The Maintenance of Uncertainty

Leonard A. Smith[a]
Mathematical Institute, University of Oxford,
Oxford, OX1 3LB, U.K.

October 14, 1,,,

Abstract

It is important to remain uncertain, of observation, model and law.

For the Fermi Summer School, Criticisms Requested

[a]email : lenny@maths.ox.ac.uk,

Contents

1 Introduction

All theorems are true[1]. All models are wrong[2]. And all data are inaccurate. What are we to do?

We must be sure to remain uncertain. In 1901, the year of Enrico Fermi's birth, it was well known that the sun could be only a few years old, inasmuch as a back of the envelope calculation showed that even if the sun were made of the highest quality coal, its chemical energy and gravitational energy would both be exhausted well before the time-scales claimed by geologists. Newton's Laws had successfully prophesied the existence of Neptune from irregularities in Uranus's orbit, and the planet Vulcan had been observed (between Mercury and the sun) which might explain irregularities in Mercury's orbit. While Neptune is still with us, Vulcan was, perhaps, a misinterpreted sunspot. Throughout Fermi's lifetime, astrophysical phenomena and physical experiments, often by his hand, repeatedly did things which could not happen, at least according to the "Laws of Physics" of the day. An unshakable belief in the applicability of those laws would have made progress impossible.

What has this to do with nonlinear dynamics and the analysis of time series? Nonlinear time-series analysis often resembles an experimental science: some technique is applied to a data set, an interesting observation is made, and a discussion ensues as to whether or not the observation is sound. Are we following Le Verrier in naming Vulcan in the hope of bringing Mercury's orbit closer into agreement with Newton's Laws, or are we following him in discovering Neptune? Is the uncertainty in the available data? or in our current understanding of the Physics? In these lectures we will examine methods which aim to maintain our uncertainty rather than adopt unsubstantiated conclusions. Applications range from testing the reliability of algorithm by analysing data of known origin, to propagating uncertainty in an initial condition under forecast models in order to examine the reliability of a particular forecast.

Nonlinearity plays a central role in data analysis, modelling, and predicting physical systems. We are often faced with questions like:

- Are these two signals related?
- Is there a deterministic/periodic component in this signal?
- Did this data set originate from a strange attractor?
- Is this system chaotic?
- What is the "limit of predictability" of this system?

[1] Regrettably, their premises are never fulfilled in reality.

[2] Unless they are "perfect," in which case they are theorems[1].

- Which is the better model for this system?

Our goal will be to examine the feasibility of answering these questions, rather than to demonstrate the current crop of algorithms for doing so.

Figure 1 shows two data sets with "similar dynamics." Is there a causal connection between these two series? Most likely not. Is there a statistically significant relationship between just these two series? For almost any simple null hypothesis: yes. In these lectures, we will examine methods which attempt to quantify the significance of a variety of data analysis techniques in the context of nonlinear, perhaps chaotic, phenomena. There are limits, of course, to our ability to determine whether or not a given observation is significant. Sometimes we simply must require more data. The important thing is to remain uncertain!

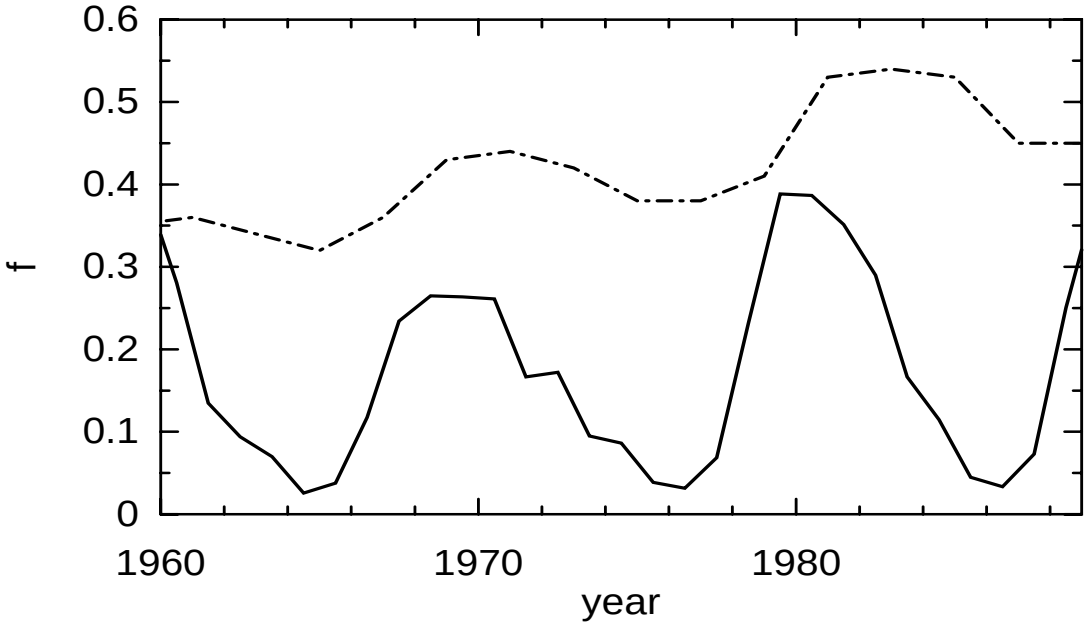

Figure 1: Simultaneous series of sunspot number (solid) and the fraction of the U.S. Senators who were Republicans on the day of their election (dot-dashed), from 1960 to 1989. The sunspot number has been rescaled by a constant.

One useful role for simple models is to help us maintain our uncertainty in the light of "promising" results. The historical record of sun-spots is one of the most studied time-series, and we will draw heavily from the work of Spiegel and Wolf [1], Weiss [2], and Casdagli *et al.* [3]. A wide ranging report on the relationship between sunspots and a variety of phenomena can be found in Stetson [4], which includes a number of interesting (then) out-of-sample forecasts. Figure 2a shows the sunspot record while Figure 2b is a particular sample from the stochastic sunspot simulation of Barnes *et al.* [5], which will be described in Section 4.1. How can we use this model to inform our uncertainty? Figure 3a shows a three-dimensional reconstruction of the sunspot data, produced with

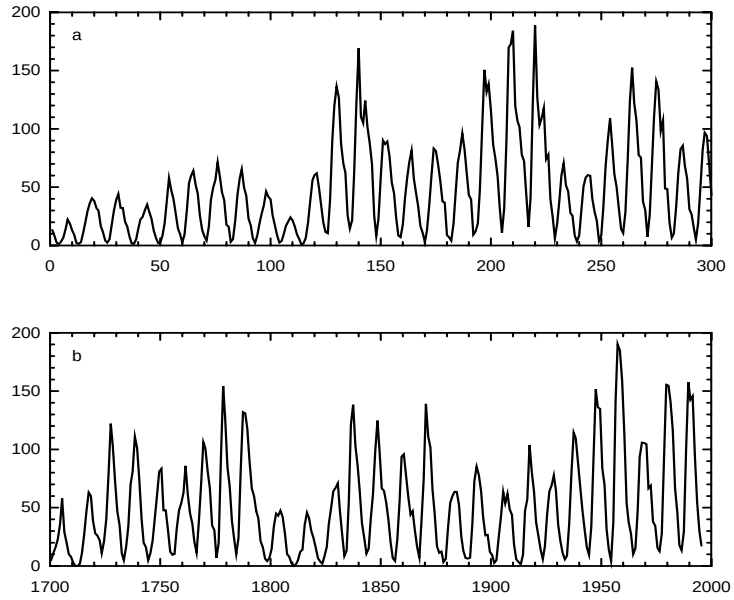

Figure 2: Time-series of (a) annual sunspot number and (b) data from the stochastic Barnes model of Section 4.1. Up-to-date sunspot data may be found at `http://www.oma.be/KSB-ORB/SIDC/sidc_txt.html`.

the techniques of Singular Spectrum Analysis (SSA), which is discussed in references [6, 7, 8] and the contribution of Ghil and Taricco to this volume. It has been observed that this view is reminiscent of the chaotic Rössler attractor: Is this observation evidence that the dynamics of sunspots are low dimensional deterministic chaos? To try to find out, we may, for example, repeat the experiment with data from the Barnes model, which we know (by construction) is stochastic and hence does not display deterministic chaos. The result is shown in Figure 3b, where again we recover structure reminiscent of the Rössler attractor. We conclude that such structure will occur in the analysis of any data set that "looks like" those of Figures 2, whether they arise either from a stochastic or from a deterministic processes; hence this observation provides little additional information on the dynamical process governing sun-spots. Our uncertainty is maintained.

In the following section, we introduce the basic framework for nonlinear dynamics. Dimensions and Lyapunov exponents are introduced and it is proven, by example, that chaos need not be difficult to predict. By chaos I shall mean deterministic chaos. A dynamical system is deterministic in the sense of Laplace when the future trajectory of the system is completely determined by the exact initial condition and the equations of motion. If the *effective* growth-rate of infinitesimal uncertainties is exponential in time, such a system is chaotic. This exponential-average-growth is reflected by positive Lyapunov exponents,but as illustrated in Section 2.2.1, positive Lyapunov exponents *per se* place no prac-

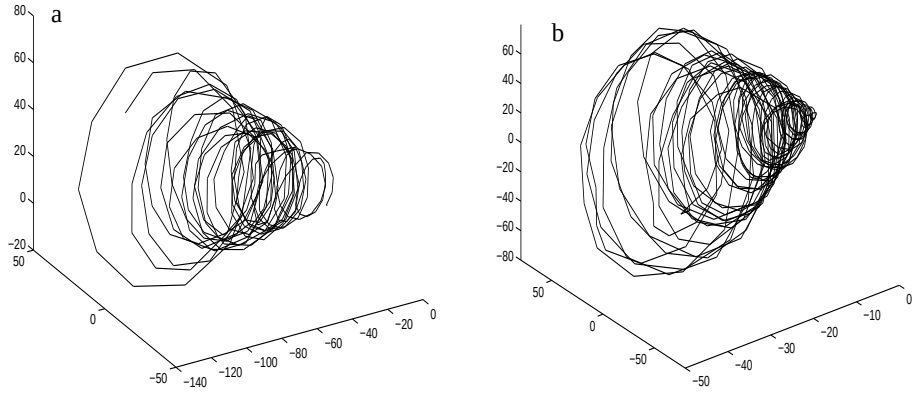

Figure 3: Three-dimensional trajectories from the SSA principle components of (a) the observed sunspot record and (b) a 512 year series from stochastic Barnes Sunspot model. In both cases, the view suggests structure similar to that of the Rössler attractor, even though the underlying process generating (b) is known to be stochastic.

tical limits on predictability. Takens' Theorem is stated in Section 2.4, and the encouragement it provides for methods of reconstructing dynamics from data is discussed.

Section 3 contrasts the various meanings of "prediction." In these lectures we are primarily concerned with forecasts either from data-driven models or from full simulations; the difference between developing the best model and extracting the best forecasts from a given model are explored, as are the extreme limitations of employing least square error criteria to *define* the best model. The initial condition is a different beast from an observation of the initial condition: observational data are never exact. Given the true initial condition of a chaotic system, the probability of an event is either zero or one. Determinism yields uniqueness. But given only an (inexact) observation, this probability may take on other values, even if our model is perfect. For this reason we are encouraged to make probabilistic forecasts even given good models of deterministic systems. If our models are not so good, the situation is even more interesting. Ensemble forecasts for perfect models, laboratory systems, and the Earth's atmosphere are discussed in Sections 3.5, 3.7 and 7.3 respectively. The relevant probability distribution functions (PDFs) often display complicated non-Gaussian structure. This makes model evaluation less trivial than taking the model with the least squared prediction error. Alternatives are discussed in Section 5.

In practice, it is often the case either that we do not understand the underlying physics of a system well enough to build first-principles models, or that such models would be too complex to be deployed. If we are lucky enough to have

a great deal of data from such a system, the techniques of Section 3 can be used to reconstruct its dynamics directly from the data. But how can we know that we have "a great deal of data" ? Section 4.1 begins with the presentation of tests for data sufficiency and the robustness of scaling exponents estimates, and concludes by suggesting tests for the self-consistency of dynamical models.

There are many applications we pass over without comment, and the nonlinear filtering of signals [9, 10] was almost one of these. The study of nonlinear systems, like the systems themselves, has too many interesting degrees of freedom. It is important to keep the driving question in sight, and distinguish between the distinct goals of studying a phenomena, testing an algorithm, analysing a data set and making the best forecast given the current state of the art(s).

For those who read only introductions while scanning figure captions, the gist of these lectures are (1) that statistics play an important role in helping us recognise the shortcomings of data analysis, and a dubious role in locating strengths; (2) that algorithms should be tested to destruction, so that at least some of their weaknesses are learned; (3) that tests of self-consistency are more accessible than tests of absolute truth, which is unsurprising if we consider even the "Laws of Physics" as the analogies of physics while we probe their limitations; and (4) that truly deep insights can only be supported by data not considered in the analysis. Until such data are obtained, we must remain uncertain, if hopeful. Regardless of the level of statistical skill and physical insight at hand, and regardless of the high level of statistical significance at which, for example, two data sets can be shown unlikely to be unrelated, promising results often evaporate given a glimpse of out-of-sample data. In the case of sunspots and the number of Republicans in the Senate, additional data can be obtained; contrast Figure 1 and Figure 22. As noted by Robert Boyle in the quotation that introduces Section 4, this was the case 300 years ago. And it will most likely be the case 300 years hence. Yet we may hope to reduce, in both magnitude and number, the disappointment of our expectations through the careful maintenance of our uncertainty.

2 Preliminaries

There is a fundamental difference between the physical processes which generate phenomena, the data recorded by measurement, and the models we construct to explore, explain and simulate the phenomena. Yet it is easy to confuse the map with the territory, especially since it is common to assign the role of the process to a particular model and then study the data it generates. In this **perfect model scenario**, a perfect model does indeed exist (the one which generated the data in the first place) and improving a given model of the correct functional form is equivalent to determining the original parameters. With physical phenomena, we never have access to a perfect model, nor are we privy to whether or not such a model exists.

Imperfect models are, of course, of tremendous value in suggesting new observations, evaluating our analysis techniques and testing our algorithms. When analysing data from a model, we know *a priori* whether the model is stochastic or deterministic. Sometimes, we may even know whether it is chaotic. We can use the model as a straw man, test our algorithm on data generated by it, and thereby determine whether or not the algorithm works (in this case), estimate the amount of data it requires, and develop tests of internal consistency for the analysis.

If a process contains a random element, then it is a **stochastic process**. The simplest stochastic process is a series of independent and identically distributed (**IID**) random variables. A wide class of autocorrelated stochastic processes can be developed by including autoregressive or moving average terms (see, for example, Chatfield [11]) and their nonlinear generalisations (see Tong [12]). By definition, chaotic processes are deterministic. Yet many interesting nonlinear processes contain elements of both chaotic and random nature, those that do not fall into the linear classification scheme of traditional statistics might be called **aleatoric** [13], since they contain both complex deterministic elements and random elements, like a human hand throwing dice. The SEQUIN model of Borland [14] suggests a framework within which to build explicit models of aleatoric dynamics, while the RAP approach of Paparella *et al.* [15] suggests a purely data-based alternative. Given only a finite data set, we can never determine whether the generating process was deterministic or stochastic - or even know whether or not the data series forms part of a long periodic orbit[3].

[3]Like the ones all digital computer experiments tend to (inasmuch as these computers are all finite state machines).

2.1 State-space Dynamics

We often envision a deterministic dynamical system as a set of m autonomous nonlinear ordinary differential equations (ODE)

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}) \tag{1}$$

The components of the vector $\mathbf{x}$ define the state space variable (*e.g.* position, velocity, acceleration, ...) in an m dimensional state space. The flow $\mathbf{f} : \mathbf{R}^m \to \mathbf{R}^m$ determines how these variables change with time. Given an initial condition $\mathbf{x}_0$, equation 1 defines the future trajectory $\mathbf{x}(t)$ for all time.

The extent to which any physical system truly corresponds to this vision is an open question, but we are free to define a dynamical system in this way and see where it leads us. The Moore-Spiegel [16] system evolves in a 3–dimensional space

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= z \\ \frac{dz}{dt} &= -z - (T - R + Rx^2)y - Tx \end{aligned} \tag{2}$$

providing a model for the height x of a parcel of ionised gas in the atmosphere of a star, where the parcel's velocity is y and its acceleration is z. In symbols $m = 3$, $\mathbf{x} \in \mathbf{R}^3$, and $\mathbf{x} = (x, y, z)$. A survey of several chaotic flows in $\mathbf{R}^3$ with quadratic nonlinearities is provided by Sprott [17]. Of course, our models are not restricted to $m = 3$, although for $m > 3$ plotting orbits becomes difficult. A useful method of visualisation when $m = 4$ or 5 is to plot a surface of section: for some component s of $\mathbf{x}$, record the $m - 1$ other components whenever s decreases through some particular value, say $s = s_0$. This results in a set of vectors in $\mathbf{R}^{m-1}$, in this case we have take a section on a plane perpendicular to one of the coordinate axes, but other surfaces may suffice[4]. The left-hand panel of Figure 5 is a 2-D section of the attractor in Figure 4 for $x = 0$. Practical considerations in obtaining surfaces of section are discussed by Hénon [18].

Ideally, each point on a surface of section evolves into a new one: the dynamics resembles a map of $\mathbf{R}^{m-1} \to \mathbf{R}^{m-1}$. Since the Moore-Spiegel equations are deterministic, we know there exists some (unknown) function that maps each point on the section to a new point after a time $t(0, y, z)$. Rather than integrating the equations, or attempting to approximate a true surface of section, we can write down the equations for a discrete-time map directly and study

[4]If the system is subject to a periodic driving force, the stroboscopic observations with the driving period are often used to obtain a section.

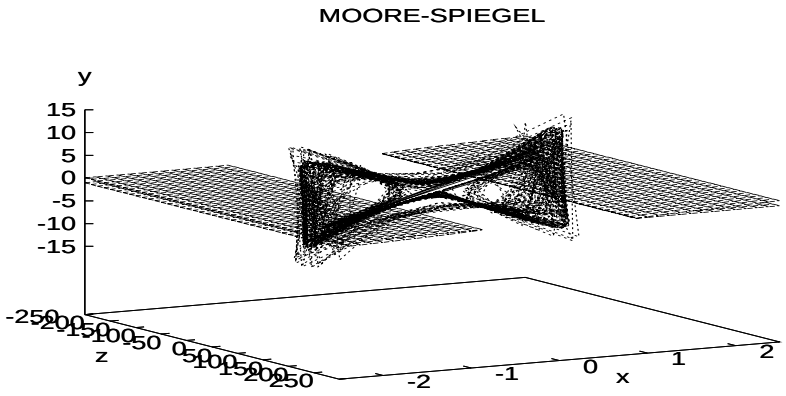

Figure 4: A strange attractor of the Moore-Spiegel system ($R = 100, T = 26$). The planar surfaces separate wedge-like regions, within these an uncertainty would eventually shrink with time, were the local Jacobian to remain relevant (see Section 2.1.2).

the dynamics of the resulting iterated dynamical system. In such maps, time takes on only integer values. Numerically, maps can be evaluated more quickly than surfaces of section. Maps, however, make less robust straw men than true surfaces of section: there need be no simple functional form for a true section and the return time in a map is "1" for all points, while the return time (if any!) on the cross-section of a flow may be of interest in itself.

One of the most studied 2-d maps is the Hénon Map[19]:

$$\begin{aligned} x_{i+1} &= 1 - ax_i^2 + y_i \\ y_{i+1} &= bx_i \end{aligned} \tag{3}$$

with $a = 1.4$ and $b = 0.3$ (see Figure 5b). Hénon constructed this map with simplicity in mind. It is the most general quadratic map with constant Jacobian determinant ($= -b$), which he considered a welcome property as the natural counterpart of the constant negative divergence in the Lorenz system[20]. The map is straightforward to manipulate analytically and thus of great value for analysis. These same properties make it a weak straw man.

A slightly more robust straw man is provided by the Stiletto Map:

$$x_{i+1} = (x_i + \frac{1}{a})e^{a(1-x_i)-1} - \frac{1}{a} + y_i$$

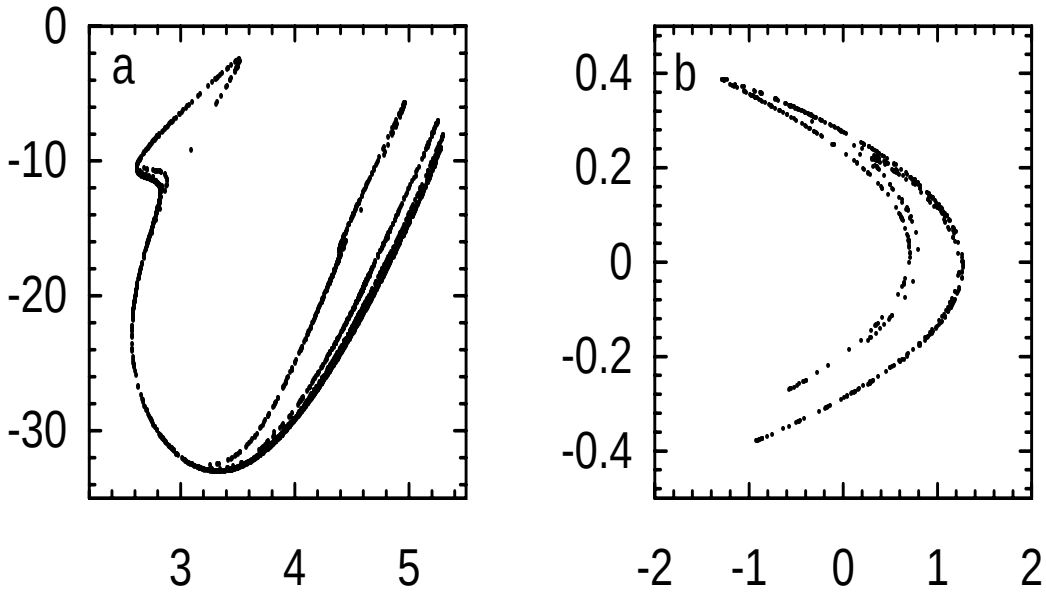

Figure 5: (a) An $x = 0$ surface of section of the Moore-Spiegel attractor from Figure 4, and (b) a trajectory from the Hénon attractor.

$$y_{i+1} = bx_i - y_i e^{-(x_i^2+y_i^2)/c^2} \tag{4}$$

which, for $c = 0$, generalises the 1-D Moran-Ricker Map in a manner analogous to the way in which the Hénon Map generalises the 1-D logistic map. An attractor of the Stiletto Map with parameters $a = 3.0$, $b = 0.3$, $c = 0.0$ is shown in Figure 6. While the functional form of equations 4 are much simpler than either the true surface of the Moore-Spiegel equations, or any map arising from real data, it may provide a bit more of a challenge to analysis techniques than the Hénon Map.

2.1.1 Linearized Dynamics of Infinitesimal Uncertainties

Given $\mathbf{f}$ and an initial condition $\mathbf{x_0}$, the trajectory $\mathbf{x}(t)$ is uniquely determined. But suppose there is an uncertainty, $\boldsymbol{\epsilon}_0 \in \mathbf{R}^m$, in the initial condition of a chaotic system. In this case, the future is uncertain even if we have a perfect model. For ordinary differential equations, the evolution of an *infinitesimal* uncertainty is governed by the linearization of the flow, that is

$$\dot{\boldsymbol{\epsilon}} = J(\mathbf{x})\boldsymbol{\epsilon} \tag{5}$$

where $J(\mathbf{x})$ is the Jacobian of the flow $\mathbf{f}$ at $\mathbf{x}$. For the Moore-Spiegel system

$$J(\mathbf{x}) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -(T+2Rxy) & -(T-R-Rx^2) & -1 \end{pmatrix} \tag{6}$$

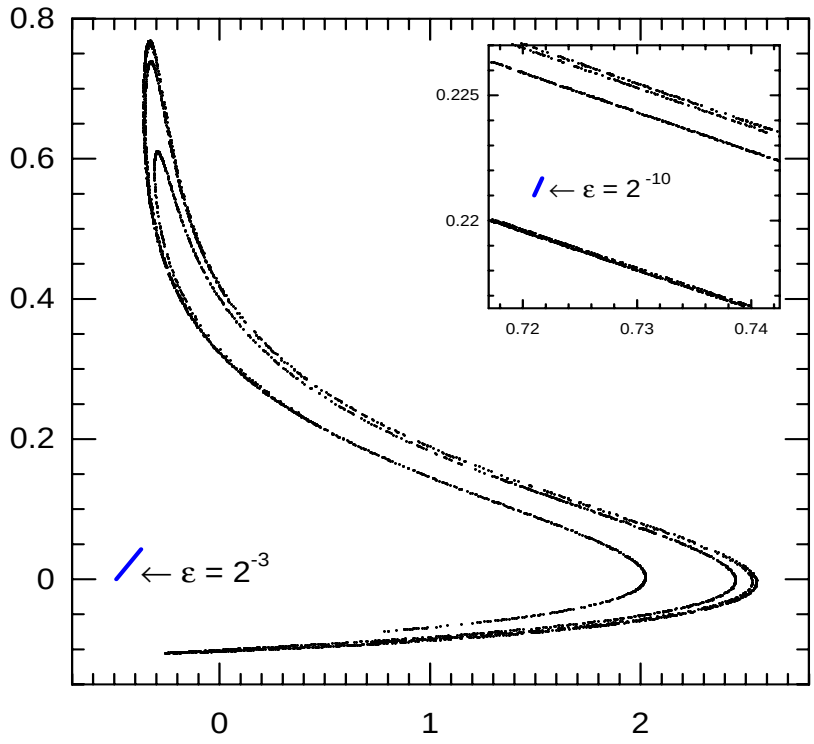

Figure 6: An attractor of the Stiletto Map ($a = 3.0, b = 0.3, c = 0$). The inset shows a blow-up of the region near a fixed point. The bars labelled $\epsilon = 2^{-3}$ and $\epsilon = 2^{-10}$ indicate two characteristic length-scales at which macroscopic "bands" of the attractor are separated, these length-scales are reflected in the correlation integral (see Section 2.3 and Figure 9).

This gives us a handle on the dynamics of infinitesimal uncertainties, which is discussed below along with their use in defining Lyapunov exponents. In Section 3.5 we will consider the dynamics of more realistic (finite) uncertainties.

2.1.2 Instantaneous Infinitesimal Dynamics

In a flow where the Jacobian is constant, the eigenvalues of the Jacobian matrix determine the long time behaviour of infinitesimal uncertainties; in particular if these eigenvalues have negative real parts, all perturbations will eventually die away. We can determine whether or not this is the case by evaluating the Routh-Hurwitz criterion for the Jacobian (see [21, 22]). The surfaces in Figure 4 show two thin wedges containing a fraction of the Moore-Spiegel attractor within which this is the case. Of course, a local Jacobian is only relevant for an instant in time, since the trajectory is advected past each point. Further, when a Jacobian matrix is not normal, there may be orientations in which an infinitesimal perturbation will grow for a finite time[23, 24]. Hence to apply the above criterion, we find ourselves in need of simultaneously assuming both the

limit $t \to 0$ and $t \to \infty$ as well! The resolution of this dilemma is to consider not the eigenvalues of the matrix J, but its singular values which are defined in the next section. As shown below, if the singular values of J are all less than one, then all infinitesimal uncertainties will decrease (instantaneously). If this result holds for a finite region of the state space, then all such perturbations will decrease for a finite time (as long as they remain within that region). There is no such region for this Moore-Spiegel attractor, however Ziehmann [25, 22] has shown that roughly 30% of the Lorenz attractor lies within such a region. Rather than introduce singular values here in the context of the Jacobian, we first lift the restriction to instantaneous dynamics by defining the linear propagator in the next section. We will still be restricted to infinitesimal uncertainties, but at least we will be able to consider their dynamics over a time interval $\Delta t > 0$.

2.1.3 Finite Time Evolution of Infinitesimal Uncertainties

The evolution of an infinitesimal uncertainty over a finite time Δt is determined by the linear propagator $M(\mathbf{x}_0, \Delta t)$ along the trajectory $\mathbf{x}(t)$, that is

$$\boldsymbol{\epsilon}(t_0 + \Delta t) = M(\mathbf{x}_0, \Delta t)\boldsymbol{\epsilon}(t_0) \tag{7}$$

where $\mathbf{x_0} \equiv \mathbf{x}(t_0)$ and, for a flow,

$$M(\mathbf{x_0}, \Delta t) = \exp\left(\int_{t_0}^{t_0+\Delta t} J(\mathbf{x}(t))dt\right). \tag{8}$$

For discrete time maps, the linear propagator is simply the product of the Jacobians along the trajectory

$$M(\mathbf{x_0}, k) = J(\mathbf{x}_{k-1})J(\mathbf{x}_{k-2})\ldots J(\mathbf{x}_1)J(\mathbf{x}_0). \tag{9}$$

This greatly simplifies the overhead implied by equation 8 ! And it is for this reason we use maps in the next subsection to dispute the relationship between Lyapunov exponents and limits to predictability.

$M(\mathbf{x}_0, \Delta t)$ evolves a spherical shell of radius $|\boldsymbol{\epsilon}_0|$ into an elliptical shell about $\mathbf{x}(t_0 + \Delta t)$, as illustrated in Figure 7. Of course for a finite uncertainty, M only approximates the true nonlinear dynamics. Figure 7 illustrates this point by showing the images of a circle under both the linear approximation and the full nonlinear function. For a fixed Δt, we can define a **maximum linear range** δ, such that the error in the linear approximation at $t = t_0 + \Delta t$ is less than some specific tolerance for any perturbation $\boldsymbol{\epsilon}$ with $|\boldsymbol{\epsilon}| < \delta$. δ will, of course, vary with $\mathbf{x}$.

The linear dynamics are most easily examined through the singular value decomposition (SVD) of M (see [21, 26]) which indicates the directions at t_0 that

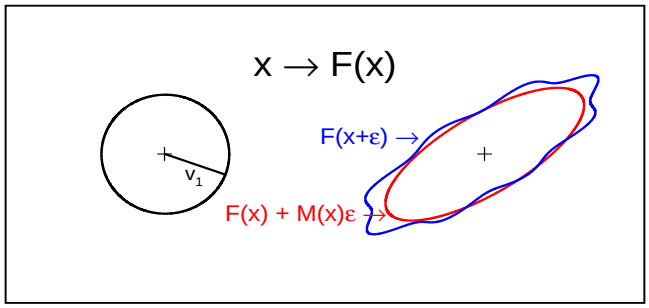

Figure 7: Schematic of the linear propagator. Under the linearized dynamics a circle of radius ϵ about the point $\mathbf{x}$ is mapped into the ellipse about $F(\mathbf{x})$. Of course for any finite radius this is only an approximation to the image of the circle under the nonlinear dynamics of F. The first left singular vector, $\mathbf{v}_1$ of the linear propagator, M, gives the direction which will be mapped into the semi-major axis of the ellipse while stretched by a factor σ_1.

will have evolved into the axes of the ellipse at $t_0+\Delta t$. In symbols, $M = U\Sigma V^T$ where the right (left) singular vectors, $\mathbf{v}_i$ ($\mathbf{u}_i$), form the columns of the orthogonal matrix V (U) and the entries of the diagonal matrix Σ are the singular values σ_i in rank order (*i.e.* with $\sigma_i > \sigma_j$ for $i < j$). With this ordering, the first singular vectors correspond to the direction which will have grown the most between t_0 and $t_0 + \Delta t$. While it is tempting to say $\mathbf{v}_1$ is the "fastest growing direction," it is also misleading since the SVD of M only indicates total change over the entire duration Δt and says nothing as to how this change came about. The right singular vectors (the $\mathbf{v}_i$) are often referred to as "initial time" singular vectors, and the left singular vectors (the $\mathbf{u}_i$) as "final time" singular vectors, inasmuch as $\mathbf{v}_i$ are defined at t_0 and evolve into the $\mathbf{u}_i$ which are defined at $t_0 + \Delta t$. Under the action of M, each right singular vector $\mathbf{v}_i$, is rotated into the corresponding left singular vector $\mathbf{u}_i$ and multiplied by the factor σ_i. That is

$$M\mathbf{v}_i = \sigma_i \mathbf{u}_i. \tag{10}$$

If all the singular values are less than one ($\sigma_1 < 1$ is sufficient, since none of the others are greater than σ_1) , then all infinitesimal perturbations will have shrunk at time $t_0 + \Delta t$. Negative numerical estimates for σ_1 in the Lorenz system were noted by Mukougawa *et al.* [27], and Nese [28] who examined variations in predictability discussed in [22].

For a given $\mathbf{x}$ and Δt, the σ_i define the **finite-time Lyapunov exponents**, $\lambda_i(\mathbf{x}, \Delta t) = \frac{1}{\Delta t} \log_2 \sigma_i$, first discussed by Lorenz [29][5] Since the σ_i are positive,

[5]Lorenz works out a numerical example in a 28 dimensional system and illustrates the variation of finite time exponents with initial state. Using a different nomenclature, of course, since this paper pre-dates that of Oseledec [30]. Additional discussion may be found in

we are free to write $\sigma_i = e^{\lambda_i \Delta t \ln 2}$ and define λ_i as an *effective* growth rate. An infinitesimal perturbation with orientation $\mathbf{v}_i$ is often said to grow as

$$\epsilon(\Delta t) \propto e^{\lambda_1 \Delta t} \tag{11}$$

For finite Δt this formulation can be misleading since in this equation λ_1 is a function of Δt (as is σ_1 and even $\mathbf{v}_1$). Thus Equation 11 holds for *any* positive definite function $\epsilon(\Delta t)$ and does not imply exponential growth.

In other words, while the *effective* rate would be informative if the growth of uncertainty were uniform and exponential, nothing needs be growing either uniformly or exponentially: effective rates remain well defined for each Δt as long the uncertainty remains non-zero and finite, irrespective of how the uncertainty actually increases with time. Interpreting them with the added assumption of uniform exponential growth is rarely justified in practice[6], although it is common. It is for this reason that we use the awkward phrase "average-exponential-growth rate." Further discussion of this point may be found in the introductory text of Nicolis [34].

The maximum average-exponential-growth rate over Δt corresponds to $\lambda_1(\mathbf{x}, \Delta t)$ and occurs when the initial (infinitesimal) uncertainty is aligned with $\mathbf{v}_1$. In the limit $\Delta t \to \infty$, $\lambda_1(\mathbf{x}, \Delta t)$ approaches the largest **global Lyapunov-exponent**, Λ_1, for almost all $\mathbf{x}$ and almost all initial orientations $\boldsymbol{\epsilon}_0$. We will define the local orientation of the **global Lyapunov vector** (LV) at $\mathbf{x}$ as the orientation of $\mathbf{v}_1$ at $\mathbf{x}$ in the limit $\Delta t \to \infty$. The remainder of (global) Lyapunov-spectrum, $\Lambda_i, i = 2, \ldots, M$, quantifies the growth of perturbations in subspaces where the effective rate is less than Λ_1, and are defined in a similar way, assuming these limits exist (Oseledec [30]). Note that for finite Δt, one can write down differential equations for the singular vectors themselves (see Green and Kim [35]). Also note that in large numerical models, direct manipulation of M is less straightforward; alternative methods for approximating the SVD in large numerical models are investigated by Barkmeijer [36].

Do positive Lyapunov exponents place an *a priori* limit on predictability? It is often argued that, since an initial uncertainty grows exponentially *on average*, then the uncertainty will tend to double (increase by one bit) after the **Lyapunov time**

$$\tau_\Lambda = \frac{1}{\Lambda_1} \tag{12}$$

where Λ_1 is expressed in bits per second. The **uncertainty doubling time**, τ_{double} provides a more direct measure of a prediction time scale through the average of the minimum time required for an uncertainty to increase by a factor

[31, 32, 22, 33] and references therein.

[6] It is justified in dynamical systems with constant Jacobians, which are rare in practice!

of two. More generally, define the q-pling time $\tau_q(\mathbf{x}_0, \boldsymbol{\epsilon}_0)$ as the smallest time at which the initial uncertainty $\boldsymbol{\epsilon}_0$ about $\mathbf{x}_0$ has increased by a factor q

$$\tau_q(\mathbf{x}_0, \boldsymbol{\epsilon}_0) = \min_{t>0}\{t \Big| \ \|F_t(\mathbf{x} + \boldsymbol{\epsilon_0}) - F_t(\mathbf{x})\| \geq q\|\boldsymbol{\epsilon}_0\|\} \tag{13}$$

Whenever an initial orientation of $\boldsymbol{\epsilon}$ is well defined for each $\mathbf{x}$, we have

$$\tau_q(\|\boldsymbol{\epsilon}\|) = \Big\langle \tau_q(\mathbf{x}, \boldsymbol{\epsilon}) \Big\rangle_{\mathbf{x}} \tag{14}$$

where the average is taken over all points $\mathbf{x}$ on the attractor. To allow a fair comparison with Lyapunov exponents, we will define τ_{double} as the limit of $\tau_2(\|\boldsymbol{\epsilon}\|)$ as $\|\boldsymbol{\epsilon}\| \to 0$, with the orientation of the uncertainty at $\mathbf{x}$ defined in the local direction most relevant to the first global Lyapunov exponent, specifically the limiting $\mathbf{v}_1$ of the SVD of $M(\mathbf{x}, \Delta t)$ as $\Delta t \to +\infty$, assuming this limit exists.

Like the Lyapunov exponents, τ_{double} reflects only the dynamics of infinitesimals, but unlike the Lyapunov exponents it reflects the time taken directly, something an effective rate simply cannot do[7]. Both finite time Lyapunov exponents and global Lyapunov exponents reflect average rates, *not* average times; being effective rates, their definition requires a time-scale be defined *a priori* . In general the inverse of the arithmetic average of the inverse of a variable yields a poor estimate of the variable!

2.2 Lyapunov Exponents and Predictability

Since it is widely held that positive Lyapunov exponents place a fundamental limit on predictability, we will prove, by example, that this is simply false. The failure of Lyapunov exponents to reflect predictability is shown in a system where the linearized dynamics are exact even for finite uncertainties. Things will only be more interesting in physical systems.

2.2.1 The Baker's Apprentice Map

The Baker's Map, a traditional [37] two-dimensional map of the unit square, may be written as:

$$x_{i+1} = \begin{cases} \frac{1}{\alpha} x_i & \text{if} \quad 0 \leq x_i < \alpha \\ \\ \beta \ (x_i - \alpha) & \alpha \leq x_i < 1 \end{cases}$$

[7]Effective rates and q-pling-times are contrasted in [22].

(15)

$$y_{i+1} = \begin{cases} \alpha y_i & \text{if} \quad 0 \le x_i < \alpha \\ \\ \alpha + \frac{1}{\beta} y_i & \alpha \le x_i < 1 \end{cases}$$

with $\alpha = \frac{1}{2}$, $\beta = 2$. The name originates from the similarity between the dynamics under the map and the stretching and folding when kneading dough. The unit square is stretched horizontally by a factor of 2, and compressed vertically by the same amount, then cut in half and the rightmost portion is stacked on top of the left, resulting in an area preserving map. In this case, both the left and right first singular vectors always point in the horizontal direction for any Δt. For any initial uncertainty with this orientation $\frac{\epsilon_{i+1}}{\epsilon_i} = \frac{1}{\alpha} = \beta = 2$ for each i and

$$\Lambda_1 \quad = \quad \lim_{n\to\infty} \frac{1}{n} \log_2 (\prod_{i=0}^{n-1} 2) \quad = \quad \lim_{n\to\infty} \frac{1}{n} \log_2 (2^n) \quad = \quad 1 \text{ bit per second.} \quad (16)$$

Thus $\tau_\Lambda = 1$. Similarly, $\tau_{\text{double}} = 1$ for all uncertainties with this orientation. In this case, the Lyapunov time accurately reflects the limits of predictability. This follows from the uniform stretching throughout the state space.

The family of Baker's Apprentice Maps are defined as

$$x_{i+1} = \begin{cases} \frac{1}{\alpha} x_i & \text{if} \quad 0 \le x_i < \alpha \\ \\ \beta \ (x_i - \alpha) \mod 1 & \alpha \le x_i < 1 \end{cases}$$

(17)

$$y_{i+1} = \begin{cases} \alpha y_i & \text{if} \quad 0 \le x_i < \alpha \\ \\ \alpha + \frac{1}{\beta} (\lfloor \beta (x_i - \alpha) \rfloor + y_i) & \alpha \le x_i < 1 \end{cases}$$

where $\alpha = \frac{2^N - 1}{2^N}$, $\beta = 2^{2^N}$ and $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z. Different apprentice maps are defined for each value of N, $N = 1, 2, 3, \ldots$, and in each case a small fraction $(1-\alpha)$ of the "dough" is stretched a great deal (2^{2^N}), before being cut and stacked, while the majority of the initial conditions are displaced only slightly $(\frac{1}{\alpha})$.

Each Apprentice Map is area preserving, and all first singular vectors are again oriented in the horizontal from which we have

$$\Lambda_1 = \alpha(\log_2(1/\alpha) + (1-\alpha) \log_2 \beta \quad (18)$$

or, with the choices of α and β given above,

$$\Lambda_1 = 1 - \alpha \log_2 \alpha > 1 \ \text{ bit per iteration} \quad (19)$$

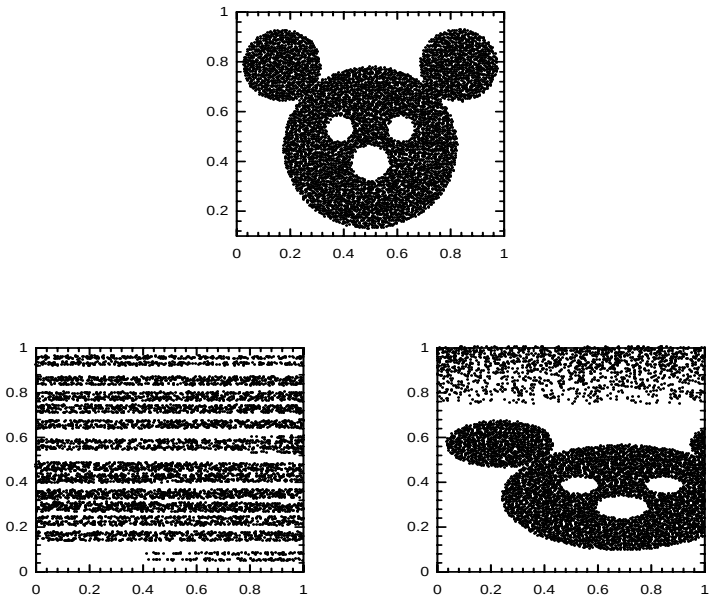

Figure 8: A ensemble of 2^{13} initial conditions (upper panel), its image after 4 iterations of the Baker's Map (lower left), and the image of the same initial ensemble after 4 iterations of the $N = 4$ Apprentice Map. Note that there are regions of relatively high predictability in the Apprentice Map, even though its Lyapunov exponent is larger than that of the Baker's Map.

Since $\alpha < 1$, $\log_2 \alpha < 0$ and $\Lambda_1 > 1$ for every N and the Lyapunov time of each of these maps is *less* than that of the Baker's Map. Yet for the majority of initial conditions, these maps are much more predictable than the Baker's Map; the difference is visible in Figure 8. The upper panel of Figure 8 shows an ensemble of initial conditions while lower panels provide the the fourth iterate of this initial ensemble either under the Baker's Map (left) or under the $N = 4$ Apprentice Map (right). The degree of detail remaining in the lower right frame illustrates the point that, for most initial conditions, the Apprentice Map is much more predictable than the Baker's Map.

These results are reflected in $\tau_{\rm double}$ for these maps. An infinitesimal uncertainty in the horizontal direction will grow on every iteration[8]. Even if the trajectory remains in the slow stretching region $x < \alpha$, the horizontal component of the initial uncertainty will double after j iterations with $j = \lceil -\frac{1}{\log_2(\alpha)} \rceil$ where $\lceil z \rceil$ is the smallest integer greater than or equal to z. In other words, j is the smallest integer such that $(\frac{1}{\alpha})^j \geq 2$. Summing the fraction of the initial conditions with

[8]Unlike the Lorenz system which has a region in which all uncertainties shrink, uncertainty always increases under the Baker's Apprentice Maps, it is only the rate of increase which varies.

Table 1: Baker's Apprentice Properties (approximate)

N	j	Λ_1	τ_{double}	τ_Λ
1	1	1.500	1.0000	0.666
2	3	1.311	2.3125	0.763
3	6	1.169	4.4096	0.856
4	11	1.087	8.1331	0.920
5	22	1.044	16.0850	0.958
6	45	1.022	32.4931	0.978
7	89	1.011	64.3126	0.989

each doubling time (from 1 to j) we have

$$\tau_{\text{double}} = \sum_{i=0}^{j-1} \alpha^i = \frac{1-\alpha^j}{1-\alpha} \tag{20}$$

Estimates of τ_{double} and τ_Λ for small N are given in Table 1. As $N \to \infty$, $\tau_{\text{double}} \to 2^{N-1}$, while $\tau_\Lambda \to 1$. Thus the large N chaotic maps are easily predicted (most of the time) regardless of the fact that the Lyapunov exponent of each is greater than that of the Baker's Map.

2.2.2 Infinitesimals and Predictability

As average rates, Lyapunov exponents need not reflect predictability. For this reason the τ_q better quantify the growth of uncertainty; yet if the τ_q are computed from infinitesimals, they too will require linear approximation to hold; the large-scale structure of the flow may either increase or decrease an infinitesimal "Limit of Predictability" limiting their usefulness. Worse yet, a global "Limit of Predictability" is only useful on time-scales over which uncertainty growth is relatively uniform. For the τ_q, this requires $\tau_{q^2} \approx 2\tau_q$ for q small enough that the linearization remains relevant. Even simple chaotic systems like the Moore-Spiegel system often fail this test for, say, $q = 64$, implying we must be able to lose more than 6 bits of information and still consider the uncertainty infinitesimal.

Realistic bounds to predictability must consider the nonlinear dynamics of finite uncertainties, a task to which we return with ensemble forecasts in Sections 3.5 and 7. As long as our uncertainty is infinitesimal, it can hardly pose a limit to predictability! And once it becomes finite, the linearization, and hence Lyapunov exponents are, in general, irrelevant to error growth.

2.3 Dimensions

Over the past decade, a great deal of effort has been devoted to determining the *number of active degrees of freedom* in a system by estimating dimensions. This effort is often justified by the observation that establishing low dimensional deterministic dynamics proves the existence of a model with only a few degrees of freedom. In practice, it may be easier to actually construct such a model from the data than it is to obtain a reliable dimension estimate directly. Some of the difficulties in obtaining good estimates are discussed below; in general, any estimate should be treated with suspicion when the remaining uncertainty in the estimate is not carefully discussed. The significance of dimension estimates may be investigated using surrogate data, which is discussed in Section 4.1.

Poincaré [38] provides a good discussion of what constitutes a "dimension," developing the intuitive idea of dimension inductively in terms of "cuts." The topological dimension of a set is one greater than the topological dimension of the set which can bound it. Here the cuts represent sets which might divide the set in question into disjoint pieces. Let isolated points have topological dimension zero. Any segment of a curve is isolated (bounded) when it is cut by two points, thus a line segment has dimension one. An area can be cut into isolated areas by a curve, and thus is assigned dimension two. And so on.

Introductions to dimension in the context of chaotic attractors are provided by Farmer *et al.* [39], Smith [40] and Theiler [41]. While an entire spectrum of dimensions, d_q, was introduced by Renyi 50 years ago (see Ruelle [42] and the many references therein), we will focus on the box-counting (fractal) dimension, d_0, and the correlation dimension, d_2. d_0 is most easily defined through the variation in the number of cubes required to cover the set as a function of the diameter of the cubes, the **correlation dimension** d_2 reflects how the probability that the distance between two randomly chosen points will be less than ℓ varies, as a function of ℓ.

Consider a "large" set of points distributed on a circle which itself is embedded in a 3-dimensional space; if we choose any point and ask, *for a sufficiently small distance* ℓ, how the number of points within a distance ℓ varies with ℓ, we will find $N \propto 2\ell^1$; if our "large" set is distributed uniformly on a circular disk, $N \propto \pi\ell^2$ as long as the chosen point is not too near the edge of the disk. These illustrations suggest that our intuitive notion of the dimension (1 for the circle, 2 for the disk) is reflected as a power law behaviour of $N(\ell)$. They also indicate obvious effects limiting us to small length scales, and therefore large data sets, even in these simple examples. The length scale must be small both to avoid macroscopic structure, due to curvature in the case of the circle, and to avoid edge-effects[9]. Macroscopic structure, such as sheets and folds, may have effects

[9]In line segment, for example, this is required to avoid the transition from growth proportional to $(2\ell)^1$ to growth proportional to ℓ^1 as ℓ increases through the distance between

at length-scales much smaller than the diameter of the set[43].

2.3.1 The Grassberger Procaccia Algorithm

The correlation dimension d_2 of an infinite set of points may be defined from the correlation integral:

$$C_2(\ell) = Probability(||\mathbf{x_i} - \mathbf{x_j}|| < \ell) \tag{21}$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are two randomly chosen points in the set. For any finite collection of points, there is some nearest pair distance, ℓ_{np}; $C_2(\ell) = 0$ for $\ell < \ell_{np}$, and $d_2 = 0$ as expected. Given an infinite set confined to a bounded region, we expect $C_2(\ell)$ to be scaling for small ℓ, that is

$$C_2(\ell) \sim \chi(\ell)\ \ell^{d_2} \tag{22}$$

where $\chi(\ell)$ accounts for both macroscopic effects at large scales and organised lacunarity effects, if any, at small scales. For the circle and the disk, $d_2 = 1$ and $d_2 = 2$, respectively, and $\chi(\ell)$ approaches a constant as $\ell \to 0$. If $\chi(\ell)$ is not a constant, and for a fractal it need not be, then Equation 22 does not restrict $C_2(\ell)$ to a power law any more than Equation 11 implies exponential growth.

At finite length scales, one can inspect the local slope of $\log_2 C_2(\ell)$ as a function of $\log_2(\ell)$ for a "scaling range" over which to estimate d_2. That is, we may estimate d_2 from a linear fit of

$$\log_2 C_2(\ell) = \log_2 \chi(\ell) + d_2 \log_2 \ell \tag{23}$$

with the assumption that $\chi(\ell)$ is either constant or rapidly oscillating over the range of the fit (which, to some extent, can be verified by examination). This is the approach of the Grassberger Procaccia Algorithm[44]. To obtain the correlation integral from a finite data set, we assume

$$C_2(\ell) \approx \frac{Number\ of\ pairs\ of\ points\ separated\ by\ less\ than\ \ell}{Total\ number\ of\ pairs\ of\ points} \tag{24}$$

for $\ell > \ell_{np}$. Thus

$$C_2(\ell) \approx \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta(\ell - |\mathbf{x_i} - \mathbf{x_j}|) \tag{25}$$

where close pairs are counted via the Heaviside function, $\Theta(x)$, which is equal to zero for negative argument and one otherwise.

a given point and the end of the line segment.

The Grassberger Procaccia Algorithm has become a widely used tool [45, 46]. The estimated correlation integral will be biased if small spatial separations between pairs of points simply reflect that they were observed close in time [47, 48]. In time-series analysis of a finite data set, $\mathbf{x_i}$ and $\mathbf{x_j}$ are *not* taken at random from an infinitely long trajectory, and $|i-j|$ indicates their separation in time if the sampling rate is constant. To limit the effect of points very nearby in time, one often restricts the sums over i and j so that $|i-j| > W$ for some constant W (see [47]). Yet even for $i >> j$, the distance between $\mathbf{x}_{i+1}$ and $\mathbf{x}_{j+1}$ is rarely independent of the distance separating $\mathbf{x_i}$ and $\mathbf{x_j}$; in a chaotic system, whenever one such pair is sufficiently close, the other pair will be close as well, giving rise to the "railroad tie" pairing of points often visible in numerically generated images of strange attractors. This is avoided in Figure 4 by not selecting the points to be plotted with a fixed sampling time. We return to these issues in section 2.3.3 after considering limitations which remain even when points randomly distributed on the attractor are available.

2.3.2 Towards a better estimate from Takens' Estimators

An alternative to determining the slope of the correlation integral has been suggested by Takens [49, 50]. Takens' estimator provides a maximum likelihood estimate, $T_2(\ell)$, of d_2 without estimating the slope of the correlation integral directly. We shall use the fact that the distribution of $T_2(\ell)$ is known to determine a coherent estimate both of d_2 and of the uncertainty in this estimate.

For a given separation ℓ, consider all pairs of points separated by a distance ℓ_{ij} less than ℓ and compute the average

$$\alpha = \langle log(\ell_{ij}/\ell) \rangle \tag{26}$$

The **Takens' estimate** for the correlation dimension is then $T_2(\ell) = -\frac{1}{\alpha}$. There remains the choice of ℓ, the implications of which is illustrated in Figure 9 which shows several independent $T_2(\ell)$ of the Stiletto Map. We will take the same fixed number of pairs of points for each curve $T_2(\ell)$; in Figure 9 each curve is based on 2^{33} pairs. For a given value of ℓ, the independent estimates $T_2(\ell)$ have a Gaussian distribution with mean equal to the Takens' estimate for d_2 at that length scale, $\overline{T_2}(\ell)$, and a standard deviation which increases due to sampling uncertainty as $\ell \to 0$. As usual, we want to consider an inconvenient limit: as $\ell \to 0$, $\overline{T_2}(\ell) \to d_2$ (neglecting lacunarity effects discussed below), while the standard deviation of the $T_2(\ell)$ increases in the same limit. Having an estimate of the distribution of $T_2(\ell)$ is of use: we can compute the largest value of ℓ for which $\overline{T_2}(\ell)$ is consistent with the observed distributions at all smaller length scales. This approach to dimension estimation automatically selects the best length-scale given the limitations of the data, and provides an

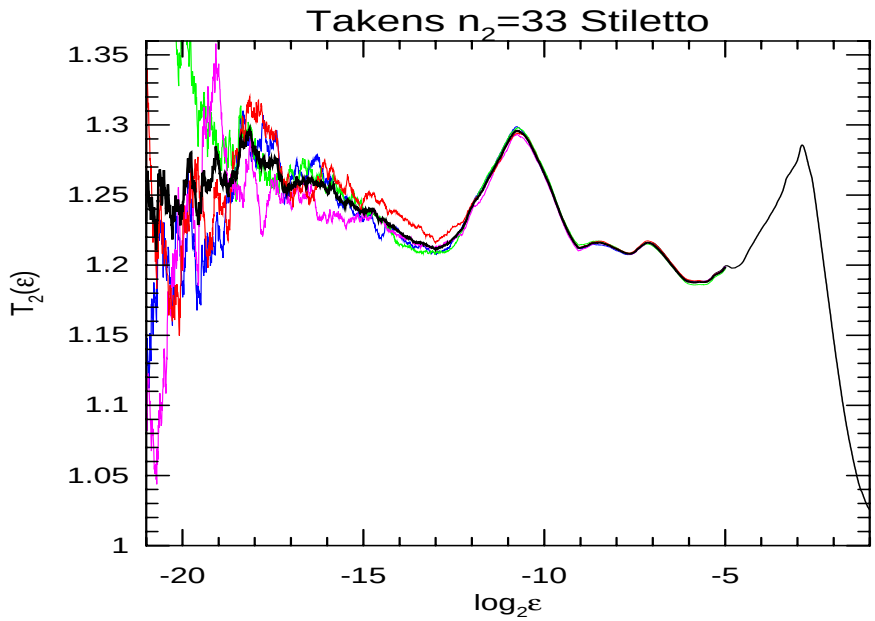

Figure 9: Takens' estimators for the Stiletto Map. Note the structure in $T_2(\epsilon)$ at $\epsilon \approx 2^{-3}$ and $\epsilon \approx 2^{-10}$, corresponding to the macroscopic structure visible in Figure 6.

uncertainty estimate as well. Of course, if the uncertainty estimate includes the embedding dimension, then it places no upper bound on the dimension of "the attractor". When the system is known there is, in principle, no fixed upper bound on the amount of data which can be considered; smaller and smaller length scales could be examined in search of an internally consistent estimate of a potentially limiting value. One could then attempt unbiased estimates, using each point only once! The curves in Figure 10 indicate a step in this direction. A trajectory is sampled randomly to form two large data-bases of sample of points; a base point is chosen at random from one of the databases, and its distances from a small number of points drawn from a the other database are recorded. The base point is then discarded. Both data-bases are continually updated by replacement so that there is no intrinsic limit to the number of independent pairs that can be considered.

Figure 10 shows that at large scales ($\log_2 \ell \approx -14$), the Takens' estimates $T_2(\ell)$ are all in close agreement. This allows a good estimate of $\overline{T_2}(\ell = 2^{-14})$, but we see from the systematic variation in the $T_2(\ell)$ as ℓ decreases that $\overline{T_2}(2^{-14})$ is an unreliable estimate of d_2. The mean of this distribution increases as ℓ decreases from 2^{-14} to 2^{-16}, making a precise estimate of $\overline{T_2}(2^{-14})$ irrelevant. At length scales $\ell < 2^{-16}$, the distribution spreads out in a manner not obviously inconsistent with $d_2 \approx \overline{T_2}(2^{-16}) \approx 1.250 \pm 0.003$. And using a t-test, it is is straightforward to determine whether or not a distribution at some smaller length scale is inconsistent with this result. Specifically, for a given confidence

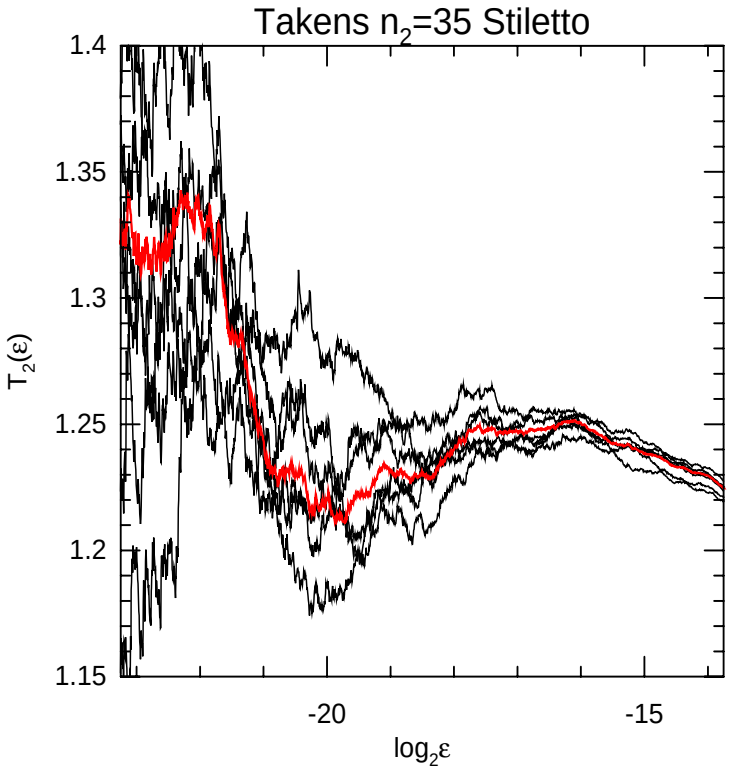

Figure 10: Six independent Takens' estimators at small scales for the Stiletto Map. Each $T_2(\epsilon)$ curve is based on 2^{35} two-point separations, no two points both being used twice. Independent data sets are used to compute each curve.

level[10] can we reject the null hypothesis that the observed distribution at a length scale of , say, $\ell \approx 2^{-22.0}$ (where the estimated mean and standard deviation are 1.31 ± 0.05), was drawn from a population with the same mean as the distribution at $\ell = 2^{-16}$? If so, we must accept the less precise (but potentially relevant) result at the smaller length scale.

2.3.3 Space Time Separation Diagrams

How can we know that we have enough data for reliable dimension estimates? In general we cannot, just as in Fourier Analysis a continuous spectrum can result from the analysis of a periodic signal if the data segment has a duration less than one full period. We can however, often determine with confidence that we do *not* have enough data. One simple test is to construct a **space-time-separation diagram** [51, 48].

[10]One must take care, of course, in computing significance levels since multiple tests of the same null hypothesis are being made and, independently, because each individual Takens' estimator is autocorrelated for similar values of ℓ.

Each pair of points on a trajectory is separated by a distance ℓ_{ij} in state space and a distance t_{ij} in time; by plotting a scatter diagram of the spatial separation against the temporal separation, we can determine a minimum time separation for which points from a trajectory might be considered independent. The scatter plot of Figure 11 shows the expected behaviour: pairs

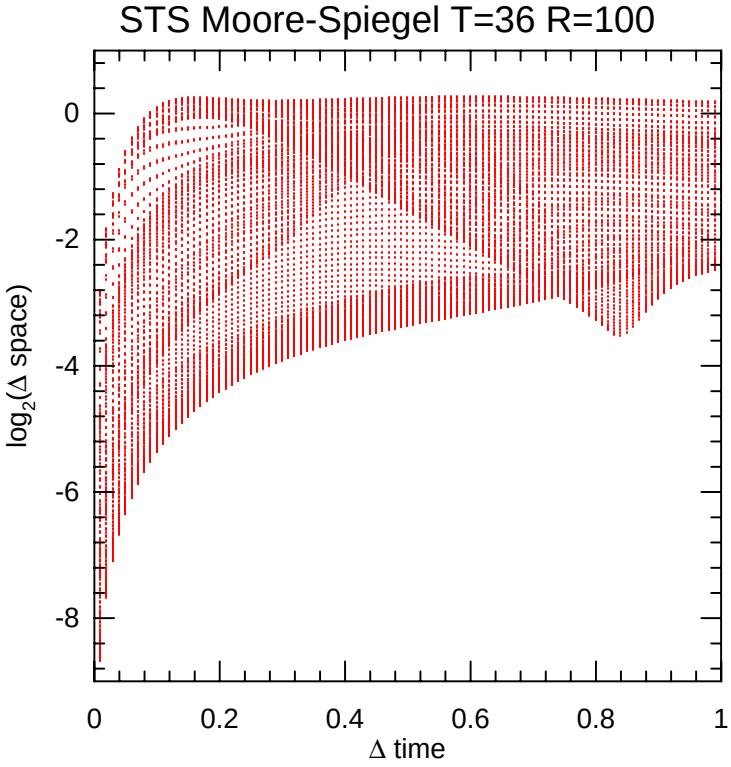

Figure 11: Space-time-separation scatter plot for the Moore-Spiegel system with $R = 100$ and $T = 36$. Each dot reflects the separations of one pair of points in state space, their separation in time in the horizontal and their separation in state space in the vertical.

which are very close in time have small separations in state space and initially the state space separation tends to increase with increasing time *along* a trajectory. The details in the structure of the space separation with time separation vary with the initial condition, as indicated by the wide range of state space separations observed even for small time separations. Structure in the space-time-separation diagram indicates a time scale at which points have not "forgotten" their initial condition, the minimum near $\Delta t = 0.8$, for example, indicates macroscopic structure in the attractor which results in relatively near returns for a set a initial conditions after this duration. The definition of the correlation integral assumes the points constituting each pair are chosen at random on the attractor, ideally the duration of the observations will be *much* greater than any time-scale on which the space-time-separation diagram reveals significant structure.

This technique is also useful for identifying non-stationary processes, or at least data series whose statistics have not yet converged. Given data from a nonstationary process in which the *only* points close in state space are also close in time, there is no recurrence and thus no evidence for an attractor. If pairs of points nearby in time are retained then one will detect the dimension of the trajectory itself, the very bias which led Theiler to introduce the minimum time separation W in the first place [47, 52, 53, 54] Examples of space-time-separation (STS) diagrams for such systems are shown in Provenzale *et al.* [48], along with those for the Lorenz system. Contours of the STS structure for the for the Moore-Spiegel system including relatively large time separations are shown in Figure 12. Note that structure in the central contours (*e.g.* the mean)

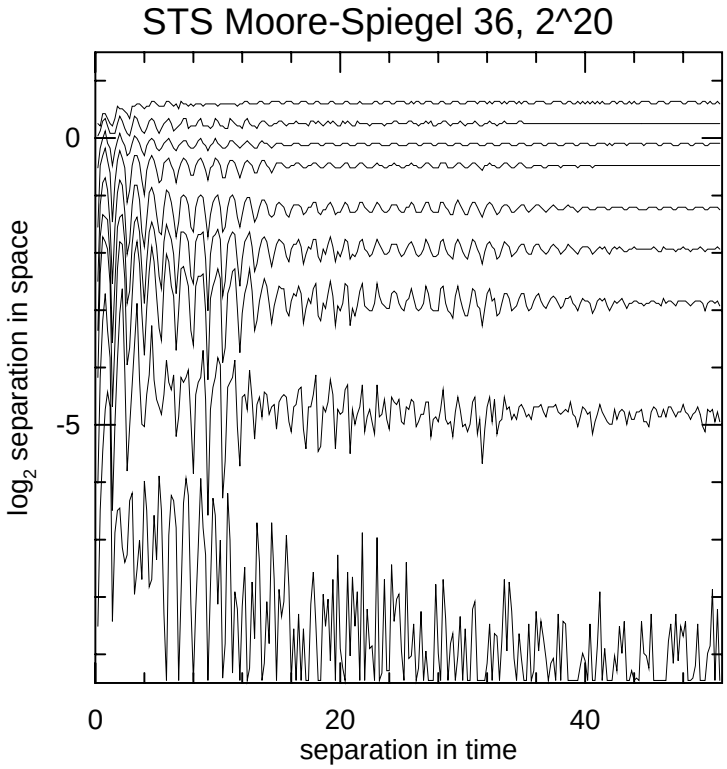

Figure 12: Space-time-separation diagram (contours) for the Moore-Spiegel system with $R = 100$ and $T = 36$. Each contour shows the state space separation within which a certain fraction of the data falls as a function of time, where the contours correspond to a fraction of .0001, .01 .1 .25 .5 .75 .9 .99 and 0.9999. For small Δt all pairs are relatively close in space as seen in the 0.9999 contour.

take a long time to disappear; long, say, compared with the "Lyapunov time." This is not surprising. Recall that Lyapunov exponents reflect the growth of infinitesimal separations, and as *effective rates* they need not reflect the growth of even infinitesimal separations if only a finite time is considered.

2.3.4 Intrinsic Limits to the analysis of Geometry

There are many limitations to dimension estimates regardless of how they are made. Even in an optimal case, computing a dimension requires a large quantity of data. Research based on a number of different assumptions [43, 55, 56, 57] suggests a rapid increase in the low bound on the data required with increasing dimension of the set. Some lower bounds consider only the limitations due to macroscopic structure, while others consider only the restrictions imposed by a minimal scaling range; unsurprisingly, calculations with include both effects suggest a larger lower bound[43]. A weak point in all general estimates of the amount of data required to estimate a dimension is that they assume some "typical" properties, often a space-filling distribution or a uniform density. Distributions with these properties are rarely encountered in practice.

Given two time series from the same dynamical system, the amount of data required will depend on the measurement function[11] chosen. These effects arise from the macroscopic structure of the embedded data. In the limit of zero length scales the measurement function is not important, macroscopically it is. In addition, the "number of points" useful for dimension calculations is determined by the length of the series in terms of the intrinsic time scale of the system (*e.g.* near returns in state space [43, 58]) rather than by the sampling rate. Space-time-separation diagrams quantify the combination of these two effects.

Also note that d_2 provides a *lower* bound on d_0, and what we are usually after is an upper bound on d_0. If the data set is small enough to be recorded, then obtaining a direct estimate of d_0 is not significantly more difficult than estimating d_2. For very low dimensional dynamical systems, the number of degrees of freedom indicated by d_2 and d_0 are similar, but then most integers less than four and greater than one are similar. It is not clear that this relation will hold for higher dimensional systems.

The fractal nature of strange attractors is reflected in the property that their dimension(s) may be fractions. Determining whether the dimension is an integer or a fraction can be very difficult since the uncertainty estimates for d_2 will often include an integer value. As stressed by Broomhead and Jones [59], determining the value of the fractional part of the dimension requires observing the fine structure of the geometry of the reconstruction, which is easily obliterated by observational noise. If there is sufficient data to probe the fine scales, it is the characteristics of the noise (which is usually space filling in this context) which are reported. Alternatively, when highly organised fine structure is observed, it can contribute to the uncertainty in the dimension itself, through lacunarity, as in Figure 9.

[11] and the delay time, and so on.

The lacunarity of a set reflects how gaps (or lacuna) appear as the range of scales changes. For strictly self similar sets, like the middle thirds Cantor set, this happens in a (logarithmically) periodic fashion[60, 61, 43, 50] and can be seen in $\chi(\ell)$. Such regular oscillations are not common in reconstructions of experimental time series. Yet maps often display similar effects (*e.g.* the Zaslavski attractor [62] analysed in the original GPA paper [63]), although this might only reflect macroscopic structure and have no impact once sufficiently small length scales are considered. For the Stiletto Map, oscillations in $\chi(\ell)$ due to lacunarity are observed at length scales analogous to those indicated in Figure 6 hamper a precise estimate of d_2. In general, figures similar to 9 and 10 can indicate whether "sufficiently small length scales" have been reached. In less organised lacunar sets, the $\chi(\ell)$ associated with the correlation integral will provide less information on the lacunarity of the set [64]. Yet in the case of the maps discussed above, at least, lacunarity oscillations reflecting, say, 10% variations in the estimate remain for some time.

But why are we measuring dimensions? If it is to identify low dimensional behaviour with an eye toward forecasting, it is more straightforward both to construct and to evaluate forecast models directly, than it is to verify dimension estimates. In fact, forecasting may provide a verification scheme for dimension calculations: if the data are of sufficient quality that accurate dimension estimations are possible, then the fine structure of the attractor is discernible and this structure must be preserved under the dynamics in order for the dimension estimate to converge. In this case, any reasonable prediction scheme will work, out-of-sample. Conversely, if no local prediction scheme works, then the dimension estimate was not accurate. Thus we have the following conjecture:

> A set of observations from a deterministic system which is of sufficient duration and quality that the correlation dimension can be accurately estimated is also sufficient for the construction of accurate local forecast models.

2.4 Takens' Theorem

In reality, we never have access to all the state space variables[12]. Often only one component is measured, the position of the gas parcel in the Moore-Spiegel system, for example. In the absence of observational noise the measurement will be a function of the state vector $\mathbf{x}$. In this context, each scalar signal (or observable) s will correspond to some *measurement function* $h(\mathbf{x})$ which defines s for each state of the system, that is

$$s(t) = h(\mathbf{x}(t)). \tag{27}$$

[12]Nor can we be certain that the underlying process is equivalent to a set of ODEs!

The measurement function is usually chosen for experimental convenience; how can we be sure it contains the information we need to reconstruct the dynamics? And if we measure only a scalar value, how can we construct a higher dimensional state space in order to mimic the true state space? Enter Takens' Theorem [65].

A delay coordinate function H simply builds an m dimensional vector, $\mathbf{y}(t) \in \mathbf{R}^m$, from m measurements separated by a delay time τ_d. In symbols

$$\begin{aligned} \mathbf{y}(t) &= H(\mathbf{x}(t)) \\ &= (h(\mathbf{x}(t)), h(\mathbf{x}(t-\tau_d)), \ldots, h(\mathbf{x}(t-(m-1)\tau_d))) \\ &= (s(t), s(t-\tau_d), \ldots, s(t-(m-1)\tau_d)). \end{aligned} \tag{28}$$

Takens' Theorem[13] tells us that, given a continuous time dynamical system with a compact invariant smooth manifold A, such that A

- is of dimension d_A,
- contains only a finite number of equilibria,
- contains no periodic orbits of period τ_d or $2\tau_d$
- contains only a finite number of periodic orbits of period $p\tau_d$, $3 \leq p \leq m$

and if the Jacobians of the return maps of those periodic orbits have distinct eigenvalues, then *with probability one*, a C^1 measurement function h will yield a delay coordinate function H which is a differentiable embedding from A to $H(A)$ for $m > 2d_A$.

What does this mean?

The theorem tells us that we do not have to measure all the state space variables of the system; that in fact almost any one will do. We can then reconstruct an equivalent dynamical system using delays as illustrated in the next section. This suggests that the particular measurement function chosen is not crucial, and neither is the delay time used. Many of the scaling exponents we would like to measure in the true state space, are preserved in the delay reconstruction. Note that the theorem applies to the manifold A, not the attractor; the Lyapunov exponents of the reconstruction $H(A)$ are those of A plus $m - d_A$ "spurious" exponents, which must somehow be identified. The issues involved are discussed in Darbyshire and Broomhead [32], Parlitz [67] and references therein.

Of course, the "measurement functions" of Takens' Theorem are found in a function space, not the laboratory. Even noise-free real data (observations)

[13] This summary is taken from Sauer *et al.* [66] which provides a good introduction.

when recorded by a digital computer would be truncated, so that real measurement functions are piece-wise constant, they are not continuous and yield a delay coordinate functions which are many to one. Thus the conditions for Takens' Theorem are not met by any real (finite resolution) observations. While we may draw comfort from Takens' theorem, it simply does not apply to finite accuracy observations. We return to this point when discussing quantisation noise in Section 2.6

2.5 The Method of Delays

Returning to the scalar signal, $s(t)$ recorded in discrete time, we have

$$s_i = s(i\tau_s) \qquad i = 1, 2, \ldots, n_s \tag{29}$$

where τ_s is the sampling time and each s_i is digitised to one of a finite number of values. A trajectory, $\mathbf{x}(t)$, is reconstructed in m dimensions from s(t) by the method of delays [68, 69] to yield

$$\mathbf{x}(t) = (s(t), s(t - \tau_d), \ldots, s(t - (m-1)\tau_d)) \tag{30}$$

where the delay time need not equal τ_s (although it must, of course, be an integer multiple of τ_s). In fact the $m-1$ delays used in defining $\mathbf{x}(t)$ need not be equal. Methods for choosing τ_d vary [70, 71, 66]; it is usually related to the decay of information in the signal with time, either from linear statistics like the autocorrelation time (τ_{auto}) or information theoretic statistics like mutual information [72]. When constructing nonlinear predictors, the delay may be chosen to optimise the predictor and need be neither constant nor uniform over the attractor, as illustrated in an experimental system in reference [73].

Reconstructions may also be based on the singular value decomposition of a (SVD) matrix whose rows are delay reconstruction vectors (see [74, 75, 7, 8]). Multi-variate series may also be employed, and often perform well with significantly less data, in terms of the total duration of the "experiment". This is easily understood as multivariate probes can add crucial information on the state of the system and thus distinguish states which would appear similar to univariate probes due to projection effects in state space.

As a simple example, let s_i equal the first variable of the Stiletto Map and $\tau_s = 1$. The method of delays then yields a series of vectors

$$\mathbf{x}_i = (s_i, s_{i-\tau_d}, \ldots, s_{i-\tau_d(m-1)}) \tag{31}$$

The results for $m = 2, \tau_d = 1$ are shown in Figure 13. If our aim is dimension calculations, we can repeat the approach of Section 2.3.2, including safety checks via space-time-separation distribution, using the delay vectors in the

place of the complete state vectors. The macroscopic structure will change depending on the measurement function, but our consistency tests remain valid. When Takens' Theorem holds, there is a diffeomorphism between the delay reconstruction and the state space dynamics; hence we can estimate Lyapunov exponents in delay-space, if we can determine the local linearization of the dynamics. This is best done by estimating the local dynamics through a prediction model, the construction of which is taken up in Section 3. Once a model is in hand, the approach is the same as when the equations of motion are known (recent advances in determining confidence limits are discussed in [33]). Before discussing model construction, a few words on noise are in order.

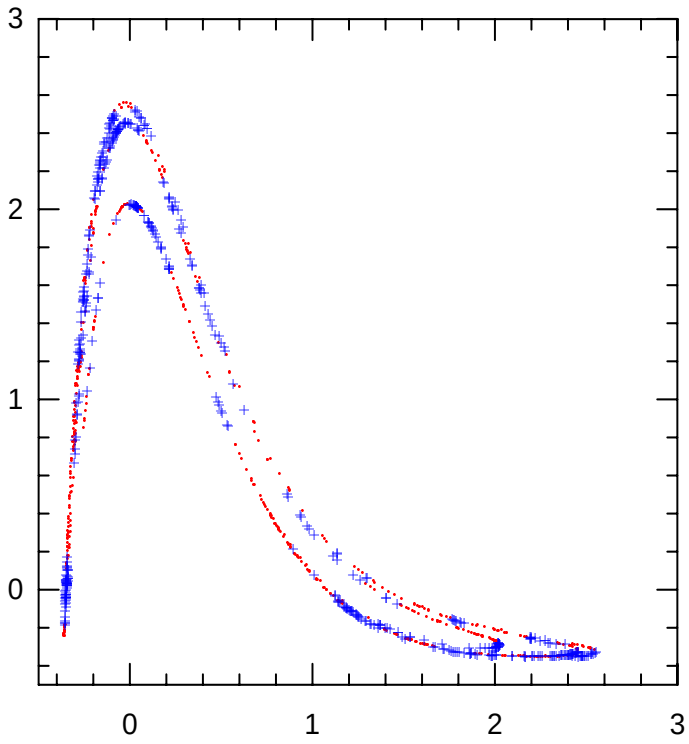

Figure 13: A delay reconstruction of the Stiletto Map, the similarity with Figure 6 is apparent. The macroscopic structure in the grouping of the '+' and '·' symbols reflect residual predictability of a local quadratic predictor, discussed in Section 5.3.3.

2.6 Noise

Thus far we have avoided noise. Let $\nu(t)$ be a stochastic process. In general, we need to distinguish two classes of noise. **Observational noise** alters the

observations by introducing a stochastic element into the measurement function

$$h_\nu(\mathbf{x}(t)) = h(\mathbf{x}(t)) + \nu(t) \tag{32}$$

but does not affect the state of the system. To the extent that we can obtain a perfect model, we may use this model to reduce observational noise in a self-consistent manner[76]. **Dynamical noise** directly affects the state of the system, and hence alters the future trajectory of the system. For a map

$$\mathbf{x}(t+1) = \mathbf{f}_\nu(\mathbf{x}(t)) = \mathbf{f}(\mathbf{x}(t)) + \nu(t) \tag{33}$$

destroying the deterministic framework of the dynamics. This is not just a technical problem: without additional knowledge of $\nu(t)$ the question of "removing" dynamical noise may not be well posed as there may be several distinct and equally valid solutions to the same inverse problem. Of course, real experiments usually have both observational and dynamical noise, but often provide some information on the structure of the noise. We take up the question of whether a given data set should be modelled as a deterministic process with observational noise, or as a stochastic process in Section 6. Questions regarding the existence of a general Takens-like Theorem for stochastic systems are of much interest; one might keep an eye out for progress in this direction [77].

As noted above, most measurements also include **quantisation noise** (also called **truncation noise**), since only a finite number of digits are recorded: most observational data sets consist of integers. Quantisation noise limits the application of Takens' Theorem, and places a lower bound on the length scales that can be considered in dimension estimates. If macroscopic effects persist at this length-scale, then the scaling range is inaccessible. Quantisation noise also plays an interesting role in delay reconstructions. Given a continuous signal of finite duration which produces a delay reconstruction trajectory of finite length, a reconstruction in three dimensions will *not* self-intersect when the dynamics are high dimensional. In the limit of infinite time, self-intersection will occur if the embedding space is not of large enough dimension, but for any finite time dimension 3 is big enough to avoid intersections (almost certainly). If a signal is quantised, however, then it will self-intersect in finite time even if $m > 2d_A$; in this case, finite time self-intersections will occur for any m if the process is either bounded or stationary.

This ends our discussion of estimating scaling exponents in low dimensional dynamical systems. Other gentle introductions can be found Eubank and Farmer [78] and the books of Nicolis [34] and Ruelle [42], in addition to references in the text. Low dimensional dynamics provides a nice theoretical structure, which we can provisionally assume to be true and then test for internal consistency. Either result tells us something about the system. In the next sections we move on to estimate the dynamics itself, and examine the predictability of finite uncertainties in nonlinear dynamical systems.

3 Prediction, prophecy, and pontification

This is only true when small variations in the initial circumstances produce only small variations in the final state of the system (This implies that it is only in so far as stability subsists that principles of natural law can be formulated: it thus perhaps puts a limitation on any postulate of universal physical determinacy such as Laplace was credited with.). In a great many physical phenomena this condition is satisfied; but there are other cases in which a small initial variation may produce a very great change in the final state of the system, as when the displacement of the "points" causes a railway train to run into another instead of keeping its proper course (We may perhaps say that the observable regularities of nature belong to statistical molecular phenomena which have settled down into permanent stable conditions. In so far as the weather may be due to an unlimited assemblage of local instabilities, it may not be amenable to a finite scheme of law at all.).

James Clerk Maxwell, Matter and Motion, 1877

3.1 Introduction

Prediction forms a vital role in science. In its simplest form, a prediction simply forecasts the state of a system, sometimes in real time. Prediction may also play the role of prophecy, as the prediction of the planet Neptune, based on the observation that the orbit of Uranus was at odds with the forecasts of Newtonian mechanics. In this section we are more concerned with predictions as forecasts: if a small initial variation may produce a very great change in the final state, then how can a forecast model be evaluated? How might we determine whether or not the weather is amenable to a finite scheme of law?

We first survey the ground rules for a statistically reliable analysis of data, and discuss the construction of data-based models. The fair evaluation of a deterministic model requires probabilistic forecasts, a point illustrated with ensemble forecasts of a fluid dynamics experiment. The Section concludes with a few comments on the extent to which one should trust the prophesies of a good forecast model; we delay an example of ensemble weather forecasts to Section 7.

Of course, any human endeavour is potentially at risk from predictions in the guise of pontification. Science is more immune that most endeavours, due in some part to consistency tests, but in larger part to the timely inoculations provided by out of sample data and the advancement of theory. Between 1859 and 1878 there were a number of reported observations of intra-Mercurial planets; with Einstein's general relativity the need for Vulcan vanished.

3.2 Simulations, Models and Physics

Prediction models come in a variety of flavours. Simple models attempt to capture the essence of a phenomena within a system of manageable size. Their merit lies in being interesting while tractable, and in that the origins of their behaviour are comprehensible. The Moore-Spiegel system extracts the essential dynamical processes governing the motion of a parcel of stellar atmosphere. There is no attempt to simulate all the physical processes of a stellar atmosphere in detail. An alternative approach is to build just such a simulation, including all known physical processes as accurately as possible: a **kitchen-sink model**, designed around computational constraints and with little regard for the comprehensibility of the model as a whole. This is the approach adopted with some success by modern weather forecasters. Yet another approach is to construct a model directly from the observations[14] (*i.e.* with no explicit model of the dynamics). Traditionally, this was the domain of the statisticians who, for the most part, constructed stochastic models.

The initial choice between a deterministic or stochastic model often hinges on the background of the person making the choice. Given observations from an interesting dynamical system, one may assume a stochastic model and bring the considerable resources of statistical modelling to bear, with the aim of generating "synthetic data" [79]. Or one may attempt a nonlinear deterministic model, construct time delay vectors using the techniques of the previous Section, and attempt predictions in time via interpolations in this **reconstruction space**. This second approach is the primary focus of this Section. In Section 6 we take up the question of how to best make an operational choice between a stochastic model and a deterministic model, in the presence of observational uncertainty.

3.3 Ground Rules

There is a basic framework in which we build and verify models. For simplicity, we will first consider models which predict a fixed **prediction time** into the future, τ_p. Each point $\mathbf{x}(t)$ on the reconstructed trajectory has a scalar image $s(t + \tau_p)$, a model should estimate this image for any $\mathbf{x}$. Prediction is thus reduced to interpolation, regardless of whether we consider a single **global predictor** valid for all $\mathbf{x}$, or a series of **local predictors**, each based on the behaviour of points near the particular $\mathbf{x}$ of current interest.

A forecast trajectory can be constructed either by **direct forecasts** using a different (single step) interpolation model for each prediction time, or with **iterative forecasts** where a single model is used, with model predictions being

[14] One might even build a computationally more convenient model from the output "data" of a full simulation, but we will not do so here.

used recursively as new initial conditions, in order to extend the trajectory. Nonlinear prediction schemes are extremely flexible, so much so in fact that they can over-fit the data by finding an interpolation surface which fits the noise on the signal in addition to the signal itself. When this happens, particularly small forecast errors are suggested in-sample, while particularly large forecast errors are observed in practice. Perhaps the most grievous error in this field is to both construct and evaluate a dynamic reconstruction on the same observations. To avoid this blunder, the data set may be divided into two sections: a **learning set** from which a model is derived and a **test set** on which various models are evaluated. This distinction must be maintained for **out-of-sample** evaluation of a predictor. While cross-validation may blur this simple picture of learning and test sets in questions of validating model structure, out-of-sample data are only out-of-sample *once*. Repeating runs with the same "out-of-sample" data set can be deadly.

Out-of-sample testing is the ultimate method of evaluation, and it is doubtful that the amount of data is "just on the edge" of being sufficient so that the algorithm only works if all the data are used. It is difficult to place a wager on the outcome of yesterdays weather, even if one has a very good weather model; it is in this sense that one should be hesitant to wager the merit of a research program on an analysis of data previously considered in the course of that program.

3.4 Data-based models: Dynamic Reconstructions

A wide variety of approaches has been explored with the common aim of extracting the information contained in the time ordering of points in delay-space. Examples include references [80, 81, 82, 83, 84, 85, 86, 87, 12, 88, 89, 90]. While these approaches differ in detail, they all attempt the same task, since in the context of deterministic analysis, prediction in time is equivalent to interpolation in state space. One determines the future behaviour assuming that it will be similar to that of a sample of the "nearby" points, this immediately restricts applications to systems whose trajectories are recurrent in state space. Success depends both (1) on having a sufficient number of nearby points to satisfy the smoothness assumptions of the chosen algorithm and (2) on adopting a model which is capable of reproducing (*i.e.* fitting) the surface, that is, an acceptable model structure. In the presence of observational noise, this requirement is increased so that the variations due to the noise may be, in some sense, averaged out. And in systems with chaotic attractors, predictors must also account for extreme inhomogeneities in the distribution of the learning set in state space.

3.4.1 Analogue Prediction

Perhaps the most straightforward method of prediction is to choose an analogue: simply take the point in the reconstruction nearest to the point to be predicted (its nearest neighbour) and either report the nearest neighbour's image as the prediction, or report the sum of the nearest neighbour's forward first difference and the current value as the prediction. The strengths and weaknesses of this approach were investigated in the late 60's by Lorenz [91]. Comparing the quality of results with those of other methods [92, 73] shows that these deterministic analogue predictors should not be rejected out of hand.

A stochastic variation of analogue prediction is found in Random Analogue Prediction (RAP) models [15] which select a near neighbour at random, with the probability of selecting a particular neighbour usually based on the distance between the prediction point and that neighbour. When the learning set is very large, the selection probability should be taken to reflect the observational noise. Given an exceedingly large learning set, quantisation noise will yield a number of analogues with *exactly* the same state space coordinates and different images for any fixed embedding dimension. Given this ambiguity, a RAP model forms a single forecast by choosing between these alternatives at random; another approach is to employ RAP ensembles and form probabilistic forecasts, as discussed in Section 3.5 below.

3.4.2 Local Prediction

Given k points within a neighbourhood, a local linear predictor aims at the linear map with the smallest mean square error when interpolating the future observation. A local quadratic predictor works similarly, but includes the quadratic terms. For small data sets, determining the correct size for each neighbourhood is crucial; if it is too large, higher order nonlinear effects will be included, while if it is too small the quadratic predictors may over-fit the data. A major difficulty is that "large" will vary with initial condition. Casdagli [89, 93] has investigated the variation of predictions with k using local linear maps in a variety of circumstances while Smith [90, 94] determines a local value for k based on the local structure of the interpolation surface and the data density. Deterministic local linear prediction of chaotic systems was initiated by Farmer and Sidorowich [81], while a parallel local stochastic approach was proposed by Priestley [95]. Local predictors need not be polynomial, of course; the schemes mentioned below as global predictors may be applied locally.

3.4.3 Global Prediction

Typically a new local predictor must be constructed for each initial condition, while global predictors cover the entire domain. The range of popular model structures for global prediction is varied, running, for the moment, from neural networks [96] through attempts to extract explicit equations of motion [80] and on to general, but comprehensible, interpolation methods[83].

Each m dimensional vector, $\mathbf{x}(t)$, in the learning set is associated with a (future) value to be predicted, $s(t + \tau_p)$. A predictor is then a map, $F(\mathbf{x}) : \mathbf{R}^m \rightarrow \mathbf{R}^1$ which estimates s for any $\mathbf{x}$. Radial basis function (RBF) predictors [82, 83, 97] consider $F(\mathbf{x})$ of the form

$$F(\mathbf{x}) = \sum_{j=1}^{n_c} \lambda_j \phi(||\mathbf{x} - \mathbf{c_j}||) \tag{34}$$

where $\phi(r)$ are radial basis functions [98]. Typical candidates for $\phi(r)$ include $\phi(r) = r^3$ and $\phi(r) = e^{-r^2/\sigma^2}$ where the constant σ reflects the average spacing of the centers $\mathbf{c_j}$. $F(\mathbf{x})$ is constructed about n_c centers

$$\mathbf{c_j}, \quad j = 1, 2, \ldots, n_c; \quad \mathbf{c_j} \in \mathbf{R}^m \tag{35}$$

chosen to cover the region of state space which the reconstruction explores.

Requiring

$$F(\mathbf{x(t)}) \approx s(t + \tau_p) \tag{36}$$

for all $\mathbf{x(t)}$ in the learning set yields the constants λ_j by solving a *linear* minimisation problem. Specifically, the λ_j are determined from the solution of

$$\mathbf{b} = \mathbf{A}\lambda \tag{37}$$

where λ is a vector of length n_c whose j^{th} component is λ_j and $\mathbf{A}$ and $\mathbf{b}$ are given by

$$A_{ij} = \omega_i \phi(||\mathbf{x_i} - \mathbf{c_j}||) \tag{38}$$

and

$$b_i = \omega_i s_i \tag{39}$$

where $i = 1, \ldots, n_L$ and $j = 1, \ldots, n_c$. The weights ω_i reflect the observational uncertainty of each point in the learning set, and may also be used to achieve a more uniform quality fit across different regions of the state space. This will always increase the in-sample RMS prediction error *and* may improve the model significantly. The performance of RBF predictors is often improved by including polynomial terms in Equation 34 and increasing the size of $\mathbf{A}$ accordingly, as discussed in the references.

Given the pseudo-inverse of $\mathbf{A}$, which may be computed via SVD, λ may be determined from Equation 37. Thus global predictors are efficient when

operated in real-time or evaluating large test data sets, since the computational overhead is done at the outset. A global reconstruction also provides a natural partition of the state space which can serve many uses, such as estimating the expected range of out-of-sample forecast errors in each prediction [97] or allowing parameters like the delay time to vary with location [73]. Local predictors, on the other hand, are often much faster to construct, are more easily modified to include new observations (note, however, results by Stark [99]), and tend to show less systematic bias in the data sparse regions of the reconstruction.

3.5 Accountable Forecasts of Chaotic Systems

Suppose we have a forecast model and an initial observation. For simplicity, assume that the observational noise is dominated by quantisation error, each variable being observed with 6 bit accuracy[15]. Placing the observation in our model yields a prediction: How *un*reliable is this prediction?

There are a number of ways this question can be addressed. If $2^{-6} \approx 0$ then we can evaluate the linear propagator of our model and characterise the expected uncertainty growth through that of an infinitesimal uncertainty, as discussed in Section 2.1.2. At some time, however, the linear approximation will become irrelevant. An alternative approach is to evolve an **ensemble** of initial conditions, each of which is consistent with the observation. In this case, each member of the ensemble would start out in the same quantisation "box" in state space, it would agree with the observation to 6 bits and differ in the details which are unknown due to quantisation error. The uncertainty in the initial condition is then reflected by a distribution of possible future states. But will this probability density function (PDF) provide an accurate description of the likely future? The ensemble described above will not, since it assumed a uniform (isotropic) initial distribution within the quantisation box, while the true initial condition was on the Moore-Spiegel attractor. If we form an ensemble of points both within the quantisation box and also on the attractor, we will have a **perfect ensemble**. Even a perfect ensemble evolved under a perfect model will not yield a unique forecast of "the" correct future state, but it will be **accountable**: it will reflect the true PDF. Note that, in general, the path of a maximum of the PDF need not correspond to a realizable trajectory of either the model or the system. Neither will the path of the ensemble mean, nor the median.

The accuracy of the PDF will increase as the number of members increases: the number of unforeseen events or "forecasts busts" (occurrences with low

[15]This divides the state space into a mesh of m-dimensional boxes; an observation indicates which box the system is in, but says nothing as to where within that box it is.

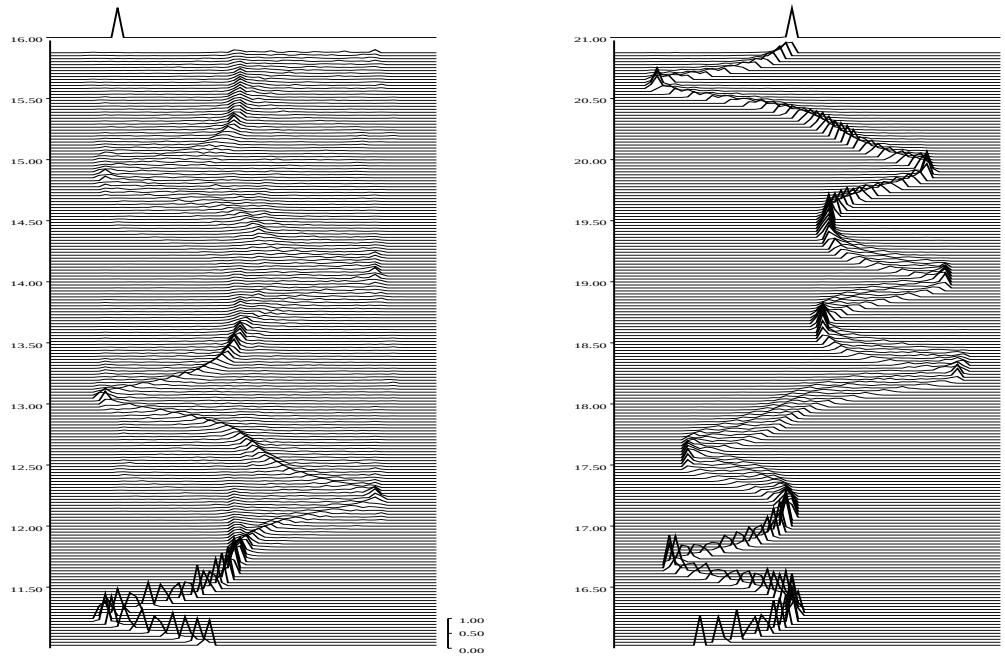

Figure 14: Perfect ensembles evolved under a perfect model of the Moore-Spiegel system. Each column shows the dynamics of uncertainty as time increases (upwards) as reflected in the evolution of the probability distribution for the variable x.

(zero) probability in the forecast), should decrease as the size of the ensemble increases in the expected way. In addition, we should be able to use ensemble members to determine the accuracy of observation required to obtain a desired forecast accuracy; this model/observation/ensemble formation scheme is fully accountable in the sense of Popper[100]. Operationally, perfect ensembles can be constructed by observing (in this case, integrating) the system and recording those points on the trajectory which are indistinguishable from the initial condition given the limited measurement accuracy. The future trajectories of these points provide a PDF forecast reflecting initial conditions within the same quantisation box, *and also* on the attractor.

This defines the optimal forecast scenario: a perfect model of the dynamics, an exact understanding of the origin of observational uncertainty, and a perfect ensemble of initial conditions. Figure 14 shows the evolution of perfect ensembles for two initial observations in the Moore-Spiegel system. Time increases

from bottom to top, in the lowest line in each panel shows the distribution of values of x consistent with the observation at the initial time. As time increases, this distribution evolves.

On the left-hand side, the ensemble initiated at $t = 11$ disperses fairly rapidly (by $t \approx 12$). Yet the location of the most likely value of x can be identified much longer. The actual location of the fiducial trajectory for which the prediction was made is shown at time $t = 16$ at the top of the panel, separated by a gap for clarity. In this case, the forecast PDF was rather disperse. The evolution of a new perfect ensemble, initialised $t = 16$, is shown on the right-hand column; it remains coherent much longer. Yet even in this case, the macroscopic structure visible at $t \approx 17.5$ makes infinitesimal prediction measures irrelevant. At $t = 21$ the forecast PDF is still rather sharp; the distribution above a gap reflects the location of the fiducial trajectory at $t = 21$ in full agreement with the forecast PDF as at $t = 16$, but this time the location is better identified by the forecast.

And that is the main point of ensemble forecasts. Both the forecasts are perfect: the forecast PDF at $t = 16$ and the forecast PDF at $t = 21$ both accurately reflect our uncertainty at time $t = t_0 + 5$, given our initial uncertainty at t_0. But the forecast on the right is much easier to use. And ensemble forecasting tells us this will be the case at the same time that the forecast is made. In this manner, we may determine whether the best single forecast is likely to prove unreliable, and make our plans accordingly. Graphs corresponding to Figure 14 are available for the Lorenz system [101] and the Hénon Map [102]. The examination of information flow through the evolution of ensembles in 1-dimensional maps dates back at least to Shaw [103]. We must make probabilistic predictions for nonlinear deterministic system, even if the probabilistic aspect comes only from the uncertainty in observation.

3.6 Evaluating Ensemble Forecasts

Ensemble forecasts can be evaluated by their calibration plots [104] as shown in Figure 20. These are constructed by dividing the range of s into, say, 64 bins. For each forecast, the probability which the PDF assigns to each bin is recorded along with the identity of the bin into which the observation actually fell. The results of many forecasts are combined by grouping together all the bins which had forecast probability of, say, between 60% and 70% and computing the fraction of these in which the observation actually fell. In this case, the fraction equals 0.65 or thereabouts, for a well-calibrated model.

A calibration plot displays the relative frequency of the observation against the forecast probability. For a well calibrated model, the forecast probability will equal the relative frequency. A well-refined forecast, on the other hand, has many forecast probabilities near either 100% or 0%. A perfect model will

be well calibrated, but refinement depends on the dynamics of the system and level of observational uncertainty as well as the model.

What happens at long forecast times? Eventually, any finite ensemble will become indistinguishable from a set of randomly drawn states[16]; this limiting distribution, $\psi_\infty(x)$, is often referred to as the **climatology**. Once this happens our forecast is definitely **useless**[17]. Unsurprisingly, the average forecast time at which this occurs cannot be determined from the Lyapunov exponents.

The optimal forecast scenario is restricted either to those systems which we construct or to those for which observation over many Poincaré return times allow the collection of good analogues. But then the entire framework of prediction via interpolation already requires a high degree of recurrence in state space. Even in low dimensional chaotic systems, the linearized dynamics are of little use and reliable forecasting requires ensemble prediction. Yet it is interesting to ponder whether this happens **even in, or only in** low dimensional systems of ODEs. Thus far our theorizing has been based solely on the output of our theories, we have examined no real data. To quote Holmes[18]: "It is a capital mistake to theorize before one has data."

3.7 The Annulus

A favourite system for which *no* perfect model exists is the thermally driven rotating fluid annulus [107]. Thermal convection in this cylindrical fluid annulus, differentially heated in the horizontal while rotating about a vertical axis, provides a laboratory analogue of the large scale circulation of the Earth's atmosphere. The experiment has a long history; its complicated, nonlinear dynamics were cited as motivation by Lorenz in 1963. Laboratory observations of the annulus hold significant advantages over both numerical models and meteorological observations. Qualitatively similar difficulties arise here as in meteorological observations, and these cannot be circumvented serendipitously, as sometimes occurs in numerical studies: the annulus is an infinite dimensional (fluid) system, imperfectly observed. Yet the physical time scales over which accurate observations can be made is much greater than the (representative) time scales for the atmosphere. We will consider results from several RBF models constructed from time series of temperature measurements from a co-rotating probe. While we shall restrict discussion to ensembles over initial conditions, it is often advantageous to consider ensembles of different model structures and over different parameter values within a given model structure [108]. The dataset discussed below was taken in a parameter regime for which

[16] In this case, the projection of the invariant measure under the measurement function.

[17] The point at which it was last useful depends on the user and may occur well before the onset of uselessness. A discussion of these issues is given references [105, 101].

[18] Holmes to Watson in A Scandal in Bohemia [106].

the flow appears to display baroclinic chaos, and has also been examined by R. Smith [109].

The ensemble evolutions in Figure 15 are determined from a 5-dimensional RBF model integrating over a short time scale. The initial *imperfect* ensemble had a Gaussian distribution with standard deviation 0.02 degrees and mean equal to the observed temperature. Every 16^{th} iteration, a new observation is incorporated, a new ensemble with Gaussian structure is initialised, (a gap is placed in the figure to mark these locations). The extent to which the predicted PDF contains the new Gaussian distribution reflects the accuracy of the prediction.

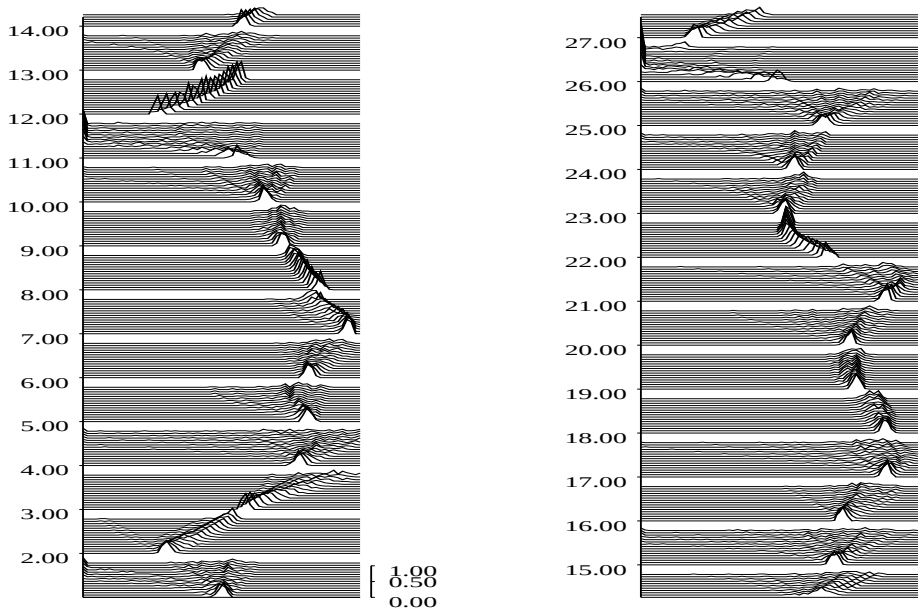

Figure 15: The evolution of the probability density functions reflecting ensemble forecasts of temperature in the rotating annulus experiment. As in the previous figure, time increases upwards. The gaps are inserted to mark the time of each new observation. With each new observation, the uncertainty collapses back to a Gaussian distribution.

Overall the ensemble forecasts reflect the evolution of the system fairly well. The $t = 1$ ensemble spreads slowly; while its mean value increases slightly,

there is noticeable probability of decreasing temperature which matches with the true observation reflected in the initial distribution at $t = 2$. Ensemble 2 (the ensemble initiated at $t = 2$) remains roughly Gaussian, indicating slowly increased temperature and comparing well with the initial location of ensemble 3. Ensemble 3 shows the rapid development of a multi-modal PDF as in ensembles 4, 11, 14, 15, and 26, among others. It is clear that predictability varies with the initial condition.

While Ensemble 11 correctly indicates an impending rapid drop in temperature without specifying when it will occur; this may be contrasted with ensemble 26, which gives a better indication of the timing of the drop as well as reflecting the initial state observed for Ensemble 27; perhaps this reflects "return of skill" [110, 111]. Model error is illustrated by ensemble 12, where we have a sharp well defined PDF which does not reflect the observation at $t = 13$. Also unlike the perfect model results, some members of the ensemble diverge immediately from the range of observations; the small waves travelling to the left (low temperatures) in ensemble 1 provide an example. In addition, the large time behaviour of the model does not match that of the system; typically the model dynamics collapses onto either an attracting fixed point or an unphysical stable periodic orbits (or "ruts" [86]), neither of which is observed in the system. Thus the PDF forecast under an imperfect model need not approach $\psi_\infty(x)$, even asymptotically.

It is, of course, impossible for an imperfect model to yield good ensemble predictions if it cannot reproduce the behaviours of the system. As a last example with the annulus, we ask whether it is *possible* for any imperfect model forecast to remain consistent with the data (*i.e.* to within the observational uncertainty) over a given period. If so, we say that the model **ι-shadows** the observations over this period.

Figure 16 shows several ensemble forecasts over longer time scales. The solid line reflects the observed temperature as a function of time; the circles reflect the initialisation of ensembles at times $t = 11, 17, 23$, and 29. After each initialisation, the ensemble is iterated for 6 steps; the distributions of dots reflect the individual ensemble members. There is good general agreement, and even an indication of return of skill at $t = 13$ and 21.

So what is the longest time for which some model trajectory will stay within the observational uncertainty of the observed temperature? The dot-dashed line in Figure 16 traces a model trajectory that ι-shadows the observations for 26 time steps, first exceeding a distance of 0.2 degrees from the observations at $t = 31$. Methods for determining this trajectory are discussed by Gilmour [112]; we return to use shadowing trajectories in model evaluation in Section 5.2.

Additional discussion of these forecasts is given in [102]. It is clear that these

predictions will not be accountable: not only is the model imperfect (*e.g.* Ensemble 12 of Figure 15), but the initial conditions are not on the attractor[19]. Since we can take large ensembles for the RBF model, we will postpone the question of selecting prefered initial orientations to the case of Numerical Weather Prediction (NWP) in Section 7.

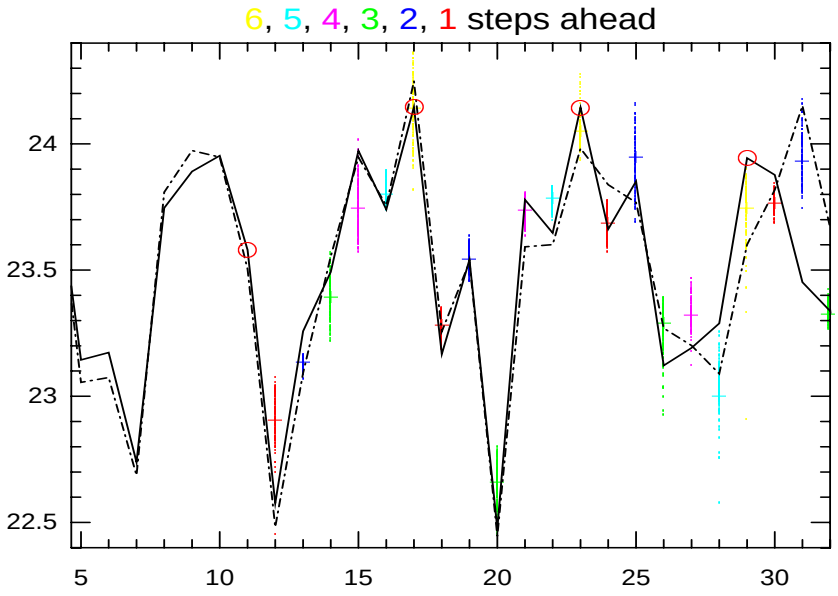

Figure 16: A temperature time series (solid) from the annulus and a shadowing trajectory (dot-dashed) from a global, iterated RBF predictor. Each ensemble contains 128 points, initially within 0.075 degrees of (each component of) the observation delay vector. A new ensemble is initiated at 6 steps. The data displayed is from an out-of-sample evaluation.

3.7.1 Prophecies

Even good models are rarely reliable in conditions which differ from those under which the model was constructed. The interpolation models discussed here, for example, will only rarely predict a value outside the range of the observations, *not* just because such a novel event (*e.g.* a new record high temperature or flood level) is unlikely, but simply because the model structure constrains most trajectories to regions well explored in the learning set. Near the (observed) global maximum all observed trajectories decrease, thus interpolation between these values will decrease as well. Analogue forecasts can *never* set new records for the variable predicted. While in both local and global interpolation models, the fraction of the forecast ensemble which "turns back" tends to increase too

[19]Indeed, a few initial conditions in Ensembles 1, 2, 4 for example, are not even in the basin of attraction!

rapidly as an observed maxima is approached, and forecast trajectories which successfully predict the magnitude of new record extremes are rare. This is not surprising. Interpolation is easier than extrapolation.

Of course, if we have observations at a variety of parameter values, we can construct RBF models which also vary as a function of this parameter, as demonstrated by Casdagli [83] in his initial paper on radial basis functions. If parameter values differ from the range in which we have observations, then the results must be interpreted with great care, including some comparison between models with different structural assumptions at the very least. Of course, the same care is required when interpreting the output of full simulation models when they are run outside the parameter range in which they have been tuned. The loss of superposition of solutions in nonlinear systems makes extrapolation outside the range of observed parameters and initial states rather adventurous, given anything less than a perfect model. We are more likely to spot Vulcans than Neptunes.

4 Aids for more reliable nonlinear analysis

> *I am very sorry, Pyrophilus, that to the many (elsewhere enumerated) difficulties which you may meet with, and must therefore surmount, in the serious and effectual prosecution of experimental philosophy I must add one discouragement more, which will perhaps as much surprise as dishearten you; and it is, that besides that you will find (as we elsewhere mention) many of the experiments published by authors, or related to you by the persons you converse with, false and unsuccessful (besides this I say), you will meet with several observations and experiments which, though communicated for true by candid authors or undistrusted eye-witnesses, or perhaps recommended by your own experience may, upon further trial, disappoint your expectation, either not at all succeeding constantly or at least varying much from what you expected.*
>
> R. Boyle, *Concerning the Unsuccessfulness of Experiments*, 1673.

The serious and effectual prosecution of experimental philosophy requires consistency tests. In this section we play three variations on this theme. First, the use of surrogate data to estimate the likelihood of having obtained an apparently interesting result fortuitously. Second, the use of strong straw men in evaluating algorithms before they are deployed on real observational data. And finally the use of simple models to estimate lenient *lower* bounds for the minimum duration of observation required for a given analysis to converge.

4.1 Significant results: Surrogate Data, Synthetic Data and Self-deception

It is an empirical fact that only interesting results are tested. Hence our first task is to estimate whether an apparently interesting result could have been avoided (and thus really is interesting). If the result could not have easily been avoided (*i.e.* if any similar looking data set would yield a similar result), we may wish to limit our interpretation of the result. Thus one goal is to construct **surrogate data** which resembles the observations, but contains none of the desired physics. This may be done through manipulating the observations themselves, or by using a model to generate surrogate data sets which "look like" the observations, but are known *not* to have the property of interest (*e.g.* they are not chaotic). We then attempt to establish the insignificance of our interesting result with the analysis of surrogate data.

A very simple example of the method of surrogate data would be to evaluate the estimated lag-one correlation coefficient between consecutive points in an observed time series, s_i: is the observed correlation between all pairs s_i and s_{i+1}

significantly greater than zero? Suppose the measured correlation coefficient is α. We can construct a surrogate series with the same distribution as the original data set but in which we know (by construction) that there is no correlation between consecutive values, by simply shuffling the original data set. Denote the correlation coefficient of the shuffled data set $\hat{\alpha}$. By constructing a large number of different surrogate series in this way, we can estimate the distribution of $\hat{\alpha}$ and then determine the likelihood of obtaining the observed value α under the **null hypothesis** that these particular data values were randomly ordered. Schematically:

i) compute α from the observations,

ii) create a surrogate set z_i by randomly shuffling the s_i,

iii) estimate $\hat{\alpha}$ from the z_i,

iv) repeat (*ii*) and (*iii*) to obtain a distribution of $\hat{\alpha}$, expecting $\langle\hat{\alpha}\rangle = 0$,

v) estimate the probability of observing α by chance from this distribution.

In effect, we determine the range of values empirically indistinguishable from zero. When estimating correlation coefficients this might be done with an analytical approach given a few assumptions about the process; when estimating Lyapunov exponents neither the approach nor the relevant assumptions are known.

Our hope is that our result is truly interesting; if this is the case then we will be able to reject the null hypothesis. That is, we will be able to conclude that there is only a small chance that the observed value of α would have been obtained *if* the data were randomly ordered in time. Of course, rejecting a null hypothesis is only interesting if it was a relevant null hypothesis. And there's the rub. While it is straightforward to dream-up an algorithm for testing the null-hypothesis "not correlated", no method is known for testing the much more interesting null-hypothesis "not chaotic."

Our target, a well formulated null-hypothesis which is also relevant to the problem at hand, can be difficult to achieve. It is easily achieved if the relevant null is that the data points are independent and identically distributed (**IID**) random variables (commonly called **white noise** due to its flat power spectrum); but most experimentalists will refine the sampling time until this null-hypothesis is no longer of interest! An extended discussion of red noise[20] null hypotheses in the context of detecting periodic oscillations via Singular Spectrum Analysis (SSA) is given in [8]. Discussions and examples of the use of surrogate data in nonlinear dynamics can be found in Theiler *et al.* [113, 114], Smith [97], and references thereof.

[20] Autocorrelated noise with more power at low frequencies, hence "red."

There are two basic approaches to constructing surrogates, we will distinguish them (in extreme cases) as generation by *equivalence* or by *similarity*. Let us call **equivalence surrogates** those which are generated by reproducing some statistic of the data set exactly, such as the autocorrelation function [115, 113]; these surrogates have the advantage of providing a well framed null hypothesis, but can be sensitive to the particular choice of statistic and the quality with which it can be estimated from the data set. This last point is particularly important with short data sets. Inasmuch as all models are wrong, it would be surprising if we could not find some statistic with which to reject any given null hypothesis. That is not our goal. Instead, we aim to gain (or lose) confidence in conclusions drawn on the evidence of our analysis of the data set, by seeing whether we might well have obtained similar evidence from the same analysis applied to a (surrogate) data set for which we know these conclusions are false.

Relaxing the requirement of a well-formulated null hypothesis may yield a more relevant null hypothesis. **Similarity surrogates** can be taken from any model whose output "looks like" the observations in question, and address the less formal question of whether a given result should be expected from "any" series which looks like the observed data. This can place the qualitative results of an algorithm in context, like the visual appearance of the sunspot reconstruction in Figure 3.

Generators for similarity surrogates can often be found in the literature on the phenomena which generated the data set of interest. A good source of stochastic surrogate data for questioning conclusions of deterministic dynamics in sunspots is provided by the Barnes model [5] which incorporates the structure of an autoregressive moving average ARMA(2,2) model with nonlinear modifications to ensure that the signal (1) remains asymmetric and positive and (2) tends to increase more rapidly than it decreases. These two well known properties of sunspot number make testing the observational data against simple "linear stochastic" surrogates futile, since linear stochastic surrogates share neither of these properties. The Barnes model is

$$z_i = \phi_1 z_{i-1} + \phi_2 z_{i-2} + a_i - \theta_1 a_{i-1} - \theta_2 a_{i-2} \tag{40}$$

$$s_i = z_i^2 + \alpha (z_i^2 - z_{i-1}^2)^2 \tag{41}$$

where $\phi_1 = 1.90693, \phi_2 = -0.98751, \theta_1 = 0.78512, \theta_2 = -0.40662$, $\alpha = 0.03$ and the a_i are IID Gaussian random variables with zero mean and standard deviation $\sigma = 0.4$. This rather pedestrian approach to surrogate data has it benefits. The ease with which 300 year segments of the output from the Barnes model can mimic the correlation integrals [2] and SVD reconstructions (see Figure 3) of the observed sunspot series, help us to maintain our uncertainty in the face of what initially appear to be very interesting results.

Framing a precise null hypothesis in the case of similarity surrogates can be quite difficult, just as insuring that equivalence surrogates are are not rejected for some secondary reason is also quite difficult. Ideally, the two approaches would converge, but equivalence surrogates often do *not* resemble the original data set, while it is difficult to determine what properties similarity surrogates are reproducing.

Returning to Figure 1, there is without doubt a statistically significant relationship between the number of sunspots and the Republican fraction of the US Senate. If we were handed these two time series at random[21], we would be hard pressed to deny the possibility that some connection existed. Of course, we were not handed these two time series at random, the time series of Republicans was chosen *because* of the correlation. Without knowing how many different time-series were tested and discarded, we simply cannot evaluate the "significance" of this observation, without more data. This bias is related to the so-called file-drawer effect in medical statistics: if the same drug is tested 20 times, there is a good chance that one set of test results will be significant at the 95% level. Publishing that result, while filing the other tests mis-represents analysis. [22] A similar confusion may arise in the analysis of a time-series, when evaluating the power contained in each of 40 different frequencies, and then noticing that one of them is significant at the 95% level. Even when good statistical tests are correctly employed, correctly setting significance levels may require information we do not possess. At some level, there is no recourse but to revert to the method of real data: acquire longer observational data sets and see if the original hypothesis is supported.

Synthetic data, in the usage of Juneja *et al.* [79], aims to be indistinguishable from the true observations. Thus the stochastic models of Juneja *et al.* provide excellent generators of surrogate data for evaluating general conclusions of determinism in fluid turbulence data. To the extent that these models cannot be rejected we have little support for deterministic dynamics; to the extent that they can be rejected, the models fall short of the goals of synthetic data. Either result would be of interest.

In concluding, we stress that the method of surrogate data can only establish insignificance. We can determine that there is no evidence that an estimated Lyapunov exponent is greater than zero, for example, by comparing it with the distribution of estimated exponents from linear stochastic series generated via phase randomisation[113, 48]. But if our estimated Lyapunov exponent is significantly different from this distribution, we can say only that it appears positive with respect to linear stochastic processes. This is a necessary but not sufficient condition to conclude that the estimated Lyapunov exponent is

[21] I was first handed these two time series by David Wark of Balliol College, Oxford.

[22] Although it has never been clear to me why statisticians, as a profession, are content to go astray about one time in 20.

positive. We are often unable to formulate, much less to reject a more relevant null-hypothesis. But it is also true that often we are unable to reject fairly simple null hypotheses. This is a useful result, as it immediately allows us to reevaluate our our expectation, thereby reducing the probability of its disappointment upon further trial.

4.1.1 Surrogate Data and the Bootstrap

Although both involve many of the same manipulations, the aims of the **bootstrap** of Efron [116, 117, 118] and surrogate data tests are fundamentally different. The bootstrap approach makes a variety of assumptions about the process which generated the data, and then computes the uncertainty of an estimated statistic by assuming those assumptions are true. Alternatively, the surrogate data approach selects a process that is *not* consistent with the type of process believed to have generated the data, and then attempts to establish that the value obtained from the observations is unlikely under this null hypothesis. Bootstrapping aims to build a distribution which is consistent with the true uncertainty in the observed result and thereby gain an estimate of this uncertainty. The method of surrogate data aims to build a distribution from a known (surrogate) process and show that the observed result is inconsistent with this distribution, thereby rejecting the surrogate process and strengthening the interpretation that the observed result is interesting (or at least, not wholly insignificant). The manipulations involved are similar, but the aims are very different.

4.1.2 Surrogate Predictors: Is my model any good?

The discussion thus far has centered on using surrogate data to establish the insignificance of an estimated statistic; a parallel approach can be taken when evaluating prediction models. It has two aspects. **Surrogate predictors** can be used to quantify the magnitude of forecast errors expected from the most naïve forecasts, the simplest being either persistence or a random choice from the observed distribution. The idea being that, to be interesting, a predictor must perform significantly better than these. Examples for the case of the rotating annulus are given in [97]. If we claim that a nonlinear model provides improved forecasts by successfully capturing a subtle nonlinearity in the data, we can evaluate the significance of this improvement by contrasting the prediction errors of the nonlinear model with those of a variety of linear models (out-of-sample, of course) and quantify the extent to which it outperforms them. It is impressive to observe a nonlinear model predicting out-of-sample more accurately than a linear model can predict the same data in-sample! The second aspect falls along the same line: if we re-fit the nonlinear model to data

generated by a linear surrogate process, for example, this advantage *should* vanish.

In comparing linear and nonlinear models, one must immediately confront the issue of parsimony: how many free parameters should be allowed in the nonlinear forecast model? Many results are available under the assumption that the process is linear; for nonlinear systems (and potentially-linear systems as well, for that matter), there is no substitute for out-of-sample evaluation. A low order nonlinear process may require a high resolution model if the model structure does not closely reflect that of the process. Out-of-sample evaluation can then detect over-fitting. Current computation power eases significant cross-validation, but some segment of the data should be used for true out-of-sample tests.

Related issues arise whenever one ignores the smaller elements of a singular value decomposition [21, 26]. The rank ordering of the singular values *assumes* bigger means more important. In the case of SSA this translates into the assumption that high variance implies high information content. In nonlinear systems this need not be the case: a low order nonlinear process[23] may project into an infinite number of singular vectors; those projections which are small on average need not be the least important. A similar quandary ensues in determining coefficients of radial basis function models when computing the SVD of the matrix $\mathbf{A}$ of Equation 38. It is not obvious how to translate knowledge of the noise level of the observations into a threshold on the singular values of $\mathbf{A}$; this threshold links positional uncertainty in state space with structural constraints in function space. Out-of-sample forecast errors often indicate that this threshold should be significantly *smaller* than linear intuitions might imply; less parsimonious nonlinear models often consistently outperform those determined with higher thresholds, out-of-sample.

4.2 Hints for the evaluation of new techniques

4.2.1 Avoiding Simple Straw Men

Given a new technique, or a new computational code implementing an algorithm, it is tempting to test it on data from well studied dynamical systems. This often leads to very misleading results, particularly when the "straw man" of choice has a simplicity lacking in the ultimate application. A classic example is to test a prediction algorithm which can fit (any) polynomial function exactly on noise-free numerical data from the Hénon Map. The source of the danger here is analogous to that in the examination of surrogate data: failure to predict the Hénon system tells you a great deal about your prediction al-

[23] One which can be described with only a few nonlinearly coupled degrees of freedom.

gorithm, while success in predicting this quadratic map tells you little about your algorithm's prospects in the real world.

To learn more, one may attempt to make the test resemble the application: use data sets of similar length and noise level. Data from a surface of section, as in the left panel of Figure 5, will provide a much better test than most maps, but even moving to something as simple as the Stiletto Map is an improvement over the Hénon Map. Recall the differences exposed by moving from the Baker's Map to the Baker's Apprentice Map, but we should not forget that neither is very realistic since the linearized dynamics hold at finite separations in both cases! Mathematical systems which are generated for their simplicity provide weak straw men. Perhaps the best straw man is a simulation model of the system. In this case the effect of varying sampling time, duration and noise level can be investigated in detail before the data itself is tainted. In all cases, aim for straw men of steel. It is a capital mistake to theorize without examining your algorithm's performance on realistic data of known origin.

4.3 Feasibility tests for the identification of chaos

We conclude with an approach for estimating a "number of data points" required for some specific application of a specific algorithm. Much has been written on how to properly analyse "small" data sets; we define a **tiny** data set as one which is too small to be properly analysed. Obviously, whether or not a given data set is tiny will depend both on the system which generated it and on the algorithm in question. The lower bounds discussed here aim to reflect necessary conditions, there is **no** implication that this amount of data is sufficient.

4.3.1 On detecting "tiny" data sets

In addition to testing new algorithms on data from simple nonlinear models, we can often use these systems to determine a lower bound on the amount of data one would expect an optimal algorithm to require. As a specific example, consider estimating the (local) linear propagator, a task which may be taken as a precondition for estimating Lyapunov exponents. Recall Figure 7 and the maximum linear range discussed in Section 2.1.3. About each point in state space, there is a sphere of initial conditions within which the exact linear propagator reflects the nonlinear dynamics to within some desired accuracy, say a one-step prediction error with a magnitude less than 5% the range of the data. Looking back at Figure 7, this just determines the largest circle such that points on the ellipse (the image of the circle under the linearized dynamics) fall within with a specific distance of their image under the full nonlinear dynamics. Since we are looking for a lower bound, we may assume that our algorithm

intuits the local linear structure perfectly given just one single observation[24] and ask: How long a string of observations is required before 99% of the initial conditions are predicted to within 5% ? Since the radius about each point will be greater than zero, the entire attractor can (usually) be covered in this way. The choice of 99% threshold is arbitrary, but the magnitude of the datasets required for any threshold in this ball park is enlightening.

For a 2-D map or a surface of section, it is straightforward to make a figure (not shown) illustrating the radius about each point along a trajectory. Plotting the fraction of the attractor covered as the length of the dataset increases is easy enough even in higher dimensional systems. The result is worth examining. Note that for reliable Lyapunov estimates, we must consider the error made in propagating the orientation of an infinitesimal under the estimated linearized dynamics in addition to the error made in propagating its magnitude. Similar programs can be constructed to test other algorithms. In this way we can gain insight into the minimal amount of data we expect to require, and can set realistic goals for the application of our algorithm.

[24] In practice, of course, more than one observation is required; the lower bound obtained in this section provides a necessary condition not a sufficient condition on the size of a data set.

5 Building Models Consistent with the Observations

An ideal model would provide an accurate representation of the system that generated the data. Imperfections in the structure of the model or inconsistencies between the model-state space and the system's state space can make this impossible. In this section we will investigate ways to evaluate imperfect models, and locate their deficiencies. When estimating the free parameters in a model, the optimal model parameters are usually defined as those which minimise a **cost function**, for example the root-mean-square (RMS) one-step prediction error[119]. Noting that relatively good models need not yield relatively small cost functions, we venture along other directions of model evaluation and improvement. Does a model trajectory exist which resembles the observations? Are prediction errors large when the predictability of the model is high? Are these errors distributed in a systematic manner in state space?

5.1 Cost functions

Attempts to minimise an RMS error are ubiquitous. In many applications where the errors are independent, Gaussian distributed random variables, it may even be optimal. But evaluation with an RMS error cost function will systematically reject perfect models of chaotic systems in favour of unphysical models with trivial asymptotic dynamics (relaxing either to a constant or to a periodic solution), when the error is evaluated after iterating a perfect model from an uncertain initial condition. In regions of state space where initial conditions with separations within the observational uncertainty diverge rapidly, a perfect model will be penalised even at short forecast times.

The Stilleto Map provides a simple example: suppose a trajectory is observed with about 2-digit accuracy in both x and y. Using these observations as initial conditions, the RMS prediction error in y of a perfect model increases with each iteration, until it exceeds the RMS error of simply predicting the mean values of x and y after a few (≈ 8) iterations. After this time horizon, an RMS cost function would reject the perfect model in favour of a model which predicted $(\langle x \rangle, \langle y \rangle)$, a point which need not be near the attractor. While this will in fact result in lower out-of-sample errors, it is much more difficult to improve a model with this kind of structure.

The analogous model in weather forecasting is one which "relaxes to the climatology." In the Stiletto Map, nonlinear noise reduction might break the analogy, but in practice, RMS cost functions are so deeply embedded in model fitting (when evaluating both data and estimating parameter values) that this may go unnoticed. What we would like is a method to evaluate the model,

given the uncertainty in the initial conditions. ι-shadowing provides a step in this direction.

The use of cost functions often assumes that the distribution of expected prediction errors is the same for all initial states, or at least known for each. When model error varies in state space, minimising a cost function may well distort a relatively good fit, in the attempt to reduce prediction errors which are unavoidable due to the structure of the model. This may reduce the similarity between model trajectories and system trajectories in regions where they might have been similar, while failing to introduce realistic looking model trajectories in the regions of high, if now somewhat moderated, prediction error. Variational data assimilation [120] schemes may degrade the accuracy of estimated states which are on trajectories which pass through regions of large model error, but which are not themselves within these regions. Shadowing trajectories simply stop when the trajectory enters such a region, thereby allowing better assimilation, when it is possible, and identifying the region of model failure for further study.

5.2 ι-shadowing: Is my model any good? (reprise)

A model **ι-shadows** the observations of a system if we can find a trajectory of the model which is consistent with all observations taken over the period in question. Obviously, a perfect model can always ι-shadow the observations, assuming we have a correct understanding of the observational noise. In fact, an entire family of imperfect models will do so as well, for any finite set of observations, and, without additional thought (or information), the choice between members of this family is arbitrary. In general they need not minimise any fixed step RMS cost function, and that is one of the reasons they are of interest to us here.

In Figure 14 the optimal ensemble forecast PDF quickly becomes highly structured, the evaluation of model quality through RMS error will reflect how well the model estimates the mean of this distribution. Minimum error criteria will select models which asymptote to the mean of the distribution of whatever quantity is being predicted, while all "realistic" models continue to oscillate. Averaged over predictions based on many different (uncertain) observations, this can lead to rejection of the perfect model in favour of the unphysical model: an RMS error criterion would reject the Moore-Spiegel equations as a good model of the Moore-Spiegel equations!

In practice, we have a series of uncertain observations of a system. In Figure 17 these are represented schematically by the circles distributed about the unknown system trajectory (dashed). Typically, low order nonlinear models are evaluated by starting on the initial observations and computing the RMS

prediction error as a function of forecast time; having taken observational uncertainty into account in tuning the model, we should take it into account in the initial condition as well. The figure shows the trajectory of an optimal RMS model (dash-dot). On average, this model will outperform a more realistic model if both are started from the current observation (or *analysis*, see Section 7.2 below). Does there exist another initial condition consistent with the current observation(s), for which the forecast of the realistic model remains within the uncertainty radius of future observations? In this case yes, the solid line. Of the two forecast systems, which is the better model of the system?

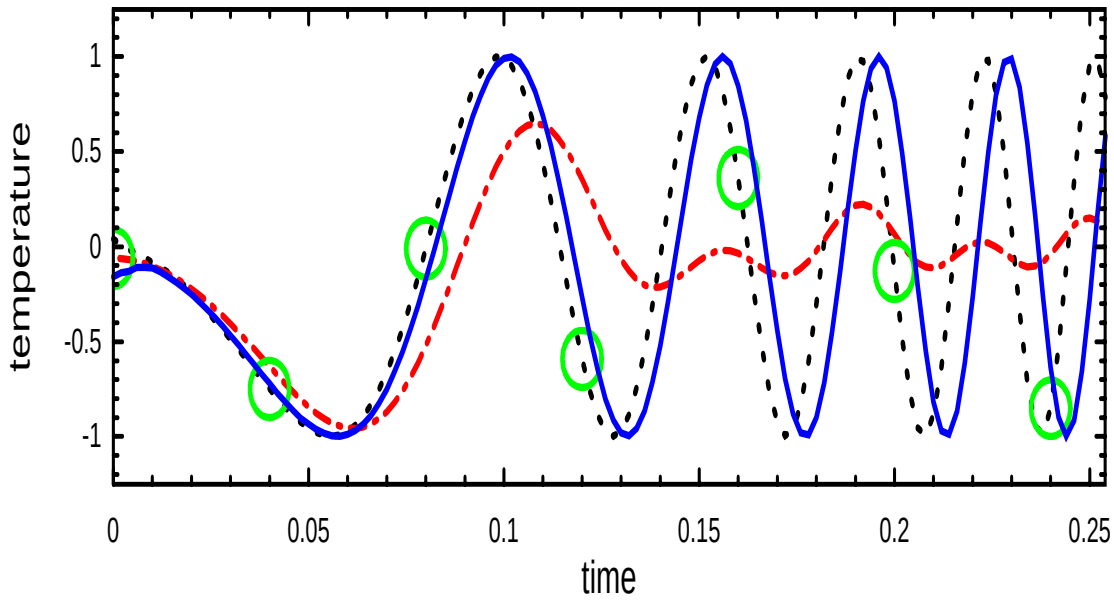

Figure 17: The shadowing dilemma. Given a physical trajectory (dashed), known only through uncertain observations (circles), an optimal RMS model (dot-dashed) which does not shadow the physical trajectory to within the observational uncertainty, and a realistic model trajectory (solid) which does shadow, but on average has a higher RMS error when *initiated* from the observations: Which is the better model of the system?

The ι-shadowing time [112, 101] is given by that initial condition consistent with the observational uncertainty which yields the model trajectory passing within the observational uncertainty of as many consecutive observations as possible. Contrasting the distribution of these ι-shadowing times (over different initial conditions) for each model provides a criteria for judging good models, longer shadowing times being preferred. Under this criteria the perfect model need never be rejected, while the optimal RMS predictor for a given observational uncertainty would be rejected early on.

A great deal of work has been done in hyperbolic dynamical systems[25] to es-

[25] An good introduction is provided in [76]. The term shadowing (and Theorem[121, 122])

tablish conditions under which it can be proven that there exists a trajectory of the *system* which shadows a trajectory of a particular model. We are interested in something altogether different, wishing to make no assumptions regarding the structure of the system (*e.g.* whether it is hyperbolic, or even deterministic) or the accuracy of the model. ι-shadowing considers the much more tractable question of whether we can locate a model trajectory which is consistent with the observations. The (unknown) properties of the system are not considered. Rather, we quantify how well imperfect models reflect the observed behaviour of physical systems given operational observational uncertainties: the shadowing of physical observations by imperfect models. This provides a new perspective for choosing between, refining, and combining the forecasts of distinct, operational nonlinear forecasting systems.

Inasmuch as determining the initial condition which will ι-shadow requires a knowledge of the future, real-time forecasts may not be improved. On the other hand, if the perturbations required to shift the current operational initial condition to the "model-correct" value contained some systematic structure, this structure could be employed to improve future forecasts.

5.2.1 Casting infinitely long shadows (out-of-sample)

Contrasting the distribution of shadowing times provides a useful approach for comparing different models and evaluating model "improvements" made using other cost functions. As a word of warning, we would note that one must take care with parameter selection via shadowing: there is one map which can shadow any set of observations.

Given a data set $s_i, i = 1, \ldots, N$ where the $s_i \in [0, 1]$ and are recorded to an accuracy of Q bits. The map

$$R(x) = 2^Q x \bmod 1 \tag{42}$$

will ι-shadow any set of observations! It will reproduce the first Q bits of each observation given the initial condition

$$x_0 = s_0 + 2^{-Q} s_1 + \ldots + 2^{-iQ} s_i + \ldots + 2^{-NQ} s_N \tag{43}$$

Equation 42 might be called the Russell Map, as it resembles an argument illustrating the futility of basing a definition for determinism upon an equation of the form $x(t) = f(x, t)$ [128]. It reflects a fundamental limit on the interpretation of shadowing, and nicely questions even (blind) out-of-sample evaluation!

is invoked in parameter estimation for *perfect* models (see, for example, Jansen and Kriegel [123] and references therein), and the evaluation of computer orbits [124, 125, 126, 127].

5.3 Distinguishing Model Error and System Sensitivity

Even if all prediction errors are equal, some are more erroneous than others. By observing where in state space the errors occur, we can identify (1) regions where the observed errors are at odds with the model sensitivity, and (2) regions of coherent residual predictability which indicate systematic model error.

5.3.1 Forecast Error and Model Sensitivity

In addition to attempting to minimise a cost function like the one-step prediction error, we can also consider where in state space they occur. If large errors only occur in regions where the *model* is sensitive to uncertainty in the initial state, then the model is at least internally-consistent. If these uncertainties are sufficiently small, their growth can be approximated through the Jacobian (or linear propagator) of the model; if not, then nonlinear ensembles are required. In either case, the basic idea is simple: if the model sensitivity reflects that of the system, then large forecast errors will only occur in regions where the model sensitivity is high. A large forecast error in such a region is unsurprising. If, on the other hand, a large forecast error occurs in a region where the model is relatively insensitive to the uncertainty in the initial state, then we have evidence for model error. For a given model, these are precisely those regions of state space in which we expect ι-shadowing to fail. Having identified such regions, the model may be refined either by including additional observations corresponding to these regions or by increasing the weight given to existing observations. This allows recursive improvement of an imperfect model, focussed on regions where its forecast errors are internally inconsistent. While the one-step prediction error will increase, shadowing times (and multiple-step prediction errors) may improve.

5.3.2 Accountability

Many models may have a trajectory consistent with the observations. But the majority of these ι-shadowing models will not have such trajectories in the correct proportions. While imperfect models will not be accountable, studying how given models fail can yield insight into how to improve them, and whether to consider ensembles over different models, as well as ensembles of different initial conditions.

5.3.3 Residual Predictability

Consistency tests on the time series of prediction errors, often called the *residuals*, provide an alternative to investigating model sensitivity. Ideally, the resid-

uals are independent, identically distributed (IID) random variables, or at least indistinguishable from a sample of IID random variables, and there are a number of more traditional [11] and more modern [129] tests to evaluate this null hypothesis. A widely applied test for IID series using the GPA has been developed by Brock *et al.* [130, 131]. Here we will consider a simpler test for residual predictability[90].

In the nonlinear prediction paradigm of Section 3, prediction in time becomes interpolation in model-state space. For each deterministic model, there will be a specific surface, optimal in the least square sense, which most closely matches the expected value of the process. Even with the additional assumptions that the process is deterministic, the embedding dimension is sufficient and the data are noise free, this surface defines a perfect model only if the expansion in model basis functions converges for a finite number of terms. The main point of this subsection is that the smooth variation of these two surfaces will result in systematic prediction error - correlated not in time but with location in model-state space. This is indicated in Figure 18 which shows the $x = 0$ slice of the prediction surface from an RBF model for the Stilleto Map along with that of the true map.

Assume that there is no residual predictability, that is, assume that the residuals are IID. Every IID sequence must remain IID under any blind rearrangement, an observation dating back, at least, to von Mises [132]. An example: given two IID variables, there is a 25% chance they will both be greater than the median of the distribution, a 25% chance they will both be less than the median, and a 50% chance one will be less and the other greater. The test of residual predictability simply finds nearest neighbour pairs of observations in the model state space, and tests whether the corresponding pairs of residuals are distributed in a manner consistent with the residuals being IID in the first place. If there are significant correlations, then the residuals were not IID, and additional predictability may be present. Tests for residual predictability are discussed at greater length in reference [94].

Figure 13 illustrates a simple test in 2-dimensions: at each point in the delay reconstruction, a '+' denotes a positive residual, while a '-' denotes a negative residual. Forming nearest neighbour pairs, one may easily evaluate the null hypothesis that these symbols are randomly assigned. In this case, the correlation is visible by eye, in many regions there are clusters of '+' symbols. In this case there is residual predictability. This is often the case, even when predicting relatively simple numerical systems, except in cases where, for example, when one attempts the equivalent of predicting the Hénon Map with a local quadratic map!

The test is easily generalised (1) to consider more than two points at a time, (2) to more than two types of error, and (3) to higher dimensions, where visual inspection may be difficult. The test also provides an indication of where in state

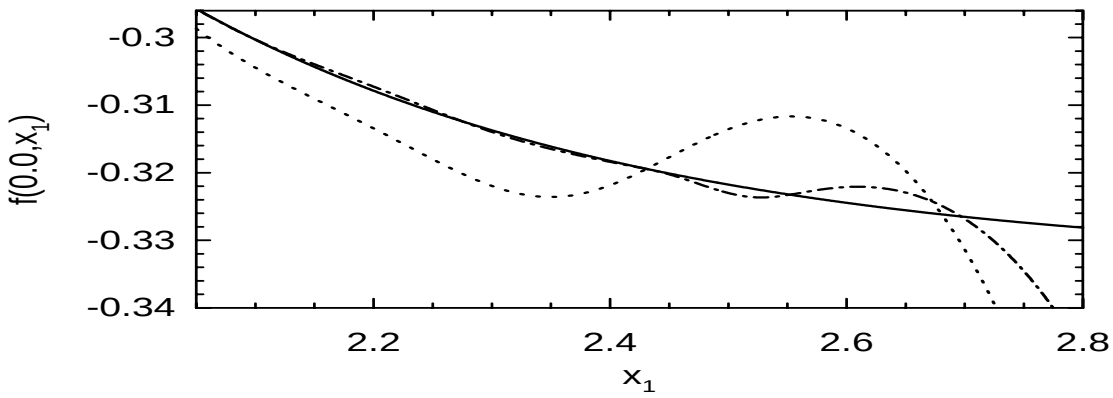

Figure 18: A one-dimensional slice at $x = 0$ of a two dimensional delay embedding of the Stiletto Map. The one step ahead prediction $f(x_0, x_1)$ is shown for $x_0 = 0$ as a function of x_1 for the the perfect model (solid) and two global RBF models using 64 centers (dotted) and 128 centers (dash-dotted).

space the model is systematically wrong. Such locations can be contrasted with regions where the model in unable to ι-shadow, or where the model sensitivity appears to be to small; this suggests a method of constructing an improved model by introducing additional freedom to the fit in these regions.

The residual predictability test can locate model error, even when one cannot reject the IID null hypothesis for the time-series of residuals on their own, because it includes information relative to their location on the prediction surface. Of course, the deterministic modelling paradigm assumes from the outset that such a surface exists. If the underlying process is stochastic, then this assumption fails. Testing for the existence of this surface with ensemble predictions provides a test for operational determinism, as discussed in the next section.

6 Deterministic or Stochastic Dynamics?

If the equivalent of the perpetual motion machine exists in time series analysis, it must be the algorithm which can distinguish deterministic dynamics and stochastic dynamics. Attempts abound.

Earman [133] provides a primer on determinism, while discussions within the context of nonlinear dynamics given in references [131, 78, 134]. In this section we will discuss the lesser goal of deciding whether a system is best modelled deterministically or stochastically given a particular data set. Hence our aim is to determine if a system is **operationally deterministic**, while remaining agnostic regarding its true nature (and ready to refine its operational status given more data or new insight).

To be useful, any test for operational determinism must be robust in the face of observational noise. For simplicity, we will limit the discussion to two one-dimensional systems, one deterministic and one stochastic. Contrast the deterministic system

$$\tilde{x}_{i+1} = \begin{cases} \frac{\tilde{x}_i}{a} & \text{for} \quad 0 \leq \tilde{x} \leq a \\ \\ \frac{1-\tilde{x}_i}{1-a} & \text{for} \quad a < \tilde{x} \leq 1 \end{cases} \tag{44}$$

with the stochastic first order autoregressive, or AR(1), system:

$$\tilde{y}_{i+1} = \alpha \tilde{y}_i + \gamma_i^{dyn} \tag{45}$$

where $\alpha = (2a - 1)$, $0 < a < 1$, γ_i^{dyn} is a normally distributed random variable with mean zero and standard deviation one and the superscript *dyn* stresses that this noise term influences the dynamics. It is interesting to note that, since the autocorrelation functions of $\tilde{x}$ and $\tilde{y}$ are identical, it is impossible to distinguish these series via *any* analysis which is based upon their autocorrelation function.

The dynamical system of Equation 44 is chaotic; at each iteration an infinitesimal perturbation will grow by a factor of either $\frac{1}{a}$ or $\frac{1}{(1-a)}$, "on average" by $e^{\Lambda \ln 2}$ with the Lyapunov exponent $\Lambda = -a \log_2(a) - (1-a) \log_2(1-a)$, although no actual perturbation ever grows by this amount (see Shaw [103]). Initial perturbations in the value of y, on the other hand, will *decrease with time* under the stochastic system of Equation 45, given the same realization of γ_i^{dyn}.

We adopt the same measurement function for each system and the same level of observational noise:

$$h(\mathbf{z}_i) = h(\mathbf{z}_i) + \nu_i \tag{46}$$

where the observational noise, ν_i, is a normally distributed random variable with mean zero and standard deviation σ_ν; $\mathbf{z}_i$ represents our state variable,

which for the moment is either $\tilde{x}$ or $\tilde{y}$. Equation 46 yields two series of observations: $x_i = h(\tilde{x}_i)$ and $y_i = h(\tilde{y}_i)$. Due to observational noise, exact prediction is impossible for either system. In both cases, as more and more observations become available, the optimal RMS error predictor should issue forecasts approaching the estimated local mean value of the conditional distribution $P(s_{i+1}|s_i)$, regardless of whether the process is stochastic or deterministic[26].

Deterministic ensemble forecasts strive to reflect this distribution. For an operationally deterministic process, this is exactly what is desired; for a stochastic process, the expected value may be defined quite precisely, and still bear no relation to the observed value. This is the distinction we aim to exploit.

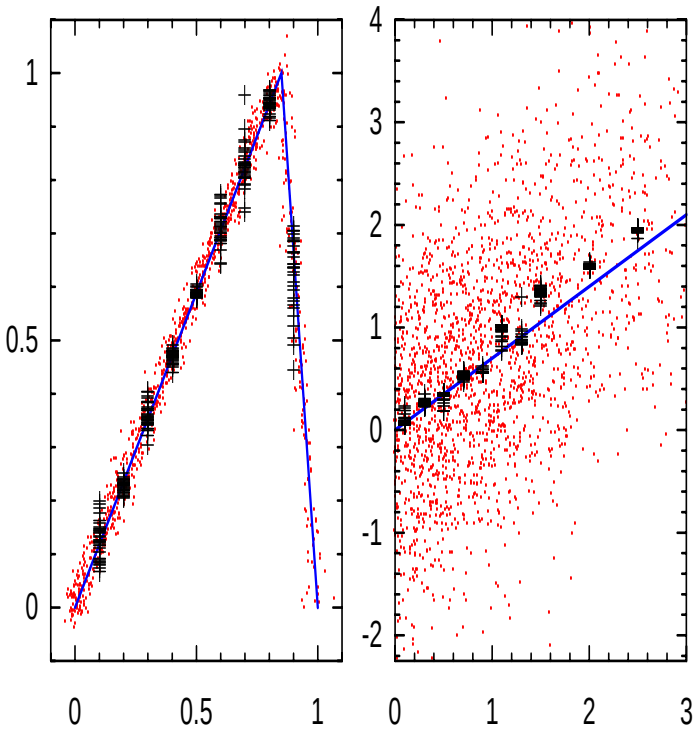

Figure 19: Observations (·), ensemble predictions (+), and expected values (solid line) for (a) the deterministic map of Equation (44) and (b) the stochastic map of Equation (45). Here $a = 0.85$. The ensembles give a rough indication of the likely distribution of values in the deterministic case, while this distribution is much to narrow in the stochastic case.

[26]For the response of an RBF model to data from an autocorrelated noise process see [97].

6.1 Using ensembles to distinguish the expectation from the expected

Delay plots for x and y are given in the left and right panels of Figure 19. Also shown are ensemble forecasts from a deterministic model. A local quadratic model was employed, in an attempt to reduce the chance of serendipitous synergy between a locally linear forecast model and either of these locally linear processes. Since the forecast model is deterministic, the ensemble spread reflects the uncertainty in the optimal RMS forecast due to observational uncertainty in the initial condition.

Knowing that the value of s_i is uncertain, the uncertainty in s_{i+1} is estimated simply by making ensemble forecasts under the deterministic model. When applied to deterministic data this should yield well calibrated forecasts, estimating the uncertainty in the expected value of s_{i+1}, and hence reflecting the distribution of likely observations. And in the stochastic case? In the stochastic case our ensemble forecast still returns the uncertainty in the expected value of s_{i+1}, but the uncertainty in the expected value is *not* the uncertainty in the future state: for a stochastic process, the expected value may be defined quite precisely and yet bear no relation to the observed value. To the extent that a perfect predictor can be approximated operationally, then in a deterministic system the uncertainty in the expected value is the uncertainty in the prediction. This is simply not the case in the stochastic system, where the (random) dynamics also makes a direct contribution to the distribution about the expected value. Contrasting the two panels of Figure 19, we see that the agreement between forecast spread and the spread in the observations correctly identifies the left panel as operationally deterministic.

If deterministic ensemble predictions are badly calibrated everywhere in state space, we adopt an operationally stochastic modelling strategy (*e.g.* SEQUIN or RAP). If the failures are localised, then we have a good indication as to where to look for model error, or physical arguments as to why those states of system may be extremely sensitive to external influences (or may be improperly embedded). If calibration failures are rare, but appear to be randomly distributed in state space, then we might argue that the system is deterministic, but not isolated: occasional external perturbations being responsible for the occasional lack of calibration.

Why might we wish to model a deterministic process as operationally stochastic? Ignorance. If we do not have enough data to fit a deterministic data-based model, and we do not understand the physics well enough to construct a simulation model, then a simple stochastic model may prove a fine choice. Consider some poorly understood chaotic phenomena: with very little data, a nonlinear, or locally linear [95], stochastic model proves best; as the duration of observation increases, we eventually observe a number of near returns in state space

sufficiently close (*e.g.* within the linear range) so as to make deterministic models optimal[27]. As still more data are obtained, the suite of deterministic forecast models improve until we hit forecast errors near the noise level of the observations; at this point we return to a stochastic model, although here we may need only assign the probability of the next observation falling into each of a few quantisation bins. A deterministic process is "operationally stochastic" either (1) when it is known so poorly that its general properties are uncertain or (2) when it is known so well that the forecasts are dominated by observational noise. Nonlinear noise reduction techniques [136] may reduce the observational noise, but we can never ascertain the ultimate noise-free dynamics from finite observational data.

We note in passing that the inhomogeneity common in chaotic dynamical systems suggests that even for a fixed prediction time, the same system may be best modelled by interpolation (*i.e.* deterministically) for most initial states, and by random analogues (*i.e.* stochastically) for other initial states. The state space may be mapped out and different schemes used in different regions, just as different delays were used for different initial conditions in reference [73]. This variation in the quality of a model with location can confound optimisation schemes based on minimising a cost function. In the Lorenz system, for example, predicting about one oscillation time ahead with ensembles with a diameter of ≈ 0.01, direct local linear prediction usually out-performs RAP given the same learning data, but the deterministic model goes badly wrong on a few occasions, which then dominate the total RMS error (and the ensemble calibration score) of the predictor. Such events contribute to the need for Olympic-style scoring rules (throwing out the worst 10% predictions when computing the mean forecast error) adopted by Casdagli[83] when documenting how forecast error scales with the size of the data set. The ability to foresee which initial conditions will yield these large forecast errors was noted by Casdagli; ensemble prediction provides additional information with which we may act on this foresight. Figures illustrating this effect will be presented elsewhere.

And if the underlying system is stochastic? Wind tunnel data reflecting the transition to boundary layer turbulence in the experiments of Gaster [137] provide an example. The data reflect fluid velocity in an open flow which is perturbed by a sinusoidal forcing of adjustable amplitude. When the forcing amplitude is small and the prediction time short, the velocity is almost periodic in time; deterministic models provide more refined ensemble forecasts than stochastic models, although both local linear ensembles and RAP ensembles are well calibrated. The calibration-refinement graphs for a RAP model are shown in Figure 20, local quadratic prediction based on the same learning set

[27]Of course, the time required to do this may be longer than the lifetime of the system. One estimate of the "return time" for the Earth's atmosphere [135] places it at 10^{30} years, significantly longer than the expected lifetime of the Universe.

yields slightly less calibrated but much more refined forecasts (more predictions near 100% and 0%), suggesting that this dataset be treated as operationally deterministic. (Calibration and refinement are introduced in Section 3.5.) For

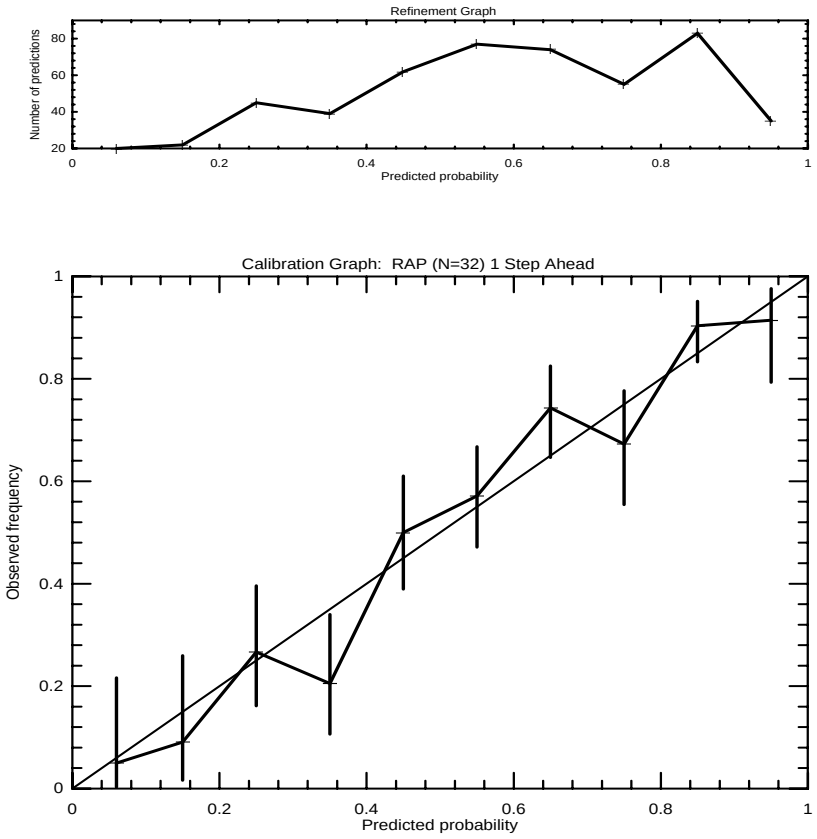

Figure 20: Calibration (lower) and refinement (upper) plots for the wind tunnel data at moderate forcing. The upper graph shows the number of predictions made clustered into bins 10% wide. The lower graph shows the relative frequency with which a given cluster of forecasts identified the outcome; if the vertical bars cross the diagonal, the observed relative frequency is consistent with the predicted probability, at the 95% confidence level. These results are for a RAP model.

a longer prediction time and larger forcing amplitude, RAP ensembles remain well calibrated, while the forecasts of all models become less refined (*i.e.* fewer forecasts are made where the PDF is well localised), and deterministic models are not well calibrated. What does this mean? RAP ensembles simply collect all relevant analogues. If the data set is large enough (*i.e.* if it contains relevant analogues), then RAP ensembles will be well calibrated for any stationary process. Deterministic forecasts assume that there is additional (accessible) information in the analogues; if this is indeed the case, then they will provide ensemble forecasts which are better refined than those of a stochastic model. That is the essence of operational determinism.

If deterministic ensemble forecasts give a good account of the observed forecast errors, then we can make a case for operational determinism. If they disappoint our expectation, "either not at all succeeding constantly or at least varying much" from what is expected, then the results should be contrasted with those of ensembles under stochastic models. Ideally one should make ensemble forecasts with both deterministic and stochastic models and judge which is the more accountable. In cases where a single forecast value is required, the choice is less straightforward.

Given a finite collection of observations, can we determine whether or not the underlying process is deterministic? In a word: No. The same process may appear either deterministic or stochastic, depending on how we observe it, and for how long. Can we decide on the best way to model a process, given a particular collection of observations? Perhaps.

7 Numerical Weather Prediction

7.1 Probabilistic Prediction with a Deterministic Model

Ensemble forecasting is now the operational standard at a number of Numerical Weather Prediction (NWP) centers world-wide. Yet even with ensemble forecasts, operational weather prediction remains difficult for at least three reasons:

- At some times, an ensemble forecast does not give a unique indication of what the weather will be;
- At some times, an ensemble forecast gives a unique, but incorrect, indication;
- And at *all* times, it requires 24 hours to verify each daily forecast.

When addressing these points, how are limited resources best distributed between addressing the first point by gathering more data for a better estimate of the initial conditions and addressing the second either by obtaining a larger computer and thus getting a better estimate of the forecast probability distribution if it is operationally accountable, or by improving the model itself if the distribution is unrealistic? These issues were discussed by Thompson in a 1957 paper entitled *Uncertainty of Initial State as a Factor in the Predictability of Large-scale Atmospheric Flow Patterns* [138], which evaluates the various factors that limit predictability and the marginal gain of improving each of them. Yet the third point is perhaps the most limiting, since arguably [135] the Earth shall not exist long enough for us to determine whether or not the weather is chaotic. The time required to obtain two observations sufficiently close that the linear approximation might be quantified may well exceed the expected lifetime of the Universe. Few can dismiss this as merely a technical constraint, although many might argue that the question is of only academic interest in that technically, chaos is irrelevant to operational forecasting inasmuch as it is defined via Lyapunov exponents.

In any case, there is no question that atmospheric dynamics are nonlinear. Applying the notions of nonlinear prediction and ensemble forecasting to the Earth's atmosphere only requires a generalisation from 3-dimensional dynamics to, say, 10^6 dimensional dynamics; similar techniques may work, but the technical (and technological) constraints of finite computers and real-time results require some consideration. Given unlimited computational power, we could adopt the approach used to form ensembles for the annulus in Section 3.7: simply sample the model-state space until a good representation of the

probability distribution of initial conditions consistent with the current observations was obtained. Then evolve each of these initial conditions under the model to obtain a forecast probability distribution at the final time.

There are a number of problems with this approach, even in principle. While our uncertainty may have a smooth distribution in model-state space, the relative likelihood of different initial conditions almost certainly does not. If the model evolves on an attractor, or even a smooth manifold of lower dimension than the model-state space, then we cannot assign the appropriate weight to each initial condition without some knowledge of this manifold. Even if our uncertainty varies smoothly in state space, the set of physically relevant states does not. The fact that the model-state space differs from that of the atmosphere only makes matters worse, a point to which we return below. In short, our ultimate ensemble forecast will not be accountable if the ensemble is chosen in this way.

In practice, the evaluation of such a large ensemble is not realistic. Operationally, one may only deploy ensembles of less than 10^2 members, in a model-state space with dimension of order 10^6. The state-of-the-art is reflected in the book *Predictability* [139]. Competing methods of ensemble formation are used in the operational forecast centers of the US, Europe and Canada. Section 7.3 provides a schematic description of the options available. First we note some of the boundary conditions on weather prediction.

7.2 The Analysis

In meteorology, an analysis is a point in model-state space. The analysis is our best guess at the model state which is in turn the best analogy to the state of the atmosphere. If we define an ideal model initial condition as that which will best ι-shadow future observations, then the analysis may be thought of as the best approximation to this model state, given the information available at the time it was constructed. Thus the analysis is a function of the time at which it was constructed as well as the time at which it represents the state of the system. If the model is not perfect, then the distinction between the different model states is important.

Technology also places a severe constraint on the effective spatial resolution of weather models if, as is often the case, we wish to have the forecast before the event. Operationally, this translates into an effective horizontal "grid scale," although in practice many of the calculations are done spectrally. Within the model, there is no spatial resolution below this scale, typically tens of kilometers, and all the physics at smaller length-scales must be parameterised. To the extent that this parameterisation is not exact[28], the weather model cannot be

[28] No parameterisation is exact[2].

perfect; if the subgridscale structure is important, then there will be a time-scale after which the model cannot reflect reality: weather models will have finite ι-shadowing times. Even if the ι-shadowing time-scale is long compared to forecast times, the coarse resolution of the model introduces another more immediate complication: there is no model equivalent of "the temperature at Heathrow." Indeed, Heathrow, Gatwick, and a large part of London may all lie within a single point. How then does one insert an observation into the model? Or evaluate a prediction?

While this point is particularly clear when the resolution of the model differs by several orders of magnitude from that of the measuring instrument, the same problem exists for almost all models: The state space in which the model evolves (*i.e.* the **model-state space**), is fundamentally different from the "true" state space of the atmosphere-ocean system within which we live and take measurements. The fact that this system is not isolated adds another dimension to the problem.

The model state which is judged to most closely correspond to a given set of observations (which may be distributed both in space and in time) defines the **analysis**; a huge effort has been put into determining the best analysis. In most cases, the analysis is based in part on predictions formulated from past observations; model predictions make a particularly large contribution to components for which few (or no) observations are available, for example small scales and particularly over the oceans. Since different models will have different model-state spaces, the analysis will depend on both the model and the observations.

In practice, the quality of a forecast is often evaluated relative to the corresponding analysis, rather than the observations themselves. While something of this sort is required, since the domain of the model differs from that of the observations, validation against observations is preferable. Interpreting the analysis as the target for a forecast introduces model error into the target, thereby complicating model evaluation: once the model has been used to validate the observations, it is less clear how to use the "improved" observations to validate or improve the model itself. The movement between "observations" and "analysis" is a field of its own (see Talagrand and Courtier [120] and references thereof); those interested in low-dimensional nonlinear dynamical systems can learn a good deal from what has been achieved (operationally!) within the meteorological community.

7.3 Constructing and Interpreting Ensembles

Constrained to an ensemble with a relatively small number ($\sim 2^5$) of members distributed in a relatively high dimensional ($\sim 2^{60}$) space: How are the best

perturbations to be chosen? This is a topic of international debate. The answers will vary with the goal of the forecast. In particular, they will vary with the relative weight given to the conflicting design goals of (a) enhancing the probability of detecting extreme events, and thereby providing better warning of the "worst case" scenario, (b) obtaining an ensemble member which ι-shadows the atmosphere for as long as possible, and (c) reflecting the true probabilities of different forecasts as accurately as possible. The interpretation of the ensemble will, of course, depend on the manner in which it was constructed. An unconstrained ensemble aims to reflect the true PDF of the model, while ensembles which selectively sample the more unstable directions often do so in an attempt to increase the variance of an ensemble with relatively few members.

Constrained ensembles are formed by restricting the ensemble members to a subspace of the full model-state space. Potential directions to be taken as components of this subspace include the orientations of:

1 Most likely static displacement, given the moments of the (local) set of physically relevant states.

2 Fastest growing infinitesimal displacement (instantaneous).

3 Local orientation of the globally fastest growing uncertainty (infinite past).

4 Fastest growing infinitesimal displacement (fixed finite future time).

5 First infinitesimal displacement past a threshold (variable finite time).

6 Most likely orientation given the variation in measurement accuracy (observational uncertainty).

7 Orientation in which the dynamics are worst represented (model error).

These seven options restrict orientation only, without suggesting a specific magnitude, and all these orientations will vary with location in model-state space. The first reflects the most likely direction due only to the local distribution of true potential initial conditions (physically relevant states). The second reflects the fastest growing direction(s) of the local Jacobian, the third the local orientation of the first global Lyapunov vector (assuming it exists), the fourth reflects the orientation of the first singular vector determined with an optimisation time which is independent of the initial condition. Option five also reflects a singular vector, but here the optimisation time is allowed to vary with location, reflecting a local τ_q where q is chosen to reflect the maximum magnitude for which the linearized dynamics are deemed to be relevant. The sixth option accounts for the fact that different components of the model-state vector are known with different accuracies, while the seventh reflects the desire

to account for model error, although it is not clear how introducing perturbations in these directions will accomplish that task.

Given an orientation, a magnitude must be selected, ideally this is done so as to produce an initial condition consistent with the long-term dynamics of the system ("on the attractor"). The basic difficulty is that while we can compute the probability of an observation $\tilde{\mathbf{x}}$ given both the true state $\mathbf{x}$ and the statistics of the observational uncertainty, we cannot compute the probability of the true state being $\mathbf{x}$ given only the observation and the noise process, since we do not know the local structure of the manifold upon which the best initial condition must lie. The selection of initial conditions consistent with the model is a nontrivial problem, even when the noise process is known.

Note that for options 2 through 5 to be of interest, not only must the models be fairly good, but the linearizations must agree at length-scales determined by the observational uncertainty. Recall Figure 7 of Section 2.1.3 and the maximum linear range, δ. Even in a perfect model, if the error in the initial condition exceeds the radius at which the linear approximation is accurate, then the linearized dynamics are irrelevant regardless of how they might be interpreted. For an imperfect model, we have the additional constraint that the linearization of the model be a good approximation of the linearization of the system. For the relevant time scale, we require (a) that the SVD of the model dynamics about the model trajectory is sufficiently similar to the SVD of the true dynamics about the systems trajectory, and (b) that the analysis (the model initial condition) lies within the the linear range of the model[29], and (c) that an ideal model initial condition exists and also lies within this radius of the analysis. We have responded to Maxwell's warning by assuming not only that the weather is amenable to a finite scheme of law, but that our approximation to this scheme is also an approximation of its first derivatives and our observational errors are sufficiently small that this linearization is relevant.

Ensemble forecasts are made routinely by meteorological centres around the world. At the European Centre for Medium-range Weather Forecasts (ECMWF), ensembles are formed based upon initial time Singular Vectors (SV) with a fixed optimisation time of about 2 days[140, 141, 142]. The American National Centers for Environmental Prediction (NCEP) employs ensembles based on Bred Vectors(BV), which reflect error growth over the recent past[143, 144, 145]. If the model was perfect and the observational uncertainty in the analysis was infinitesimal, then the bred vectors would converge toward the local orientation of the global Lyapunov Vectors (LV). Of course, neither condition is satisfied and the bred vectors also contain useful information on

[29]This can be verified to a limited extent by monitoring whether or not the nonlinear trajectory of each ensemble member still reflects the linear approximation at optimisation time. If a trajectory fails this test, it suggests itself as the source of a valid alternative linearization.

model error. Recent comparisons can be found in references [146, 147]. In the near future, the Japanese Meteorological Agency (JMA) plans to investigate bred vectors and singular vectors from the same model. The results should be enlightening. There are also attempts to construct ensembles without introducing dynamical constraints [148]. This field is evolving rapidly; two versions of the current state of play may be found at `http://www.ecmwf.int/` and `http://sgi62.wwb.noaa.gov:8080/ens/enshome.html` .

7.4 The outlook(s) for today

Once the members of an ensemble are agreed upon, each is evolved under the full nonlinear model, resulting in a collection of potential realizations of the weather over the following days or weeks. From the perspective of a week ago Wednesday, what weather was forecast for today? The ECMWF's answer to this question is illustrated in Figure 21. Here we have 33 forecast maps each giving a version of today's weather as determined from an initial condition consistent with the ECMWF analysis of the state of the atmosphere 10 days ago. Each of these "postage stamps" shows an equally valid prediction for a deterministic system; what can we learn from such an ensemble?

First, they are different: for this particular initial condition and initial uncertainty and model, this picture shows that the uncertainty in the 10 day forecast is large, even if the model was perfect. This simple fact alone is of great utility; if we look back to these 33 forecasts after only 3 days (not shown), we see they are in rather good agreement. In this way, the postage-stamp maps give a good indication of the *un*reliability of the forecast, and to a more limited extent, an estimate of its reliability. More detailed discussion can be found in *Predictability* [139] and the references therein.

7.5 Conclusion

How close are operational weather forecasting models to the ultimate limits of prediction? It would be interesting to contrast the distribution of the ι-shadowing times for operational NWP models. Observing significantly shorter shadowing times in lower resolution models would immediately address the dilemma of "higher resolution" versus "larger ensembles", in favour of obtaining higher resolution, although ensembles with (at least) as many members as those now in use will always be required simply to resolve the forecast PDF. In addition, projections of the "shadowing perturbations" into the constrained subspaces used for ensemble formation might help to resolve the questions as to how to form the best operational ensemble. In any event, the experiment would provide an estimate of the true limit of predictability of current operational models. It is all but inconceivable that either observational accuracy

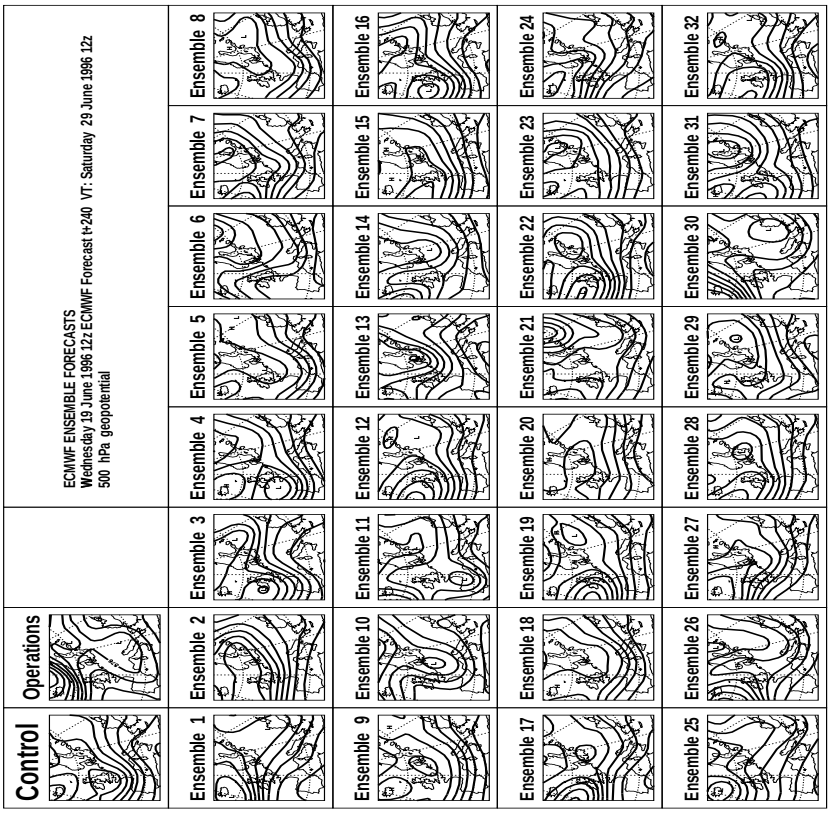

Figure 21: Today's forecast, illustrating the ECMWF ensemble over Europe after 10 days. Each small contour plot represents one member of the ensemble. In this case, the forecast PDF at this point in time is a distribution of these fields.

or modelling skill will ever reach a state where one would return to making a single deterministic forecast. Ensemble forecasts are here to stay.

The preferred strategy for ensemble formation remains an issue of debate even when the goal is agreed, and resolving the issue for the Earth's atmosphere may take some time. Undoubtedly some issues will be resolved, while others are advanced and refined until their proper resolution remains ambiguous. After all, we get only one 24-hour forecast a day, and consecutive days are highly correlated. It is at this point that laboratory systems like the rotating annulus come to the fore. Not only can we contrast ensembles based on any current (or future) formation scheme under a wide variety of different models, but sufficient amounts of data can be collected in order to evaluate these schemes. This statistical clarity is purchased at the price of considering an analogue system - from this analysis we can never be certain which approach is optimal for the Earth's atmosphere - but the relative strengths and weaknesses of competing ensemble prediction systems can be determined with a precision which will never be available for the atmosphere in any case. This insight

may shed light on the merits of various approaches to forecasting physical systems, clear of the statistical uncertainties which will remain in atmospheric prediction for much longer than the current crop of atmospheric models.

8 Summary

Quantifying uncertainty qualifies the value of scientific knowledge, whether the uncertainty lies in a forecast of the future state of a perfectly known dynamical system, an estimated scaling exponent, or today's weather from the perspective of 10 days past.

When quantifying predictability, the limitations imposed by considering only infinitesimal uncertainties in perfect models have led us to the use of ensembles to quantify the dynamics of finite uncertainties under the state-of-the-art forecast models in systems ranging from laboratory apparatus to the Earth's atmosphere. The insight that deterministic nonlinear systems require probabilistic forecasts is a major step forward in our understanding of predictability. Even in a perfect model, nonlinearity will tend to distort the distribution of uncertainty (whether it is originally Gaussian or otherwise). This non-Gaussian structure limits the applicability of least root-mean-square cost functions in defining and identifying the best model. The distribution of ι-shadowing times offers an alternative approach for contrasting skill between models.

No general "Limit of Predictability" time-scale can be derived from the dynamics of infinitesimal uncertainties since an initial uncertainty must grow to macroscopic scales before it is of practical importance. Thus the predictability-horizons of a chaotic system may differ greatly from any time-scale defined via Lyapunov exponents, in either their global or their finite-time incarnations. Even within the infinitesimal range to which their relevance is restricted, Lyapunov exponents reflect only an *effective* rate defined over a fixed time. Uncertainty $q-$pling times lift these constraints, but the average τ_q suffers from the inhomogeneity of uncertainty growth in state space. This inhomogeneity restricts the utility of *any* "Limit of Predictability." Nevertheless, chaos can be arbitrarily predictable as illustrated by the Baker's Apprentice Maps.

In the absence of any information on the state of the system, the best ensemble forecast for every initial condition is given by the climatological distribution[30], $\psi_\infty(x)$. While an ensemble forecast *can* contain usable information as long as it is distinguishable from $\psi_\infty(x)$, the question of whether or not a prediction is "useful" depends not only upon the forecast but also upon the goals of the user of that forecast. Once the image of the initial ensemble remains indistinguishable from $\psi_\infty(x)$, then the forecast is definitely "useless". The onset of uselessness poses a limit to predictability.

Given a perfect model and any specific ensemble, a well defined limit to predictability occurs when the ensemble can no longer be distinguished from $\psi_\infty(x)$. This limit is then a function of both the size of the ensemble and

[30]This is simply the projection of the invariant measure under the measurement function, to the extent that it is both well-defined and known.

the level of observational noise, as well as the initial condition and the system in question. Under a perfect model, ensemble forecast PDFs will approach $\psi_\infty(x)$ as $t \to \infty$, while under imperfect models they need never approach the correct asymptotic distribution. And imperfect ensembles need not do so accountably, even when evolved under perfect models. A limit of predictability applicable to both perfect and imperfect models can be constructed in terms of the distribution of ι-shadowing times. In this sense, the only limit on a perfect model arises from using finite ensembles.

Simple nonlinear systems may be employed to gain strategic insight into the methods used in physics-based "full" simulations (kitchen sink models). Ensemble forecasting and formation has been illustrated both in perfect model experiments with low dimensional systems, and in imperfect model experiments with laboratory systems of thermal convection in the rotating fluid annulus and boundary layer turbulence in the wind tunnel. In an environment which is data-rich and either knowledge poor or model poor, simple nonlinear models can contribute directly to our understanding of physical phenomena. Interesting examples include very short-term predictions of the surface temperature in Berlin in the range from 3 to 21 hours (see Ziehmann [25]) and very long term predictions of Ice Ages through ice volume (see Casdagli *et al.* [3]). It remains to test these results with ensemble forecasts using the same nonlinear models: if the results withstand this reasonably independent verification, they hold important implications for investigators attempting the construction of simulation models from first principles.

Most of the complications nonlinearity brings to analysis and forecasting have been illustrated to occur *even in* simple low dimensional examples; it will be interesting to learn which, if any, happen *only in* low dimensional systems. In terms of low dimensional data-driven models, it is often said that the predictions of these models are of limited interest because they "contain no physics." I would argue that they contain too much in exactly the same manner that a photograph of any particular bird may be considered an inferior representation of that species of bird when compared to the corresponding drawing by Audubon. The ultimate radial basis function model of the annulus might well be expected to out-perform a computational fluid dynamics code, as the RBF model would take into account imperfections in the Oxford annulus, and so on. The analogy would be too tight: while superb as a forecast model, as an analogy it may fail to generalise to other parameter values, much less other annuli. The dilemma becomes a value judgement between defining a good prediction either as a good forecast or as a reliable prophecy.

It is true, if tautological, that like all analogies the Laws of Physics are reliable guides as long as they are applied within their range of validity. This makes them extremely useful both pedagogically and in engineering applications. But when exploring the frontiers of science, we never know if we are within that

"range of validity" or beyond it: the success of Newton's laws in predicting the planet Neptune from observations of the perturbations from the Newtonian orbit of Uranus is often cited; the fact that Newton's Laws were invoked to support the discovery of the planet Vulcan - thereby accounting for perturbations from the Newtonian orbit of Mercury remains largely unremarked. Vulcan was, perhaps, a misinterpreted sunspot.

A general lack of faith in the existence of a large "range of validity" for data-based models is an asset in terms of maintaining our uncertainty in their results; even more so if it calls into question the range of validity of kitchen sink models "far" from the conditions for which they were tuned. Studies of the uncertainty in and limitations of climate models are underway [149].

The chance of reporting fortuitous results which, upon further trial, disappoint our expectations in that their implications of, say, predictability do not hold may be reduced by asking: "Would this analysis technique yield similar evidence even if the data had come from an uninteresting system?" We have seen that, to a limited extent, the probability of this happening can be quantified through the use of surrogate data. But the design of a relevant null-hypothesis to be rejected (*i.e.* the set of uninteresting systems) is non-trivial, and estimation of true significance levels is difficult. One can always perform multiple analyses on the same data set, but the statistical significance of the result does not increase very quickly if the tests are in any way related, regardless of whether or not we can quantify the extent to which they are related. Worse yet, appropriate significance levels may be inaccessible. Taking the series of Figure 1 as two randomly selected series of simultaneous measurements, there is without doubt a statistically significant relationship between the number of sunspots and the Republican fraction of the US Senate. One would be hard pressed to deny the possibility that some connection existed; and scientifically, it is much more interesting to "identify" a physical mechanism, even *a posteriori* , than to remain silent having failed to reject the null hypothesis. These two time series were not, however, selected at random: the time series from the Senate was chosen from among untold others *because* of this correlation, and thus appropriate significance levels cannot be determined. For these reasons, among others, it is crucial to think before running hard-won data through a black box. Data are only out-of-sample data once, and *a priori* hypothesis are much easier to evaluate. Luckily, in the case of the Senate the penultimate test of out-of-sample verification can be applied. Figure 22 shows the same two time series from 1900 until 1989, the correlation is less striking. We conclude that the apparently interesting results drawn from Figure 1 are, upon informed reflection, not so interesting. And so we look for others.

A collection of methods designed to identify the drawing of unwarranted conclusions from the nonlinear analysis of nonlinear data has been considered. In large part, these simply correspond to adopting good statistical practice. We

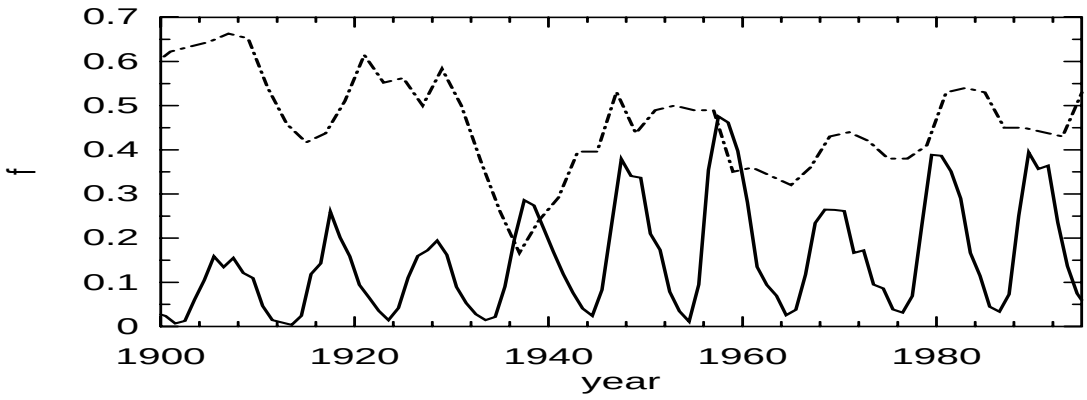

Figure 22: Simultaneous series of sunspot number (solid) and the fraction of the U.S. Senators who were Republicans (dot-dashed) on the day of their election, for the period 1900 to 1989. Note that these figures give the party division after the election that preceded each Congress, and does not reflect changes in party ratios that may have resulted from the death, resignation, or change in party affiliation of one or more senators within a Congress.

can use surrogate data to see how easily a desired result would be generated by chance. And we can employ straw men of steel to test our algorithms in hopes of identifying their limitations, before applying them in the analysis of data. And we can test the internal consistency of the failures of our models, out-of-sample. But front-line research, by design, occurs where our understanding is the most uncertain. Our best protection is to accept this fact, look closely for inconsistency and new data, and maintain our uncertainty whenever humanly possible.

Acknowledgements

Many of the results reported here, as well as the completion of these notes, were obtained due to the support of a Senior Research Fellowship at Pembroke College, Oxford. I would like to thank A. Provenzale and G. Cini for their invitation to present these lectures, and for their patience with the written version. My understanding of the results discussed herein continues to benefit from joint work with C. Ziehmann, E. Spiegel, and M. Allen. I grateful to M. Ghil for his continuing assistance in sharpening my arguments, while numerous discussions with D. Broomhead, J. Kurths, M. Muldoon, J. Stark and J. Theiler has improved them further. As did the numerous insightful criticisms I. Gilmour provided on a much rougher draft. I would like to thank M. Harrison, Y. Hayashi, T. Palmer, and Z. Toth for introducing me to the variety of alternatives available in making ensemble forecasts in numerical weather prediction. It is with pleasure that I acknowledge the assistance of J.A.M. Quatannens, Assistant Historian of the Senate Historical Office, for the timely provision of data reflecting the composition of the Senate. I learned of these series from David Wark, who supplied a graph covering the period 1960 to 1980. Finally, I would like to thank the students; I have benefited from their questions, and by putting these in writing they helped me see why I had not provided comprehensible answers. I hope this attempt comes closer.

References

[1] E. A Spiegel and A. Wolf. Chaos and the solar cycle. In *Chaos in Astrophysics*, volume 497 of *Annals of the New York Academy of Science*, pages 55–60, 1987.

[2] N. O. Weiss. Periodicity and aperiodicity in solar magnetic activity. *Phil. Trans. R. Soc. Lond.*, A 330:617–625, 1990.

[3] M. Casdagli, D. Des Jardins, S. Eubank, J.D.Farmer, J. Gibson, N. Hunter, and J. Theiler. Nonlinear modeling of chaotic time series:Theory and applications. page 335, 1992.

[4] H. T. Stetson. *Sunspots and their Effects.* McGraw-Hill, London, 1937.

[5] J. A. Barnes, H. H. Sargent, and P. V. Tryon. Sunspot cycle simulation using random noise. In R.O. Pepin, J.A. Eddy, and R.B. Merrill, editors, *The Ancient Sun*, pages 159–163, New York, 1980. Pergamon.

[6] D. S. Broomhead and G. King. Extracting qualitative dynamics from experimental data. *Physica D*, 20:217–236, 1986.

[7] R. Vautard and M. Ghil. Singular Spectrum Analysis in nonlinear dynamics with applications to paleoclimatic time series. *Physica D*, 35:395–424, 1989.

[8] M. R. Allen and L. A. Smith. Monte Carlo SSA: Detecting irregular oscillations in the presence of coloured noise. *J. Climate*, 9:3373–3404, 1996.

[9] J. Stark and B.V. Arumugam. Extracting slowly varying signals from chaotic background. *Int. J. Bif. and Chaos*, 1992. to appear.

[10] D.S. Broomhead, J.P. Huke, , and M.A.S. Potts. Cancelling deterministic noise by constructing nonlinear inverses to linear filters. *Physica D*, 89(3–4):439–458, 1996. Preprint.

[11] C. Chatfield. *The analysis of time series.* Chapman and Hall, London, fourth edition, 1989.

[12] H. Tong. *Non-Linear Time Series Analysis.* Oxford Univ. Press, Oxford, 1990.

[13] L. A. Smith. Local optimal prediction. *Phil. Trans. R. Soc. Lond.*, A 348:371–381, 1994.

[14] L. Borland. Simultaneous modeling of nonlinear deterministic and stochastic dynamics. *Physica D*, 99:175–190, 1996.

[15] F. Paparella, A. Provenzale, L.A. Smith, C. Taricco, and R. Vio. Local random analogue prediction of nonlinear processes. *Phys Lett A.* in review.

[16] D. W. Moore and E. A Spiegel. A thermally excited nonlinear oscillator. *Astrophys. J.*, 143(3):871–887, 1966.

[17] J.C. Sprott. Some simple chaotic flows. *Physical Review E*, 50(2):R647–R650, 1994.

[18] M. Hénon. On the numerical computation of poincare maps. *Physica*, 5 D(2–3):412–414, 1982.

[19] M. Hénon. A two-dimensional mapping with a strange attractor. *Commun. Math. Phys.*, 50:69, 1976.

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

[21] G. Strang. *Linear algebra and its application.* Hartcourt Brace Jovanovich, San Diego, 1988.

[22] L.A. Smith, C. Ziehmann, and K. Fraedrich. Uncertainty dynamics and predictability in chaotic systems. *Q.J.R. Meteorol. Soc.* in review.

[23] B.F. Farrell. Small error dynamics and the predictability of atmospheric flows. *Journal of the Atmospheric Sciences*, 47:2409–2416, 1990.

[24] B.F. Farrell and P.J. Ioannou. Generalized stability theory part i: Autonomous operators. *Journal of the Atmospheric Sciences*, 53:2025–2040, 1996.

[25] C. Ziehmann-Schlumbohm. *Vorhersagestudien in chaotischen Systemen und in der Praxis.* PhD thesis, Freie Universität Berlin. Meteorologische Abhandlungen. Neue Folge Serie A. Band 8 Heft 3, 1994.

[26] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipies.* Cambridge University Press, Cambridge, 1987.

[27] H. Mukougawa, M. Kimoto, and S. Yoden. A relationship between local error growth and quasi-stationary states: Case study in the Lorenz system. *Journal of the Atmospheric Sciences*, 48:1231–1237, 1991.

[28] J.M. Nese. Quantifying local predictability in phase space. *Physica D*, 35:237–250, 1989.

[29] E.N. Lorenz. A study of the predictability of a 28-variable atmospheric model. *Tellus*, 17:321–333, 1965.

[30] V.I. Oseledec. A multiplicative ergodic theorem. Ljapunov characteristic numbers for dynamical systems. *Transactions of the Moscoe Mathematical Society*, 19:197–231, 1968.

[31] H.D.I. Abarbanel, R. Brown, and M. B. Kennel. Variation of lyapunov exponents on a strange attractor. *Journal of Nonlinear Science*, 1:175–199, 1991.

[32] A.G. Darbyshire and D.S. Broomhead. Robust estimation of tangent maps and liapunov spectra. *Physica D*, 89(3–4):287–305, 1996. Preprint.

[33] L.A. Smith, C. Ziehmann, and J. Kurths. Estimating lyapunov exponents with uncertainty. in preparation.

[34] G. Nicolis. *Introduction to Nonlinear Science.* Cambridge University Press, Cambridge, 1995.

[35] J.M. Greene and J.-S. Kim. The calculation of Lyapunov spectra. *Physica*, 24 D:213–225, 1987.

[36] J. Barkmeijer. Approximating dominant eigenvalues and eigenvectors of the local forecast error matrix. *Tellus*, 47A:495–501, 1995.

[37] Halmos. *Lectures on Ergodic Theory.* Chelsea, New York, 1956.

[38] H. Poincaré. *The Foundations of Science.* Unversity Press of America, New York, 1982.

[39] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Physica*, 7D:153–180, 1983.

[40] L. A. Smith. *Lacunarity and Chaos in Nature.* PhD thesis, Columbia University, 1988. 263 pages.

[41] J. Theiler. Estimating fractal dimension. *J. Opt. Soc. Am.*, A7:1055–1073, 1990.

[42] D. Ruelle. *Chaotic Evolution and Strange Attractors.* Cambridge University Press, Cambridge, 1989.

[43] L. A. Smith. Intrinsic limits on dimension calculations. *Phys. Lett. A*, 133:283, 1988.

[44] P. Grassberger and I. Procaccia. Estimation of the kolmogorov entropy from a chaotic signal. *Physical Review A*, 28(4):2591–2593, October 1983.

[45] P. Grassberger, T. Schreiber, and C. Schaffrath. Non-linear time sequence analysis. *Int. J. Bif. and Chaos*, 1:521–547, 1991.

[46] L.M. Berliner. Statistics, probability and chaos. *Statistical Science*, 7(1):69–122, 1992.

[47] J. Theiler. Spurious dimension from correlation algorithms applied to limited time-series data. *Phys. Rev. A*, 34(3):2427–2432, 1986.

[48] A. Provenzale, L. A. Smith, R. Vio, and G. Murante. Distinguishing between low-dimensional dynamics and randomness in measured time series. *Physica D*, 58, 1992.

[49] F. Takens. On the numerical determination of the dimension of an attractor. volume 1125 of *Lecture Notes in Mathematics*, Berlin, 1985. Springer-Verlag.

[50] J. Theiler. Lacunarity in a best estimator of fractal dimension. *Phys. Lett.*, A 133:195–200, 1988.

[51] L. A. Smith. Discusion of the paper by prof r. smith : Estimating dimension in noisy chaotic time series. *J. R. Statist. Soc. B*, 54(2):456–458, 1992.

[52] A. R. Osborne and A. Provenzale. Finite correlation dimension for stochastic systems with power-law spectra. *Physica D*, 35:357–381, 1989.

[53] J. Theiler. Some comments on the correlation dimension of $1/f^{\alpha}$ noise. *Phys. Let. A*, 155:480–493, 1991.

[54] C.D. Cutler. A theory of correlation dimension for stationary time series. In *Chaos and Forecasting: Proceedings of the Royal Society.* World Scientific. To appear.

[55] D. Ruelle. Deterministic chaos: the science and the fiction. *Proc. R. Soc. Lond. A*, 427:241–248, 1990.

[56] C. Essex and M. A. H. Nerenberg. Comments on "Deterministic chaos" by D. Ruelle. *Proc. R. Soc. Lond. A*, 435:287–292, 1991.

[57] K. Judd. An improved estimator of dimension and some comments on providing confidence intervals. *Physica*, D 56:216–228, 1992.

[58] H. Isliker. A scaling test for correlation dimensions. *Phys. Lett.*, 169 A(5):313–322, 1992.

[59] D. S. Broomhead and R. Jones. Time-series analysis. *Proc. R. Soc. Lond.*, 423:103–121, 1989.

[60] R. Badii and A. Politi. Intrinsic oscillations in measuring the fractal dimension. *Phys. Lett. A*, 104(6,7):303–305, 1984.

[61] L. A. Smith, J.-D. Fournier, and E. A Spiegel. Lacunarity and intermittency in fluid turbulence. *Phys. Lett. A*, 114:465, 1986.

[62] G. M. Zaslavski. The simplest case of a strange attractor. *Phys. Lett.*, 69 A:145, 1978.

[63] P. Grassberger and I. Procaccia. Characterization of strange attractors. *Phys. Rev. Lett.*, 50:346, 1983.

[64] P. Grassberger. Finite sample corrections to entropy and dimension estimates. 128:369–373, 1988.

[65] F. Takens. Detecting non-linearities in stationary time series. Preprint.

[66] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65:579–616, 1991.

[67] U. Parlitz. Identification of true and spurious lyapunov exponents from time series. *Int. J. Bif. and Chaos*, 2(1):155, March 1992.

[68] N.H. Packard, J.P.Cruchfield, J. D. Farmer, and R.S. Shaw. Geometry from a time series. *Phys. Rev. Lett.*, 45:712, 1980.

[69] F. Takens. Detecting strange attractors in fluid turbulence. In D. Rand and L.-S. Young, editors, *Dynamical Systems and Turbulence*, volume 898, page 366, New York, 1981. Springer-Verlag.

[70] Th. Buzug, T. Reimers, and G. Pfister. Optimal reconstruction of strange attractor from purely geometric arguments. *Europhys. Lett.*, 13(7):605–610, 1990.

[71] M. Casdagli, S. Eubank, J.D. Farmer, and J. Gibson. State space reconstruction in the presence of noise. *Physica D*, 51:52–98, 1991.

[72] Fraser A. M. Reconstruction attractors from scalor time series. *Physica D*, 34:391–404, 1989.

[73] L. A. Smith. Does a meeting in Santa Fe imply chaos? In A. Weigend and N. Gershenfeld, editors, *Predicting the Future and Understanding the Past: A Comparison of Approaches*, volume XV of *SFI Studies in Complexity*, pages 323–344, New York, 1993. Addison-Wesley.

[74] D. S. Broomhead and G. P. King. Extracting qualitative dynamics from experimental data. *Physica D*, 20:217–236, 1986.

[75] G. King, R. Jones, and D.S. Broomhead. Phase portraits from a time series: a singular system approach. *Nuclear Physics B (Proc. Suppl.)*, 2:379, 1987.

[76] J.D. Farmer and J.J. Sidorowich. Optimal shadowing and noise-reduction. *Physica*, 47(3):373–392, 1991.

[77] J. Stark, D.S. Broomhead, M.E. Davies, and J. Huke. Theorems for forced and stochastic systems. In *Proceedings of the 2nd World Congess of Nonlinear Analysts, Athens, Greece*, 1996.

[78] S. Eubank and D. Farmer. An introduction to chaos and randomness. In E. Jen, editor, *Proc. SFI Summer School*. Addison-Wesley, 1990.

[79] A. Juneja, D.P. Lanthrop, K. R. Sreenivasan, and G. Stolovitzky. Synthetic turbulence. *Phys. Rev. E*, 49(6):5179–5194, 1994.

[80] J. Crutchfield and B. S. McNamara. Equations of motion from a data series. *J. Complex Systems*, 1:417–452, 1987.

[81] J.D. Farmer and J. Sidorowich. Predicting chaotic time series. *Phys. Rev. Lett.*, 59:8, 1987.

[82] D. S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *J. Complex Systems*, 2:321–355, 1988.

[83] M. Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35:335–356, 1989.

[84] A. I. Mees. Modelling complex systems. In L. S. Jennings, A. I. Mees, and T. L. Vincent, editors, *Proceedings of the Conference on Modelling Complex Systems*, Boston, 1989. Birkhauser.

[85] N. H. Packard. A genetic learnign algorithm for the analysis of complex data. *J. Complex Systems*, 4:543, 1990.

[86] L. A. Smith. Quantifying chaos with predictive flows and maps: Locating unstable periodic orbits. In N.B.Abraham, A.M. Albano, A. P. Passamante, and R. E. Rapp, editors, *Measures of Complexity and Chaos*, NATO ASI Series. Plenum Press, 1990.

[87] G. Sugihara and R. M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in a time series. *Nature*, 344:734–741, 1990.

[88] A. S. Weigend, B. A. Huberman, and D. E. Rumelhart. Predicting the future: A connectionist approach. *Int. J. Neural Systems*, 1:193–209, 1990.

[89] M. Casdagli. Chaos and deterministic versus stochastic non-linear modeling. *J. R. Statist. Soc. B*, 54(2):303–328, 1992.

[90] L.A. Smith. Local optimal prediction: Exploiting strangeness and the variation of sensitivity to initial condition. *Phil Trans R Soc Lond A*, 348(1688):371–381, 1994.

[91] E. N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, 26:636–646, 1969.

[92] E. J. Kostelick and D. P. Lanthrop. The prediction of chaotic time series: a variation on the method of analogues. In A. Weigend and N. Gershenfeld, editors, *Predicting the Future and Understanding the Past: A Comparison of Approaches*, volume XV of *SFI Studies in Complexity*, New York, 1993. Addison-Wesley. to appear.

[93] M. Casdagli. In A. Weigend and N. Gershenfeld, editors, *Predicting the Future and Understanding the Past: A Comparison of Approaches*, volume XV of *SFI Studies in Complexity*, New York, 1993. Addison-Wesley. to appear.

[94] L.A. Smith. Locally optimized prediction of nonlinear systems: Stochastic and deterministic. In H. Tong, editor, *Chaos and Forecasting*, volume 2 of *Nonlinear Time Series and Chaos*, pages 87–108, London, 1995. World Scientific.

[95] M.B. Priestley. State-dependent models: A general approach to nonlinear time series analysis. *Journal of Time Series Analysis*, 1(1):47–71, 1980.

[96] C. Chatfield. Nueral networks. *International journal of forecasting*, 4, 199.

[97] L. A. Smith. Identification and prediction of low-dimensional dynamics. *Physica D*, 58:50–76, 1992.

[98] M.J.D. Powell. Radial basis functions for multivariate interpolation: a review. In *IMA conference on "Algorithms for the Approximation of Functions and Data"*. RMCS Shrivenham, 1985.

[99] J. Stark. Recursive prediction of chaotic time series. *J. Nonlinear Sci.*, 3:197–223, 1993.

[100] K. R. Popper. *The Open Universe*. Routledge, New York, 1982. Accountability is defined on page 12.

[101] L. A. Smith. Accountability in ensemble prediction. In T. Palmer, editor, *Predictabilty*, Reading, UK, 1996. ECMWF.

[102] L. A. Smith. Visualising predictability with chaotic ensembles. In F.T. Luk, editor, *Advanced Signal Processing: Algorithms, Architectures and Implementations*, volume 2296, pages 293–304, Bellingham, WA, 1994. SPIE.

[103] R. Shaw. Strange attractors, chaotic behavior, and information flow. *Zeitschrift für Naturforschung*, 36 a:80–112, 1981.

[104] Allan H. Murphy and R. L. Winkler. Diagnostic verification of probability forecasts. *Intern. J. of Forecasting*, 7:435–455, 1992.

[105] A.H. Murphy. What is a good forecast? an essay on the nature of goodness in weather forecasting. 1993.

[106] Sir Athur Conan Doyle. *The Penguin Complete Sherlock Holmes*. Penguin, London, 1930. pg 163.

[107] P. Read, M. J. Bell, D. W. Johnson, and R. M. Small. Quasi-periodic and chaotic flow regimes in a thermally driven, rotating fluid annulus. *J. Fluid Mech.*, 238:599–632, 1992.

[108] D. Draper. Assessment and propagation of model uncertainty. *J. Roy. Stat. Soc.*, 57(1):45–97, 1995.

[109] R. L. Smith. Estimating dimension in noisy chaotic time series. *J. R. Statist. Soc. B*, 54(2):329–352, 1992.

[110] J.L. Anderson and H.M. van den Dool. Skill and return of skill in dynamic extended-range forecasts. *Monthly Weather Review*, 122:507–516, March 1994.

[111] H. Tong and R. Moeanaddin. On multi-step non-linear least squares prediction. *The Statistician*, 37:101–110, 1988.

[112] I. Gilmour. Contrasting erroneous models given uncertain data. Master's thesis, University of Oxford, 1996. Transfer of Status Thesis.

[113] J. Theiler, S. Eubank, A. Longtin, B. Galdrikan, and J. D. Farmer. Testing for nonlinearity in time series: The method of surrogate data. *Physica D*, 58:77–94, 1992.

[114] J. Theiler and D. Prichard. Constrained-realization monte-carlo method for hypothesis testing. *Physica D*, 94:221–235, 1996.

[115] A. R. Osborne, A. D. Kirwan, A. Provenzale, and L. Bergamasco. A search for chaotic behavior in large and mesoscale motions in the pacific ocean. *Physica D*, 23:75–83, 1986.

[116] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistices and Applied Probability*. Chapman & Hall, New York, 1993.

[117] B. Efron and R. Tsibirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, 1:54–77, 1986.

[118] B. Efron and R. Tsibirani. Cross validation and the bootstrap. 1997.

[119] C. Chatfield. Apples, oranges and mean squared error. *International journal of forecasting*, 4:515–518, 1988.

[120] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation: Part i: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113:1311–1328, 1987.

[121] D.V. Anosov. Geodesic flows and closed Riemannian manifolds with negative curvature. *Proc. Steklov Inst. Math.*, 90, 1967.

[122] R. Bowen. ω-limit sets for axiom A diffeomorphisms. *J. Differential Equations*, 18:333–339, 1975.

[123] W. Jansen and U. Kriegel. Some problems of the parameter estimation of strange atractors. In W. Ebeling and M. Peschel, editors, *Lotka-Volterra-Approach to Cooperation and Competition in Dynamic Systems*, pages 114–122, Berlin, 1985. Akademie-Verlag.

[124] S.M. Hammel, J.A. Yorke, and C. Grebogi. Numerical orbits of chaotic dynamical processes represent true orbits. *Bull. Amer. Math. Soc.*, 19:465–470, 1988.

[125] C. Grebogi, S.M. Hammel, J.A. Yorke, and T. Sauer. Shadowing of physical trajectories in chaotic dynamics: containment and refinement. *Phys. Rev. Letts.*, 65, no. 13:1527–1530, 1990.

[126] T. Sauer and J.A. Yorke. Rigorous verification of trajectories for the computer simulation of dynamical systems. *Nonlinearity*, 4:961–979, 1991.

[127] B.A. Coomes, H. Koçak, and K.J. Palmer. Rigorous computational shadowing of orbits of ordinary differential equations. *Numerische Mathematik*, 69:401–421, 1995.

[128] B. Russell. On the notion of cause. In H. Feigl and M. Brodbeck, editors, *Readings inthe Philosophy of Science*, New York, 1053. Appleton-Century-Crofts.

[129] W.A. Brock, D.A. Hsieh, and B. LeBaron. *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence.* MIT Press, Cambridge, MA, 1991.

[130] W.A. Brock, W.D. Dechert, and J. Scheinkman. A test for independence based on the correlation dimension. technical report 8702. Social Systems Research Institute, University of Wisconsin, Madison, 1987.

[131] W.A. Brock. Distinguishing random and deterministic systems: Abridged version. *J.Econ.Theo.*, 40:168–195, 1986.

[132] R. von Mises. *Probability Statistics and Truth.* George Allen and Unwin, London, 1957.

[133] J. Earman. *A primer on determinism*, volume 32 of *University of Western Ontario series in philosophy of science.* Reidel, Boston, 1986.

[134] D. T. Kaplan. Exceptional events as evidence for determinism. *Physica D*, 73(1–2):38–48, 1994.

[135] H. M. van den Dool. Searching for analogues, how long must we wait? *Tellus*, 46 A(3):314–324, 1994.

[136] J.D. Farmer and J. Sidorowich. Exploiting chaos to predict the future and reduce noise. In Y. C. Lee, editor, *Evolution, Learning, and Cognition*, page 277. World Scientific, 1988.

[137] M Gaster. The nonlinear phase of wave growth leading to chaos and the breakdown to turbulence in a boundary layer as an example of an open system. *Proc. R. Soc. Lond. A*, 430:3–24, 1990.

[138] P. D. Thompson. Uncertainty of initial state as a factor in the predictability of large-scale atmospheric flow patterns. *Tellus*, 9:275–295, 1957.

[139] ECMWF. *Predictability*, Seminar Proceedings, Shinfield Park, Reading RG2 9AX, UK, 4 - 8 September 1995.

[140] F. Molteni, R. Buizza, T.N. Palmer, and T. Petroliagis. The ECMWF ensemble prediction system: Methodology and validation. *Q.J.R.Meteorol.Soc.*, 122:73–119, 1996.

[141] T.N. Palmer. Medium and extended range predictability and stability of the pacific/north american mode. *Quarterly Journal of the Royal Meteorological Society*, 114:691–713, 1988.

[142] R. Buizza and T.N. Palmer. The singular-vector structure of the atmospheric general circulation. 1993. Submitted to J. Atmos. Sci., November 1993.

[143] Z. Toth and E. Kalnay. Ensemble forecaszing at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, 74(12):2317–2330, 1993.

[144] Z. Toth and K. Kalnay. Ensemble forecasting at NMC and the breeding method. NMC office note 407, NMC, 1995.

[145] Z. Toth and E. Kainay. Ensemble forecasting at ncep and the breeding method. *Mon. Wea. Rev*, 1997. in print.

[146] I. Szunyogh, E. Kainay, and Z. Toth. A comparison of lyapunov vectors and optimal vectors in a low resolution gcm. *Tellus*, 49A:200–227, 1997.

[147] Y. Zhu, G. Iyengar, Z. Toth, S. M. Tracton, and T. Marchok. Objective evaluation of the ncep global ensemble forecasting system. In *Preprints, 15^{th} AMS Conference on Weather Analysis and Forecasting*, 1996. Norfolk, Virginia.

[148] J. Anderson and V. Hubeny. A reexamination of methods for evaluating the predictability of the atmosphere. *Nonlinear Processes in Geophysics*, 1997. in review.

[149] M. R. Allen. Quantifying uncertainty in climate analysis and prediction. NERC Advanced Fellowship Proposal, Available from Dept. of Physics, AOPP, Oxford University, Oxford OX1 3PU, 1997.