [9,10], Ridout and Judd [12], and Judd [8]; in particular it is shown that the above gradient descent algorithm always converges to a trajectory of the model, that is, $L(x(\tau))$ converges to zero, and furthermore,

[2] We will not specify how this closeness is measured. Indeed the particular cost function to be discussed does not take this into account, but rather it relies on the analysis being close to the observations so that a trajectory obtained by minimal perturbation of the analysis will naturally be close to the observations. It is possible to devise cost functions that take closeness or consistency into account, but this can result is undesirable effects, for example, convergence to a pseudo-orbit rather than a trajectory. We return to this point later.

[3] Strictly speaking $L(x)$ is half the indeterminism. The half is introduced for later convenience.

the method is optimal in the sense that for bounded observational noise and increasingly large $n$, the algorithm converges to the true trajectory for a perfect model of a hyperbolic system.

Throughout it is assumed that the model $f$ is a diffeomorphism on $\mathrm{R}^d$, which implies in particular that the Jacobian derivative matrix $\mathrm{d}f(x)$ has rank $d$ everywhere.

Writing out the gradient in Eq. (2) explicitly for each component we have that

$$\frac{\partial L}{\partial x_i} = \begin{cases} \mathrm{d}f(x_1)^{\mathrm{T}}(x_2 - f(x_1)), & i = 1, \\ -(x_i - f(x_{i-1})) + \mathrm{d}f(x_i)(x_{i+1} - f(x_i)), & 1 < i < n, \\ -(x_n - f(x_{n-1})), & i = n, \end{cases} \tag{3}$$

where $\mathrm{d}f(x_i)$ is the Jacobian derivative of $f$ at $x_i$. This can be all written a little more compactly by defining the $(n-1)d \times nd$ block diagonal matrix,

$$A = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \vdots \\ & A_{n-1} & 0 \end{pmatrix}, \tag{4}$$

and another $(n-1)d \times nd$ block diagonal matrix,

$$S = \begin{pmatrix} 0 & I & & \\ & \ddots & \ddots & \\ & & 0 & I \end{pmatrix}, \tag{5}$$

where $I$ is the $d \times d$ identity, 0 is a $d \times d$ zero matrix, $A_i = \mathrm{d}f(x_i)$ and unspecified entries are zero. Defining $\delta_i(x) = x_{i+1} - f(x_i)$, and $\delta(x) = (\delta_1(x), \ldots, \delta_{n-1}(x))$, where $\delta(x)$ is considered a column vector in $\mathrm{R}^{(n-1)d}$, then Eq. (2) becomes

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = -(S - A)^{\mathrm{T}}\delta(x(\tau)). \tag{6}$$

It should be emphasized here that $A$ is dependent on $x$ and ought to be denoted $A(x(\tau))$, but for convenience we drop the explicit notation.

## 4. Descent with limited derivative information

A difficulty with applying gradient descent in high-dimensional models is the computation of the derivatives of $f$, that is, the elements of $A$. One might wonder what would happen if $A$ were replaced by some approximation $\tilde{A}$, perhaps obtained by some approximation of each $\mathrm{d}f(x_i)$ using a small ensemble of points around each $x_i$? Consider solving

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = -(S - \tilde{A})^{\mathrm{T}}\delta(x(\tau)). \tag{7}$$

Does the solution $x(\tau)$ of this equation converge to a trajectory of $f$, that is, does $L(x(\tau))$ converge to zero, as it can be shown to with Eq. (2) and its equivalent form (6)? We address this question now.

**Result 1.** Eq. (7) has fixed points where $L(x) = 0$ and nowhere else.

**Proof.** It is clear that Eq. (7) can have fixed points only where $\delta(x) = 0$ or $\delta(x) \in \ker(S - \tilde{A})^{\mathrm{T}}$. Now $\delta(x) = 0$ if and only if $L(x) = 0$. On the other hand, note that any matrix of the form $(S - \tilde{A})$ has rank $(n-1)d$, and hence $(S - \tilde{A})^{\mathrm{T}}$ has trivial kernel by the rank-nullity theorem. □

**Result 2.** $L(x)$ is a Lyapunov function of Eq. (7) if $(S - A)(S - \tilde{A})^{\mathrm{T}}$ is positive definite for all $x$.

**Proof.** $L(x)$ is a Lyapunov function of Eq. (7) if the vector field of Eq. (7) always points inward on level sets of $L(x)$. This is equivalent to requiring that the vector field of Eq. (6), which are the inward normal vectors of the level sets of $L(x)$, has positive inner product with the vector field of Eq. (7) at every point. This can be assured if

$$w^{\mathrm{T}}(S - A)(S - \tilde{A})^{\mathrm{T}}w > 0 \quad \forall w \in \mathrm{R}^{(n-1)d}, \quad w \neq 0, \tag{8}$$

that is, $(S - A)(S - \tilde{A})^{\mathrm{T}}$ is positive definite. □

Note that Result 2 is a sufficient condition, it is not necessary. Also, recall that $L(x) = 0$ is a manifold of fixed points of Eqs. (6) and (7). This manifold is a global attractor of Eq. (6): this follows from Result 2 because if $\tilde{A} = A$, then $(S - A)(S - \tilde{A})^{\mathrm{T}}$ is symmetric and of full rank and so positive definite. If $L(x)$ is not a Lyapunov function for Eq. (7) it might be useful to know when $L(x) = 0$ is locally stable, although we will not use this next result.

**Result 3.** The set of fixed points $L(x) = 0$ can only be locally stable for Eq. (7) if $(S - \tilde{A})^{\mathrm{T}}(S - A)$ has no negative eigenvalues.

**Proof.** Let $x$ be a fixed point and consider the linearization of Eq. (7) about this point. Now,

$$\delta_i(x + w) = x_{i+1} + w_{i+1} - f(x_i + w_i), \tag{9}$$

$$\delta_i(x + w) = \delta_i(x) + w_{i+1} - \mathrm{d}f(x_i)w_i + \mathrm{O}(w^2), \tag{10}$$

and so the linearization about $x$, in terms of $w \in \mathrm{R}^{nd}$, is

$$\dot{w} = -(S - \tilde{A})^{\mathrm{T}}(S - A)w. \tag{11}$$

Hence, stability at $x$ requires at least that $(S - \tilde{A})^{\mathrm{T}}(S - A)$ has no negative eigenvalues. If the eigenvectors with zero eigenvalues correspond to $\ker(S - A)$ and these vectors span the tangent plane of the manifold of $L(x) = 0$, then $L(x) = 0$ must be locally stable by the generalized Hartman–Grobman theorem. In the proof of Result 1 it was shown that $\ker(S - \tilde{A})^{\mathrm{T}}$ is trivial, hence the eigenvectors with zero eigenvalues are exactly $\ker(S - A)$. □

It is useful for the following to note that the $(n-1)d \times (n-1)d$ matrix $(S - A)(S - \tilde{A})^{\mathrm{T}}$ has the form

$$(S - A)(S - \tilde{A})^{\mathrm{T}} = \begin{pmatrix} I + A_1\tilde{A}_1^{\mathrm{T}} & -\tilde{A}_2^{\mathrm{T}} & & & \\ -A_2 & I + A_2\tilde{A}_2^{\mathrm{T}} & -\tilde{A}_3^{\mathrm{T}} & & \\ & -A_3 & \ddots & \ddots & \\ & & \ddots & I + A_{n-2}\tilde{A}_{n-2}^{\mathrm{T}} & -\tilde{A}_{n-2}^{\mathrm{T}} \\ & & & -A_{n-1} & I + A_{n-1}\tilde{A}_{n-1}^{\mathrm{T}} \end{pmatrix}, \tag{12}$$

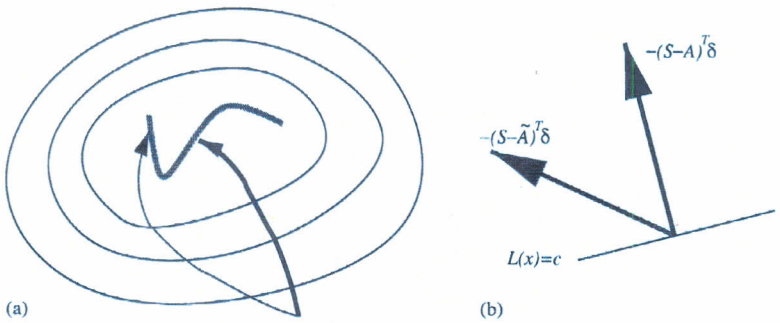


Fig. 1. Schematic representation of gradient descent with alternative descent. (a) The (red) shape in the middle represents the set of trajectories in $\mathrm{R}^{nd}$, that is, where $L(x) = 0$. The concentric thin loops represent the level sets of $L(x)$. The thick (black) arrow represents the path of steepest descent from an initial position, where as the thinner (blue) arrow represents the descent path, which is not the steepest descent path, but does at least descend to a trajectory. (b) Schematic of the relationship between the vectors $-(S - A)^{\mathrm{T}}\delta$ and $-(S - \tilde{A})^{\mathrm{T}}\delta$. The vector $-(S - A)^{\mathrm{T}}\delta$ is always perpendicular to the contour $L(x) = c$. If the inner product of the two vectors is positive, then $-(S - \tilde{A})^{\mathrm{T}}\delta$ is also pointing downhill, although not in the direction of steepest descent. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

and the $nd \times nd$ matrix $(S - \tilde{A})^{\mathrm{T}}(S - A)$ has the form

$$(S - \tilde{A})^{\mathrm{T}}(S - A) = \begin{pmatrix} \tilde{A}_1^{\mathrm{T}}A_1 & -\tilde{A}_2^{\mathrm{T}} & & & \\ -A_1 & I + \tilde{A}_2^{\mathrm{T}}A_2 & -\tilde{A}_3^{\mathrm{T}} & & \\ & -A_2 & \ddots & \ddots & \\ & & \ddots & I + \tilde{A}_{n-1}^{\mathrm{T}}A_{n-1} & -\tilde{A}_{n-1}^{\mathrm{T}} \\ & & & -A_{n-1} & I \end{pmatrix}. \tag{13}$$

Fig. 1 gives a schematic representation of how the descent algorithm works with approximate gradients. Result 2 ensures that vector field defined by $-(S - \tilde{A})^{\mathrm{T}}\delta$ always points downhill, where as the vector field $-(S - A)^{\mathrm{T}}\delta$ points in the steepest descent direction, Fig. 1(b). Such vector fields result in paths that lead to the set $L(x) = 0$, although the paths will be different, that is, lead to different trajectories, Fig. 1(a). The steepest descent direction is not necessarily the quickest path to a trajectory, on the other hand, it is not necessary that alternative descent fields may yield shadowing trajectories which are closer to the observations.

Although Eqs. (2), (6) and (7) seem to imply one must solve a very high-dimensional ordinary differential equation, the preceding results suggest that one does not have to be too careful about following a particular path. Thus the differential equations may often be integrated by a simple iterative Euler step method, $x(\tau+\mathrm{d}\tau) = x(\tau)-\mathrm{d}\tau\nabla L(x(\tau))$, etc.

### 4.1. A sufficient substitution

Consider the particularly simple instance of Eq. (7), obtained by setting $\tilde{A}_i = \lambda I$, $i = 1, \ldots, n-1$, for some $\lambda \in \mathrm{R}$. Result 2 requires finding $\lambda$ such that $(S - A)(S - \tilde{A})^{\mathrm{T}}$ is positive definite for all $x$. From the structure of Eq. (12) it is easily shown that in this simple case the principle block minors $M_{bj}$, $j = 1, \ldots, n-1$, that is, the determinant of the upper-left $jd \times jd$ matrix, of this matrix have the form

$$\begin{aligned} M_{b1} &= \det(I + \lambda A_1), \qquad M_{b2} = \det(I + \lambda A_1 + \lambda^2 A_1 A_2), \\ M_{b3} &= \det(I + \lambda A_1 + \lambda^2 A_1 A_2 + \lambda^3 A_1 A_2 A_3), \end{aligned} \tag{14}$$

This is strongly suggestive, by a standard matrix result, that $(S - A)(S - \tilde{A})^{\mathrm{T}}$ is positive definite if $|\lambda|$ is sufficiently small, because it would appear that the principle minors are strictly positive. However, one needs to show *all* the principle minors are strictly positive, not just the block minors.

**Result 4.** If $\tilde{A}_i = \lambda I \ \forall i$, then $(S - A)(S - \tilde{A})^{\mathrm{T}}$ is positive definite for sufficiently small $|\lambda|$.

**Proof.** It is sufficient to show all the principle minors are strictly positive. Let $M_k$ be the $k \times k$ upper-left sub-matrix of $(S - A)(S - \tilde{A})^{\mathrm{T}}$. From Eq. (12) it can be seen that

$$M_{k+1} = \begin{pmatrix} M_k & \lambda b_k \\ c_k^{\mathrm{T}} & 1 + \lambda a_k \end{pmatrix} \tag{15}$$

for some scalar $a_k$ and column vectors $b_k$ and $c_k$, which may depend polynomially on $\lambda$. It follows that

$$\det M_{k+1} = \lambda d_k + (1 + \lambda a_k)\det M_k \tag{16}$$

for some scalar $d_k$, which can only depend polynomially on $\lambda$. Since $M_1 = 1 + \lambda a_0$ for some scalar $a_0$ it follows that $\det M_k = 1 + e_k\lambda$ for some scalar $e_k$, which can only depend polynomially on $\lambda$. Hence, $\det M_k > 0$ for all $k$ for sufficiently small $|\lambda|$. □

It would appear that a suitably small value of $\lambda$ is less than the $d$th power of the reciprocal of the largest modulus of the eigenvalues of the $A_i$. On the other hand, Result 2 provides a sufficient condition and is stronger than necessary, which means a larger value of $\lambda$ may be sufficient.

### 4.2. *Discussion of $\lambda I$-substitution and shadowing*

Although convergence to a trajectory is assured with sufficiently small $\lambda$, one should not be overly enthusiastic about the prospects of obtaining a shadowing trajectory in close proximity to the observations. It is certainly the case that such a simple substitution for $\tilde{A}$ could have enormous computational advantages, but recall that convergence to a trajectory does not imply convergence to a shadowing trajectory; it was only assumed that the trajectory would remain close to observations by making minimal perturbations of the initial analysis. Furthermore, it might well happen that a $\lambda I$-substitution converges to a trajectory more costly than computing a full adjoint or some other approximation.[4]

First note that if $\lambda = 0$ one obtains a trivial solution trajectory, that is, one obtains, after considerable computation, the trajectory with $x_1 = s_1$, which is, the trajectory through the unperturbed initial point. Such a trajectory will almost certainly not be a long-term shadowing trajectory.

To understand the $\lambda I$-substitution better it might be useful to consider the algorithm in more geometric terms. Consider Eq. (3) with $\mathrm{d}f(x_i)$ replaced by $\lambda I$. It is seen that all but the first and last $x_i$ have two sources of correction: a *forward* correction in the direction of the mismatch in determinism $x_i - f(x_{i-1})$, and a *backward* correction in the direction $x_{i+1} - f(x_i)$, which is the mismatch in determinism at the next point $x_{i+1}$. The $\lambda I$ substitution applies a simple scaling to the backward correction, where as the original gradient descent (Eq. (3)) scales and rotates this vector by projecting onto the adjoint $\mathrm{d}f(x_i)^{\mathrm{T}}$. Solution of the descent equation (7) propagates determinism errors forward and backward along the sequence of states.

*Aside*: It might be thought that the $\lambda I$ substitution decouples the components, indeed computationally the $\lambda I$ substitution is extremely efficient because the error corrections are computed for each component separately, there

[4] For weather models a full moist adjoint is generally not available, so this is a moot point. However, the theory developed here also suggests that if a dry adjoint is available, then this might be successfully used as an approximation to the moist adjoint.

are no matrix multiplications. The coupling between components is not removed, however, as the coupling appears through $\delta(x(\tau))$, because $\delta(x(\tau))$ is updated at each integration step. In general, if $\lambda$ is too small, then corrections propagate faster in the forward direction than the backward direction, and can lead to trajectories that are almost exactly a trajectory through the initial $x_1 = s_1$.

If $\lambda$ is made larger, so as to balance the propagation of errors in forward and backward directions, then the condition of Result 2 may fail. In practice, if the initial sequence of states is far from being a trajectory, then a large value, say $\lambda = 0.5$, can result in considerable movement toward determinism before the condition in Result 2 fails. Failure of this condition is easily recognized as an increase in $L(x)$.

By stopping the descent when $L(x)$ attains a local minimum for a fixed value of $\lambda$, one then obtains a shadowing pseudo-orbit $x$ with $L(x) > 0$; a pseudo-orbit being a sequence of states that do not quite form a trajectory. For a perfect model, the original observations are a proximate pseudo-orbit of the true trajectory, but the hope is the new pseudo-orbit obtained from the descent has smaller indeterminism $L(x)$, and is closer to *truth* in some sense.

There are several reasons why one might accept a pseudo-orbit rather than pursue a shadowing trajectory. A good reason, for imperfect models, is simply that there does not exist a shadowing trajectory. In both perfect and imperfect models the fact remains that the "true" states cannot be known; there are many states indistinguishable from the true states [9,10]. A pseudo-orbit might be as good a sequence of states as one might hope to obtain up to indistinguishably. In the next section we discuss an application of the $\lambda I$-substitution algorithm to a climate model, and observe that pseudo-orbits are useful in this application.

## 5. Shadowing in a quasi-geostrophic climate model

The $\lambda I$-substitution in the shadowing equation (7) has been used with a simplified general circulation model of Weisheimer et al. [14]. This is a model of large-scale atmospheric flow in the troposphere and stratosphere. The model is a three layer quasi-geostrophic (QG) model, with circulation forced by a meridional temperature gradient between the equator and poles, which simulates solar forcing in a simple way, and schematic orography consisting of two major mountain up-lifts. The model has been run with horizontal spectral resolution T21 corresponding to $32 \times 64$ grid points or a $5.2° \times 5.2°$ resolution. The Runge–Kutta integration time step was 1 h, with the map $f$ being a 24 h integration.

The shadowing experiments were done in the perfect model scenario, that is, the model is identical to system from which the observations are obtained. The observations were simulated by perturbing a *true* trajectory (as spectral variables of dimensionless units) with uniform Gaussian noise. The test trajectories had 100 points, that is, 100 days. Pseudo-orbits were found by letting $\lambda = 0.5$ and solving the descent equation (7), with the raw observations as the initial condition, using Euler steps of size 0.2 until $L(x)$ attained a local minimum. The final pseudo-orbit obtained this way will be called the *best* pseudo-orbit, which means the best we obtained by this simple descent. A better pseudo-orbit almost certainly could be obtained by a more sophisticated adjustment of $\lambda$ or better approximation of the derivative $\tilde{A}$.

Observational noise can have the effect of obliterating information about spectral components with small variance. From a 1000-day run it was found that the standard deviation of the 1518 spectral components have an approximate power-law distribution. For reference the maximum standard deviation of any component was $4 \times 10^{-4}$. The experiments reported here have a noise level of $\sigma = 4 \times 10^{-5}$ for which we find 72% of spectral components have standard deviation less than the noise level.

Fig. 2 shows the change in determinism between the initial observed sequence and the best pseudo-orbit. The mismatch in determinism is displayed by plotting $\|x_{i+1} - f(x_i)\|$ for $i = 1$–99. There is a significant improvement in mismatch along the pseudo-orbit, approximately an order of magnitude improvement. The best pseudo-orbit has

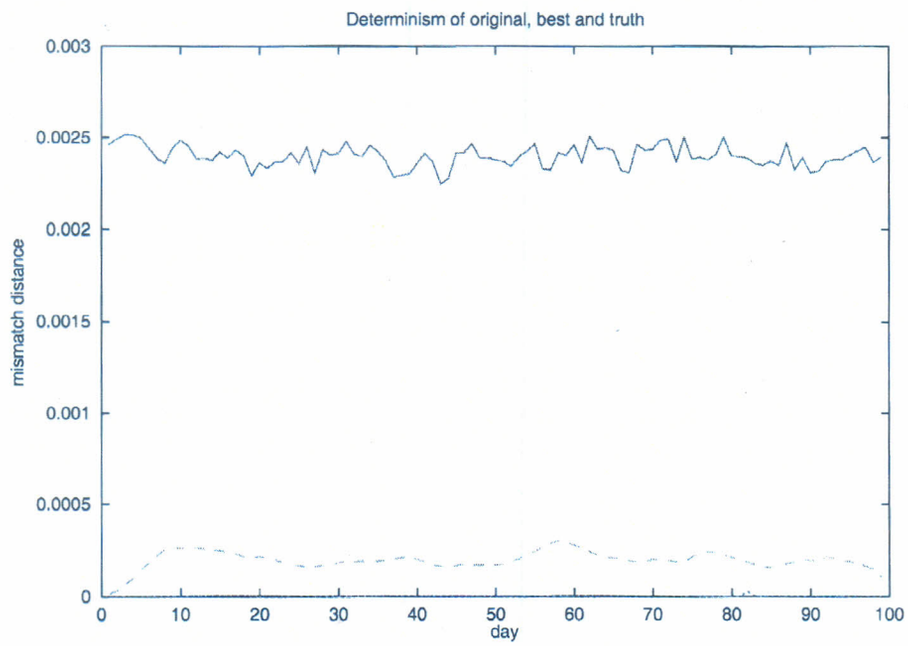


Fig. 2. The mismatch from determinism $\|x_{i+1} - f(x_i)\|$ at each $i$ for the initial observed sequence (solid line), and the best shadowing pseudo-orbit as described in the text (dashed line). This is the QG model described in the text with observational noise $4 \times 10^{-5}$, which exceeds the standard deviation for approximate 72% of spectral components.

more or less uniform mismatch, although at the beginning, $i = 0$, and end, $i = 99$, there is smaller indeterminism (the mismatch dips to nearly zero), and around $i = 10$ and 55 there are periods of larger indeterminism (local maxima in the mismatch). The smaller indeterminism at the beginning and end is an effect of the descent algorithm, because here errors are only modified in one direction and tend to converge faster. As we will see this is generally not a good effect, certainly not at $i = 0$. Will we also see later that the two periods of larger indeterminism (greater mismatch) correspond to atmospheric states that are particularly unpredictable.

Fig. 3 shows the distance between the initial observed sequence and the true trajectory, and also the distance between the best pseudo-orbit and the true trajectory. The separation distance is displayed by plotting $\|x_i - \hat{x}_i\|$ for $i = 1$–100, where $\hat{x}_i$ is the true trajectory and $x_i$ the sequence of states compared. There is a significant reduction in the distance from truth, mostly by a factor of 4, except near the beginning where the initial point of the best pseudo-orbit differs very little from the observations. This can be understood using the theoretical analysis of Judd and Smith [9], and Ridout and Judd [12], where it is noted that gradient descent trajectories deviate from truth at the beginning of the trajectories in the stable direction, and at the end of trajectories in the unstable direction. Why the effect is much more noticeable at the beginning is not entirely clear, there are a number of reasons why this might occur, which need to be further investigated. Note that there is considerable variation of the separation distance along the pseudo-orbit, and that there appears to be a correlation between the separation distance and mismatch, that is, for $t > 20$ larger mismatch in determinism correlates with larger separation from truth.

The next experimental results shown in Fig. 4 investigate the relationship between the best pseudo-orbit and shadowing trajectories. The best pseudo-orbit is not a trajectory, but we can compute the trajectories of each state in the pseudo-orbit. We expect these trajectories to approximate truth fairly well, and at least shadow longer than the states corresponding to the initial observations. In Fig. 4 we compute the trajectory of each state of the best

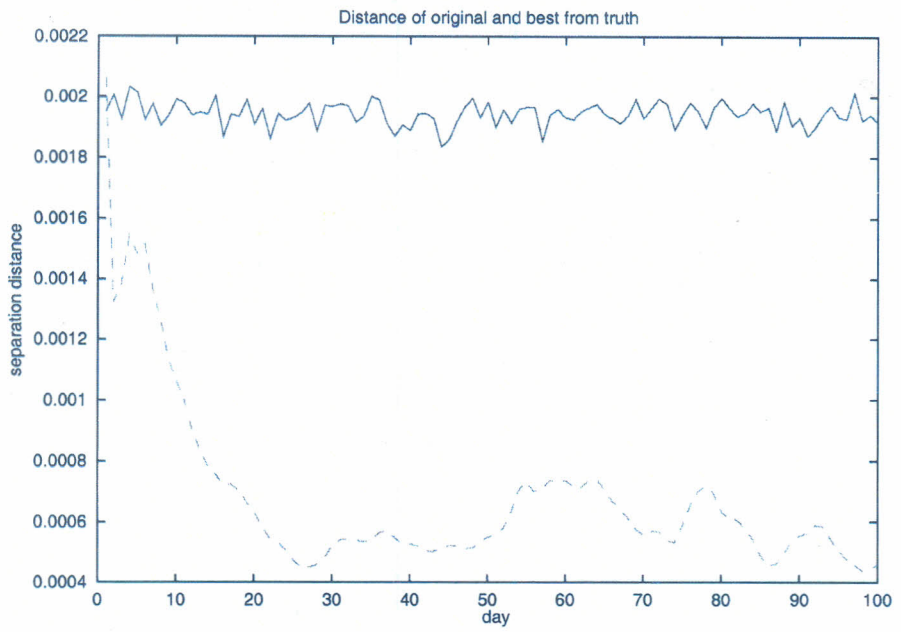


Fig. 3. Separation distance $\|\hat{x}_i - x_i\|$ between true trajectory $\hat{x}_i$ and initial observed sequence (solid line) and the best shadowing pseudo-orbit (dashed line). Other details as in Fig. 2.

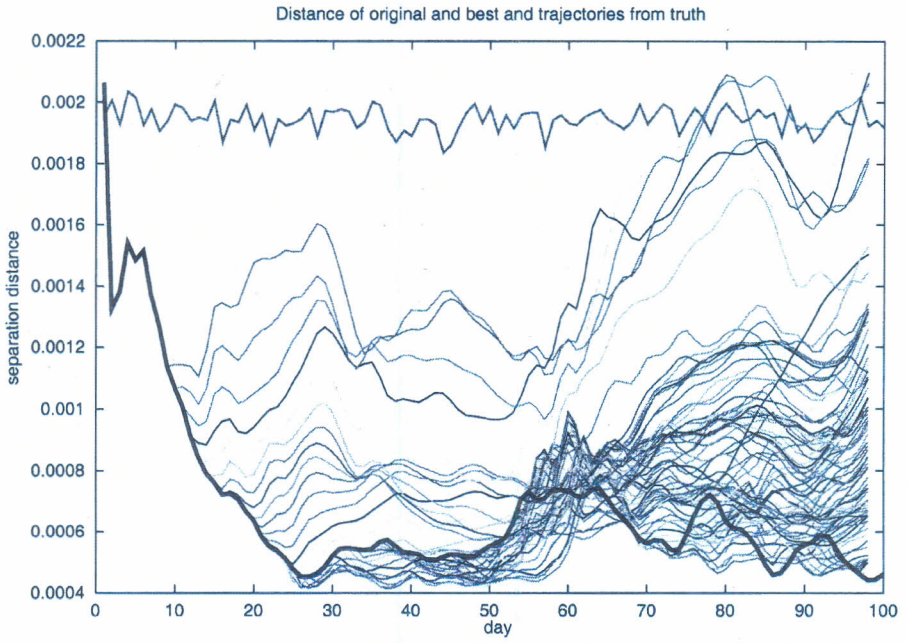


Fig. 4. The trajectories of each state of the best shadowing pseudo-orbit have been computed and the separation $\|\hat{x}_i - x_i\|$ of these trajectories from the true trajectory $\hat{x}_i$ are shown here. The separation distances of the initial observed sequence and best shadowing pseudo-orbit is replotted from Fig. 3 with thick lines for reference.

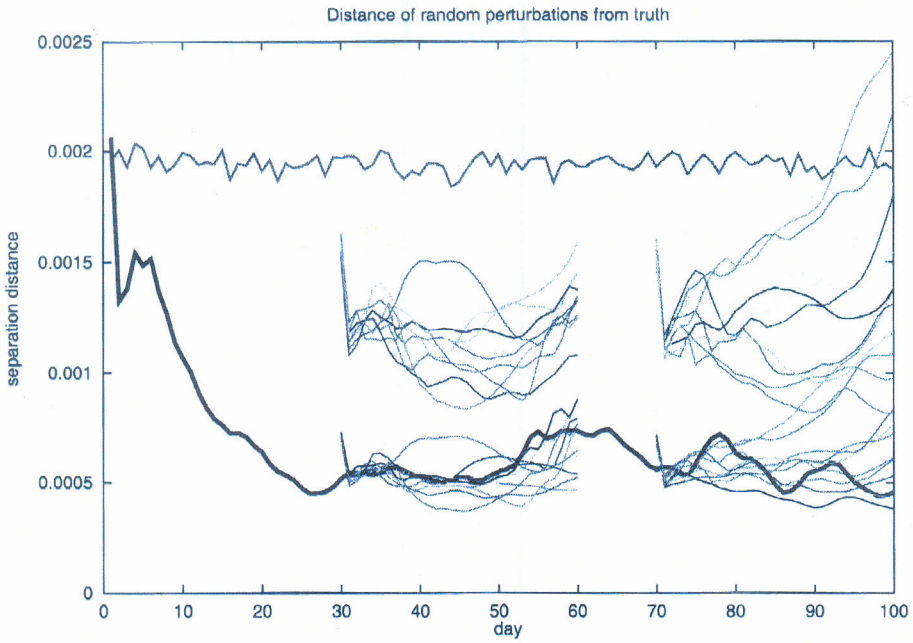


Fig. 5. Separation distance $\|\hat{x}_i - x_i\|$ between true trajectory $\hat{x}_i$ and the trajectories of random Gaussian perturbations of true states at $t = 30$ and 70. There are 10 perturbations of $\sigma = 4 \times 10^{-6}$ and $8 \times 10^{-6}$. The graph includes the information from Fig. 3 for reference. Observe that the states in Fig. 4 provide slightly better shadowing. See text for more discussion.

pseudo-orbit (for $t > 8$), and plot the separation distance of these trajectories from truth, the separation distance of the initial observations and the best pseudo-orbit are replotted from Fig. 3 for reference. We observe that the trajectories tend to maintain a similar distance from truth as the best shadowing pseudo-orbit. However, there is a period, $50 < t < 65$, where the trajectories appear to diverge rapidly from truth. Fig. 4 illustrates that although the states that make up the best pseudo-orbit are not a shadowing trajectory, the trajectories that begin at each of these states are good shadowing trajectories. For Fig. 4 all states with $t > 15$ shadow to within the observational error, and all states with $t > 20$ shadow to within 0.001, that is, half the observational error. Note that the region between $t = 50$ and 60 where there is notable increase in separation also correlates with the period of maximum indeterminism; this is discussed later in reference of Figs. 6 and 7.

Fig. 5 investigates the shadowing of a few random states close to the true trajectory; these provide a rough gauge of how good the best pseudo-orbit is. The random states are generated by a random Gaussian perturbation of two true states, $t = 30$ and 70. Randomly generated states like these tend to be "unphysical" and have a short relaxation phase that brings the state closer to the manifold of "physical" states. For the system and perturbations studied the relaxation takes less than 1 day and results in a reduction of the separation of these states from truth. Two levels of perturbations were used. One level of perturbation was chosen to give states after relaxation with approximately the same separation from truth as the corresponding states of the best pseudo-orbit, and the second level of perturbation (twice the first) gives states after relaxation that have a distance from truth a little less than half way between that of the pseudo-orbit states and the observations. We observe that the first set of random states have trajectories that shadow approximately as well as the corresponding best pseudo-orbit state, although the pseudo-orbit state trajectories are marginally better, but this is difficult to see. The second set shadow considerably worse. It is interesting to note that some of the random states have trajectories that are getting closer to truth, at

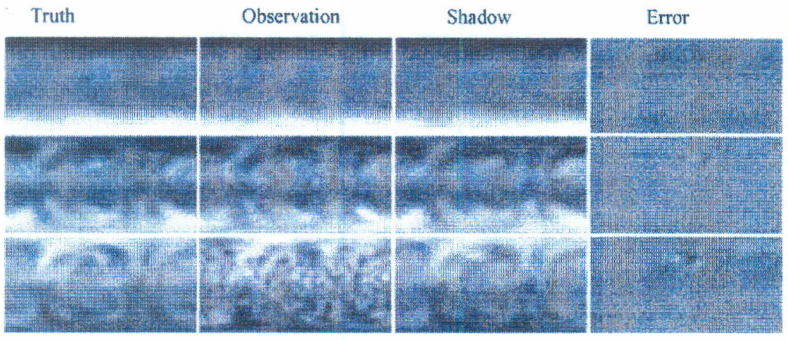


Fig. 6. Views of the stream function at day 62. In each column there are three panels corresponding to the three layers of the atmosphere, upper, middle and lower. The columns correspond to the true state, the observation on that day, the state of the shadowing pseudo-orbit, and the difference between the truth and the shadowing state. Day 62 was where the shadowing pseudo-orbit had maximum error. The error is spatially localized and results from incorrect phase of a mountain lee wave. Note that the observation error in the experiment was uniform in the spectral components, which translates into grid-point space as spatial correlated noise that is stronger in the lower layer.

least for a short period. In conclusion, the best pseudo-orbit states provide trajectories at least as good as could be expected from states as far as they are from truth. As a point of reference, the trajectories of the observations either shadow at about the same distance or diverge.

It has already been noted that there appears to be a correlation between the period of maximum indeterminism in the best pseudo-orbit and where states of the best pseudo-orbit provide the shortest shadowing times and most rapid divergence from truth. Although the results of only one set of observations has been display here, we find that different noise realizations give almost identical results, right down to the location and duration of difficult to shadow periods. This suggests that periods that appear to be difficult to shadow are not statistical accidents, but related to dynamical features of the flow. It has also been found that difficult to shadow periods can persist for 50 or more days, for example, such a 50-day period occurs during the 100-day continuation of the trajectory analyzed here. A detailed investigation of what causes difficult to shadow periods has not been done. Figs. 6 and 7 show at day 62 and day 90 the stream functions in grid-point space of the true state, observations, best pseudo-orbit, and the difference between the pseudo-orbit and the true state. Day 62 is at the peak of the poorly shadowed period, where as day 90 is a typical well-shadowed period. It is seen on day 62 that errors are very localized and they can be shown to be related to a lee wave of a topographic feature. It appears that at day 62 the pseudo-orbit's lee wave is out of

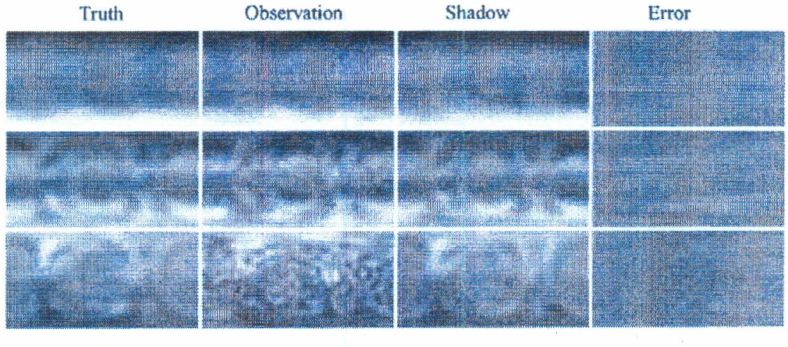


Fig. 7. Same as Fig. 6 at day 90. Day 90 has more typical errors. There are no very strong errors, but a number of weak localized errors.

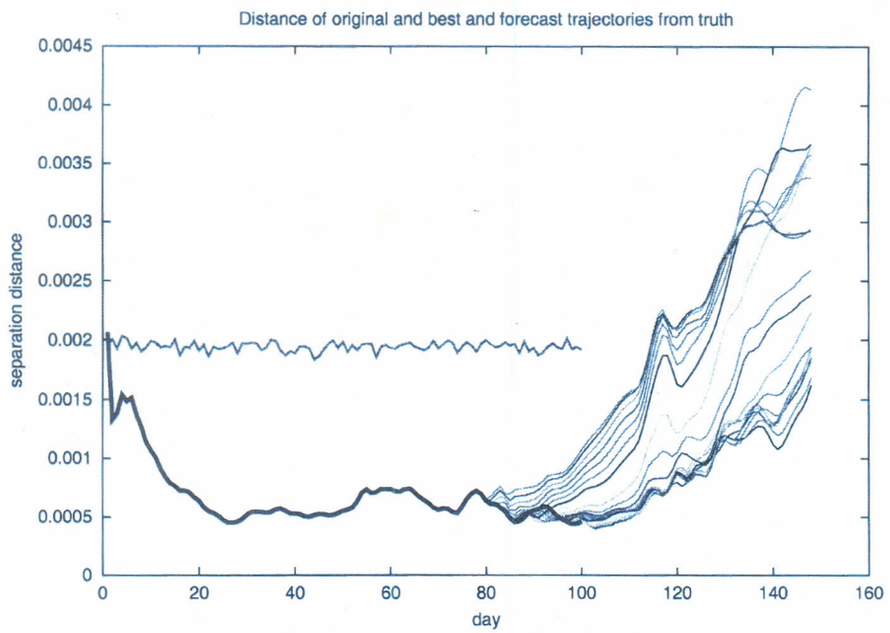


Fig. 8. The separation distance as in Fig. 4 for the trajectories initialized with the states in the last 20 days of the best shadowing pseudo-orbit and extending 50 days beyond the observed data used to construct the shadowing pseudo-orbit. The forecasts are consistent with observation up to observational error to 50 days from the initialization day.

phase with the true lee wave. Viewing an animation over several days of representations like Figs. 6 and 7, shows that in the days preceding day 62 the pseudo-orbit develops the lee wave much sooner and more strongly than the true trajectory. It might be that the uncertainty in the development of the lee wave is related to the indistinguishable states, or it might be a artifact of interaction between system dynamics and the crude descent method; a study of this phenomenon will be required.

Finally one is interested in forecasting the future, not just shadowing the past. The hope is, of course, that the states of a shadowing trajectory ought to provide good forecasts. Fig. 8 shows how the separation from truth grows when states for the last 20 days of the best shadowing pseudo-orbit are used as initial conditions for trajectories that are extended 50 days beyond the observed data. It is seen that in this case all the states provide forecasts that are consistent with observations out to 50 days from the day initialized. Comparing with Fig. 4 it is seen that this shadowing of the future is not as easy as the shadowing of the past. As mentioned above, for this particular trajectory, there is a difficult to shadow period, that just happens to begin around day 110 and extends for around 40 days. This should be compared with a similar shorter difficult to shadow period seen in Fig. 4 between days 50 and 65, for example, note how in both cases all the trajectories diverge as group at a similar rate.

## 6. Discussion

It has been argued that the gradient descent method of searching for shadowing trajectories by minimizing an indeterminism function can be modified to produce descent methods that require only limited derivative information. Indeed it seems that in a simple QG model that using no gradient information can obtain useful shadowing pseudo-orbits each of whose states can provide useful shadowing trajectories.

Clearly much work remains to be done and clearly many improvements are possible on the very crude schemes demonstrated. It should be considered whether adaptive adjustment of $\lambda$ during the descent in the $\lambda I$-substitution method is useful and whether varying $\lambda$ along the trajectory can improve results. It should be investigated whether estimation of the adjoint or rank reduced adjoints are useful. It is perhaps computationally efficient when the adjoint can be computed to initially use the $\lambda I$-substitution method then later switch over to a full adjoint gradient descent, in which case it should be studied what the optimal change-over point is.

It should also be investigated how shadowing algorithms can be used operationally. It is clearly inefficient to run the descent from zero whenever new daily observations arrive, it would be more efficient to use a "moving window" that uses the present estimate of the pseudo-orbit and the new observations.

As this paper goes to press experiments are being performed with an operational forecast model at reduced resolution and real 3DVAR analyses, using no adjoint and an available dry adjoint as an approximation to an unavailable full moist adjoint. This work is showing interesting and promising results and will be reported shortly.

## Acknowledgements

This work has been partially supported by a University of Western Australia small grant, a European Union grant EVK2-CT-2001-50012, and an Office of Naval Research DRI grant N00014-99-1-0056.

## References

[1] M. Davies, Noise reduction schemes for chaotic time series, Physica D 79 (1994) 174–192.
[2] J.D. Farmer, J.J. Sidorowich, Optimal shadowing and noise reduction, Physica D 47 (1991) 373–392.
[3] I. Gilmour, Nonlinear model evaluation: iota-shadowing, probabilistic prediction and weather forecasting, Ph.D. Thesis, Mathematical Institute, Oxford University, 1998.
[4] C. Grebogi, S.M. Hammel, J.A. Yorke, T. Sauer, Shadowing of physical trajectories in chaotic dynamics: containment and refinement, Phys. Rev. Lett. 65 (1990) 1527–1530.
[5] S.M. Hammel, A noise-reduction method for chaotic systems, Phys. Lett. A 148 (1990) 421–428.
[6] S.M. Hammel, J.A. Yorke, C. Grebogi, Do numerical orbits of chaotic dynamical processes represent true orbits?, Complexity 3 (1987) 136–145.
[7] S.M. Hammel, J.A. Yorke, C. Grebogi, Numerical orbits of chaotic processes represent true orbits, Bull. Am. Math. Soc. 19 (1987) (New Series).
[8] K. Judd, Nonlinear state estimation, indistinguishable states and the extended Kalman filter, Physica D 183 (2003) 273–281.
[9] K. Judd, L.A. Smith, Indistinguishable states. I. Perfect model scenario, Physica D 151 (2001) 125–141.
[10] K. Judd, L.A. Smith, Indistinguishable states. II. Imperfect model scenarios, Physica D, submitted for publication.
[11] E.J. Kostelich, T. Schreiber, Noise reduction in chaotic time-series data: a survey of common methods, Phys. Rev. E 48 (1993) 1752.
[12] D. Ridout, K. Judd, Convergence properties of gradient descent noise reduction, Physica D 165 (2001) 27–48.
[13] D.M. Walker, A.I. Mees, Noise reduction of chaotic systems by Kalman filtering and by shadowing, Int. J. Bifurc. Chaos 7 (3) (1997) 769–779.
[14] A. Weisheimer, M.V. Kurgansky, K. Dethloff, D. Handorf, Extratropical low-frequency variability in a three-level quasi-geostrophic atmospheric model with different spectral resolution, J. Geophys. Res. 108 (2003) 4171.
[15] L.A. Smith, Disentangling uncertainty and error: on the predictability of nonlinear systems, in: A.I. Mees (Ed.), Nonlinear Dynamics and Statistics, Birkhauser, Boston, 2000, pp. 31–64.
[16] L.M. Berliner, Statistics, probability and chaos, Stat. Sci. 7 (1992) 69–122.