School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

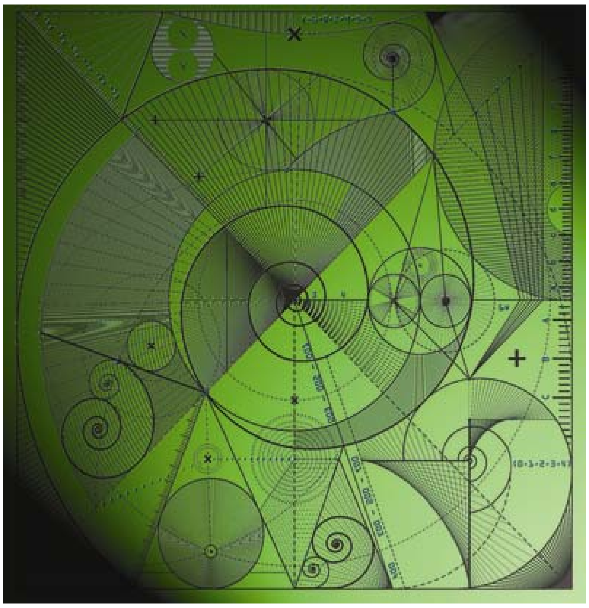
Preprint MPS_2010-28

10 September 2010

Initial Distribution Spread: A density forecasting approach

by

R.L. Machete and I.M. Moroz



Initial Distribution Spread: A density forecasting approach

R. L. Machete[a, 1]and I. M. Moroz[b]

a. Dept. of Mathematics and Statistics, P. O. Box 220, Reading, RG6 6AX, UK

b. Mathematical Institute, 24-29 St Giles', Oxford, OX1 3LB

Abstract

Ensemble forecasting of nonlinear systems involves the use of a model to run forward a discrete ensemble (or set) of initial states. Data assimilation techniques tend to focus on estimating the true state of the system, even though model error limits the value of such efforts. This paper argues for choosing the initial ensemble in order to optimise forecasting performance rather than estimate the true state of the system. Density forecasting and choosing the initial ensemble are treated as one problem. Forecasting performance can be quantified by some scoring rule. In the case of the logarithmic scoring rule, theoretical arguments and empirical results are presented. It turns out that, if the underlying noise spread of the time series exceeds the spread of one step forecast errors, we can diagnose the underlying noise spread regardless of the noise distribution.

Keywords: data assimilation; density forecast; ensemble forecasting; uncertainty

1 Introduction

Given an *initial state* of some chaotic dynamical system − examples of which include the population of an animal species in a game reserve, daily weather for Botswana or day to day electricity demands for London − we could perform a point forecast from the single state. Observational uncertainty and/or model error could limit the value of such a forecast. One can go over these hurdles by generating a discrete set of initial states in the neighbourhood of the current state and then forecasting from it. The set of initial states is called an *initial ensemble.* Forecasting from an initial ensemble is called *ensemble forecasting* [1]. The time ahead at which forecasts are made from any member of the initial ensemble is called *lead time.* Ensemble forecasting is performed mainly to account for uncertainty in the initial conditions, although it can also be used to mitigate model error. A lot of attention has been paid to generating initial ensembles (e.g see [1, 2]). Here, we present a novel approach to selecting the spread of the distribution from which an initial ensemble is drawn. The distribution from which an initial ensemble is drawn shall be called the *initial distribution.* Its covariance matrix will be taken to be diagonal and uniform over all the initial conditions considered, in tune with common practice in data assimilation and ensemble forecasting [1, 3, 4]. Note, however, that we do not assume that the initial distribution is the underlying noise distribution.

[1] Corresponding author: `r.l.machete@reading.ac.uk`, tel: +44(0)118 378 5016

Abbreviations: Initial Distribution Spread (IDS), Perfect Model Scenario (PMS), Imperfect Model Scenario (IMS), Moore-Spiegel (M-S), European Centre for Medium-Range Weather Forecasts (ECMWF)

Ideally, the initial ensemble should be drawn from the underlying invariant measure, in which case we have a *perfect initial ensemble.* A perfect initial ensemble is especially useful in the scenario when our forecasting model is isomorphic to the model that generated the data, which scenario is called the *perfect model scenario* (PMS) [2, 5]. When there is no isomorphism between the forecasting model and the model that generated the data, then we are in the *imperfect model scenario* (IMS). A perfect model with a perfect initial ensemble would give us a *perfect forecast* [6]. If either our model or initial ensemble is not perfect, then we have no reason to expect perfect forecasts.

In all realistic situations, we have neither a perfect model nor a perfect initial ensemble, yet we may be required to issue a meaningful forecast probability density function (pdf). Roulston and Smith [7] proposed a methodology for making forecast distributions that are consistent with historical observations from ensembles. This is necessary because the forecast ensembles are not drawn from the underlying invariant measure due to either imperfect initial ensembles or model error. Their methodology was extended by Broecker and Smith [8] to employ continuous density estimation techniques [9, 10] and blend the ensemble pdfs with the empirical distribution of historical data, which is referred to as *climatology.* The resulting pdf is what will be taken as the forecast pdf in this paper.

The quality of the forecast pdfs can be assessed using the logarithmic scoring rule proposed by Good [11] and termed *ignorance* by Roulston and Smith [12], borrowed from information theory [13, 14]. Here, we discuss a way of choosing the initial distribution spread (IDS) to enhance the quality of the forecast pdfs. The point is that if the spread is too small our forecasts may be over confident and if it is too large our forecasts may have low information content. Our goal is to choose an IDS that yields the most informative forecast pdfs and determine, for instance, if this varies with the lead time of interest. As is commonly done in data assimilation and ensemble forecasting (e. g. see [1, 15]), we only consider Gaussian initial distributions. In traditional data assimilation and ensemble forecasting techniques, estimation of the initial distribution is divorced from forecasting: this is the main point of departure in our approach. We revisit this later in the discussion of the results in § 5.

Our numerical forecasting experiments were performed on the Moore-Spiegel (M-S) [16] system and an electronic circuit motivated by the M-S system. Indeed electronic circuits have been studied to enhance our understanding of chaotic systems and Chua circuits [17] are among famous examples. Recently, Gorlov and Strogonov [18] applied ARIMA models to forecast the time to failure of Integrated Circuits. Hence, electronic circuits have not only been studied to enhance our understanding of chaotic systems and the forecasting of real systems, but also to understand the circuits themselves and to address practical design questions.

This paper is organised as follows: § 2 introduces the technical framework for discussing probabilistic forecasting of deterministic systems. The theoretical and empirical scores for probabilistic forecasts are presented in §3. Computations of the initial ensemble spread are discussed in § 4 for the PMS and IMS. In the PMS, the M-S system [16] is considered. For the IMS, the M-S system and an electronic circuit are modelled using radial basis function (rbf) models. The circuit was constructed in a physics laboratory using state of the art equipment to mimic the M-S system. A theoretical argument in support of the numerical results is also presented. Implications and practical relevance of the results are discussed in § 5 and concluding remarks are given in § 6.

2 Forecasting

Consider a deterministic dynamical system,

$$\dot{\boldsymbol{x}} = \boldsymbol{F}(\boldsymbol{x}(t), \boldsymbol{\lambda}), \tag{1}$$

with the initial condition $\boldsymbol{x}(0) = \boldsymbol{x}_0$, where $\boldsymbol{x}, \boldsymbol{F} \in \mathbb{R}^m$, $\boldsymbol{\lambda} \in \mathbb{R}^d$ is a vector of parameters, $\boldsymbol{F}$ is a Lipschitz continuous (in $\boldsymbol{x}$), nonlinear vector field and t is time. By Picard's theorem [19], (1) will have a unique solution, say $\boldsymbol{\varphi}_t(\boldsymbol{x}_0; \boldsymbol{\lambda})$. If $\nabla.\boldsymbol{F} < 0$, this system might have an attractor [20], which if it exists we denote by A. In particular, we are interested in the case when the flow on this attractor is chaotic.

2.1 Forecast Density

For any point in state space, $\boldsymbol{x}$, and positive real number ϵ, let $B_{\boldsymbol{x}}(\epsilon)$ denote an ϵ-ball centred at $\boldsymbol{x}$. Suppose that ϱ is some invariant measure (see appendix A) associated with the attractor A. For any $\boldsymbol{x}_0 \in A$, we define a new probability measure associated with $B_{\boldsymbol{x}_0}(\epsilon)$ by

$$\varrho_0(E) = \lim_{T\to\infty} \frac{1}{T\varrho(B_{\boldsymbol{x}_0}(\epsilon))} \int_0^T \mathbf{1}_{E\cap B_{\boldsymbol{x}_0}(\epsilon)}(\boldsymbol{x}(t))\mathrm{d}t, \tag{2}$$

where $\mathbf{1}$ is an indicator function. This measure induces some probability density function, $p_0(\boldsymbol{x}, \boldsymbol{x}_0, \epsilon)$. We will call a set of points drawn from $p_0(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$ a *perfect initial ensemble*. At any time t, the forecast of the perfect initial ensemble using the flow $\boldsymbol{\varphi}_t$ will be distributed according to some pdf $p_t(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$. The pdf $p_t(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$ will be referred to as a *perfect forecast density* at lead time t.

2.2 Imperfect Forecasts

Operationally, we never get perfect forecasts since our initial ensemble is never drawn from $p_0(\boldsymbol{x}, \boldsymbol{x}_0, \epsilon)$ and our model, $\boldsymbol{\varphi}_t(\boldsymbol{x})$, is always some approximation of the system, $\bar{\boldsymbol{\varphi}}_t(\boldsymbol{x})$, which possibly lives in a different state space. In that case, our forecast pdf would be $f_t(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$ rather than the perfect forecast $p_t(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$. Henceforth we suppress the ϵ dependence.

3 Scoring Probabilistic Forecasts

The next question would be: how close is $f_t(\boldsymbol{x}; \boldsymbol{x}_0)$ to $p_t(\boldsymbol{x}; \boldsymbol{x}_0)$? In a general sense, we consider the score of a forecast $f_t(\boldsymbol{x}; \boldsymbol{x}_\tau)$ and denote it by $S(f_t(\boldsymbol{x}; \boldsymbol{x}_\tau), \boldsymbol{X})$ [21], where $\boldsymbol{X}$ is the random variable of which $\boldsymbol{x}$ is a particular realisation. If $\boldsymbol{X}$ is distributed according to $p_t(\boldsymbol{x}; \boldsymbol{x}_\tau)$, the expected score of f_t is

$$\mathbb{E}[S(f_t(\boldsymbol{x}; \boldsymbol{x}_\tau), \boldsymbol{X})] = \int S(f_t(\boldsymbol{x}; \boldsymbol{x}_\tau), \boldsymbol{z}) p_t(\boldsymbol{z}; \boldsymbol{x}_\tau)\mathrm{d}\boldsymbol{z}. \tag{3}$$

At lead time, t, the overall forecast score on the attractor is

$$\mathbb{E}[S(t)] = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[S(f_t(\boldsymbol{x}; \boldsymbol{x}_\tau), \boldsymbol{X}^{(\tau)})]\mathrm{d}\tau, \tag{4}$$

where $\boldsymbol{X}^{(\tau)}$ is the random variable being forecast from the initial distribution corresponding to $\boldsymbol{x}_\tau$. Provided the underlying attractor is ergodic, we can rewrite (4) as

$$\mathbb{E}[S(t)] = \lim_{T\to\infty} \frac{1}{T}\int_0^T S(f_t(\boldsymbol{x};\boldsymbol{x}_\tau), \boldsymbol{X}^{(\tau)})\mathrm{d}\tau. \tag{5}$$

For each forecast, the underlying system can only furnish one verification of $\boldsymbol{X}$ and not the distribution $p_t(\boldsymbol{x};\boldsymbol{x}_0)$. Therefore, we use (5) to score forecasts rather than (4). Discretise time according to $\tau_i = (i-1)\tau_s$, for $i = 1, 2, .., N$, where τ_s is the sampling time. This gives a sequence of forecast pdfs, $\{f_t(\boldsymbol{x};\boldsymbol{x}_i)\}_{i=1}^N$, corresponding to verifications $\{\boldsymbol{X}^{(i)}\}_{i=1}^N$ and score S. We can thus discretise (5) to obtain the following empirical score to value the t-ahead forecast system:

$$\langle S\rangle(t) = \frac{1}{N}\sum_{i=1}^{N} S(f_t(\boldsymbol{x};\boldsymbol{x}_i), \boldsymbol{X}^{(i)}). \tag{6}$$

This is the same score proposed by Broecker and Smith [21].

In this paper, we shall use the score:

$$S(f_t, \boldsymbol{X}) = \mathrm{ign}(f_t, \boldsymbol{X}), \tag{7}$$

where $\mathrm{ign}(f_t, \boldsymbol{X}) = -\log f_t(\boldsymbol{X})$ is "the information deficit or *Ignorance* that a forecaster in possession of the pdf has before making the observation $\boldsymbol{X}$" [12]. An important property of this score is that it is *strictly proper.* A strictly proper score is one for which (3) assumes its minimum if and only if $f_t = p_t$ [22]. Another property of the Ignorance score, although less persuasive, is *locality.* A score is local if it only requires the value of the forecast pdf at the verification to be evaluated [21].

4 Initial Distribution Spread

The primary concern is to determine optimal initial distribution spreads for the forecasting problem. Each initial ensemble is drawn from a Gaussian distribution centred at the initial observation. The problem is then reduced to finding the optimal spread of the Gaussian distribution. An optimal spread is one that minimises the average score at the lead time of interest. In the theoretical setup this score would be the one given by equation (4) and in an operational setup we would use the empirical score given by (6). In the numerical examples considered in this section, we use continuous forecast pdfs obtained from discrete forecasts as discussed in [8].

The cases considered are the PMS and the IMS. In the PMS, numerical experiments are performed on the M-S system [16] at classical parameter values. In the IMS, the M-S system and circuit are considered and the models are constructed from data using cubic rbf's (see [23, 24] for details). We shall denote the spread of the underlying Gaussian distribution of the initial ensemble by σ_e and that of observational noise by δ. For a given observational noise level, we vary σ_e logarithmically between 10^{-3} and 1. $\delta = 0$ will represent the noise-free case. In the multivariate case, we set σ_e to be the spread of the perturbation of the ith coordinate and then set the standard deviation of the perturbation of the jth coordinate to be

$$\sigma_e^{(j)} = \sigma_e \frac{\sigma_j}{\sigma_i}, \tag{8}$$

where σ_j is the standard deviation of the jth variable.

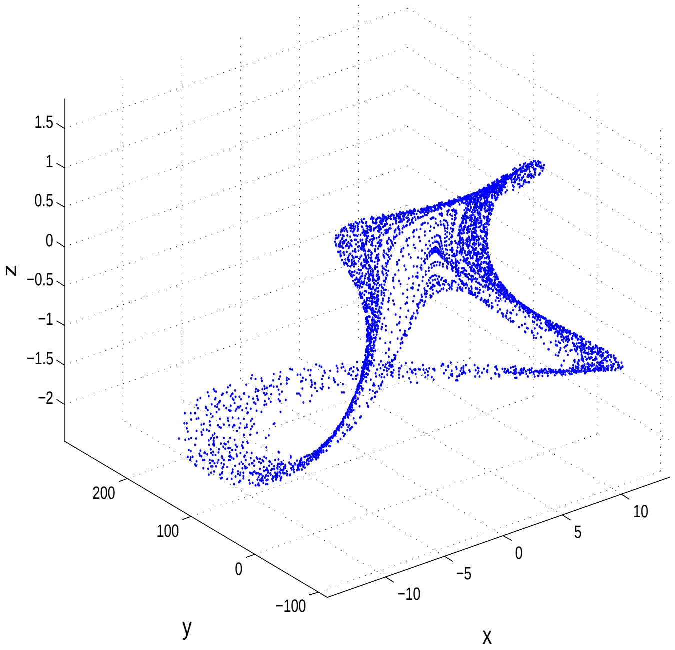


Figure 1: The graph of the M-S attractor at parameter values $T = 36$ and $R = 100$.

4.1 Perfect Model Scenario

We consider the M-S system [16], which is given by:

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -y + Rx - T(x+z) - Rxz^2, \\ \dot{z} &= x, \end{aligned} \tag{9}$$

with classical parameters $T \in [0, 50]$ and $R = 100$. This system was integrated for $T = 36$ and $R = 100$ using a 4-th order Runge-Kutta method to generate some data, which we will call *M-S data.* Transients were discarded to ensure that all the data collected were on the attractor, which is shown in figure 1. From any initial point on the M-S data, an initial ensemble is generated by perturbing the observation with some random variable drawn from a Gaussian distribution and the M-S system in (9) used as the model to forecast this ensemble.

4.1.1 Clean Data

For the case $\delta = 0$, the graphs of average Ignorance, $\langle \text{ign}(\sigma_e) \rangle$, versus the IDS, σ_e, are shown in figure 2. The different colours correspond to the different lead times of up to 32 time steps. Notice that the graphs generally yield straight lines except at higher lead times and IDS. In particular, the magenta lines (corresponding to lead times of 32) saturate at higher values of σ_e. As the IDS increases, we would expect the forecast pdfs at low lead times to be approximately flattened Gaussians. That is why all the red lines (lead times 1 and 2) grow linearly without saturating. Notice that the lead times of, say 31 and 32, score less than the 20 lead times when $\sigma_e > 10^{-1}$. When higher lead time forecasts score less than lower lead times, we say we have *return of skill.* Linear graphs

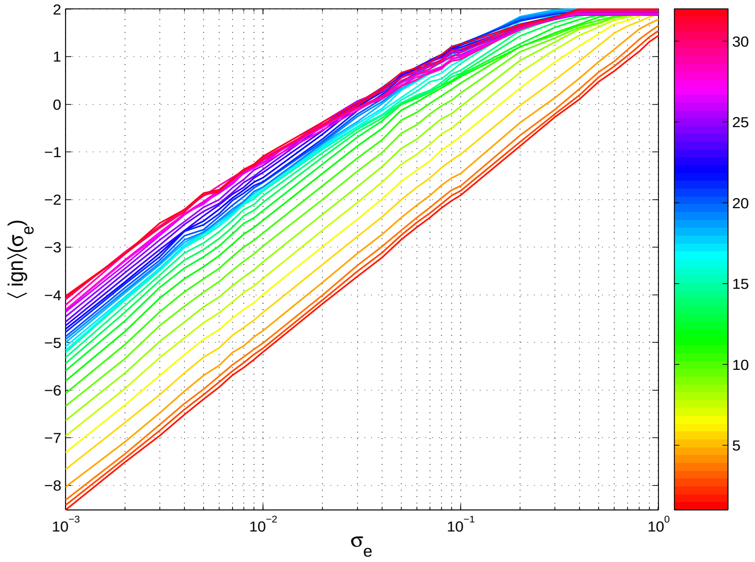


Figure 2: The graphs of average Ignorance versus initial distribution spread, σ_e, using a perfect M-S model with 512 ensembles for various lead times (according to the right colour bar), each ensemble containing 32 members. The M-S data was noise-free.

such as those in figure 2 suggest that the underlying model is perfect and the data is noise-free.

4.1.2 Noisy Data

We next consider the case when the data has observational noise of standard deviation, $\delta = 5 \times 10^{-2}$. The corresponding graphs of average Ignorance versus the IDS, σ_e, are shown in figure 3. At low IDS, all the graphs are almost flat since the initial distribution spreads are drowned by the noise. As the IDS increases, the higher lead time graphs begin to dip. This is because the forecast pdfs at higher lead times spread out, and in the process, the verifications which were initially at the tails of the distributions, tend to be encapsulated by the ensembles as we gain skill (see figure 4). At low lead times, the verifications are generally at the centre of the ensembles. As the distributions spread out and flatten, average Ignorance increases. That is why at low lead times, the graphs, initially flat then begin to increase linearly. This occurs when $\sigma_e \approx \delta$: the same value of σ_e at which graphs of the higher lead times attain their minima.

To address whether the above results could have been influenced by using too few ensemble members, we considered the graphs of the kernel width, σ_k, versus the IDS, σ_e. For the average Ignorance graphs shown in figure 3, the corresponding graphs of σ_k versus σ_e are shown on the right hand side. Since these were obtained with 32 ensemble members per forecast, we increased the number of ensemble members per forecast to 256 and then plotted the associated graphs shown in figure 5. In both cases, the graphs of σ_k versus σ_e are essentially the same, confirming that the results obtained with 32 ensemble members were not biased by the number of ensemble members.

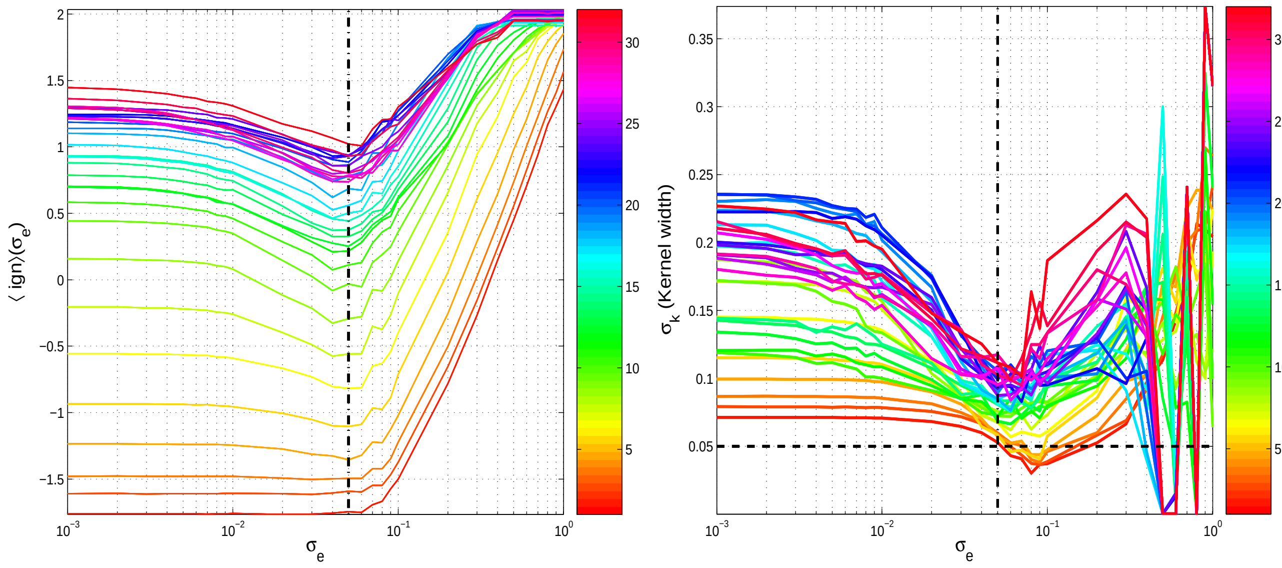


Figure 3: (left) The graph of ignorance versus initial distribution spread, σ_e, for the perfect M-S model with 512 ensembles for various lead times (according to the right colour bar), each ensemble containing 32 members. Observational noise of standard deviation 5×10^{-2}. (right) Graph of kernel width versus initial distribution spread. The vertical and horizontal thick dash-dotted lines correspond to the noise spread.

4.2 Imperfect Model Scenario

We now carry over the ideas of the preceding subsections to the IMS, considering models of the form

$$x_n = \phi(\boldsymbol{x}_{n-1}), \tag{10}$$

where $\boldsymbol{x}_{n-1}$ is a delay vector. The deterministic model, ϕ, was built from data using cubic radial basis functions.

4.2.1 The M-S system

Let us first consider M-S data with observational noise, $\delta = 10^{-2}$ and 10^{-1}. The graphs of average Ignorance versus IDS, σ_e, for various lead times are shown in figure 6. We notice that the low lead time graphs begin to rise at $\sigma_e \approx \delta$. At a slightly larger value of σ_e, graphs for the higher lead times reach their minima. This is very much reminiscent to the PMS, and suggests a way of using nonlinear prediction to detect noise level.

Suppose the noise of the underlying system is not Gaussian but is uniformly distributed with standard deviation δ. We consider this case and the noise distribution to be $U[-b, b]$, in which case $\delta^2 = b^2/3$. We have plotted graphs of average Ignorance versus IDS in figure 7 with $\delta = 10^{-1}$. Again we see graphs dipping at $\sigma_e \approx \delta$.

When the data is noise free, we have the results shown in figure 8. They are qualitatively similar to the PMS and IMS with noisy data. This case also presents striking differences from the previous noisy data scenarios. The optimal spread varies with lead time as highlighted by the black solid line on the figure. In this case, one may select the IDS that yields good forecasting performance at higher lead times.

The foregoing discussions can be summarised as follows: Whereas there is similarity in the graphs of average Ignorance for the PMS with observational noise and the IMS,

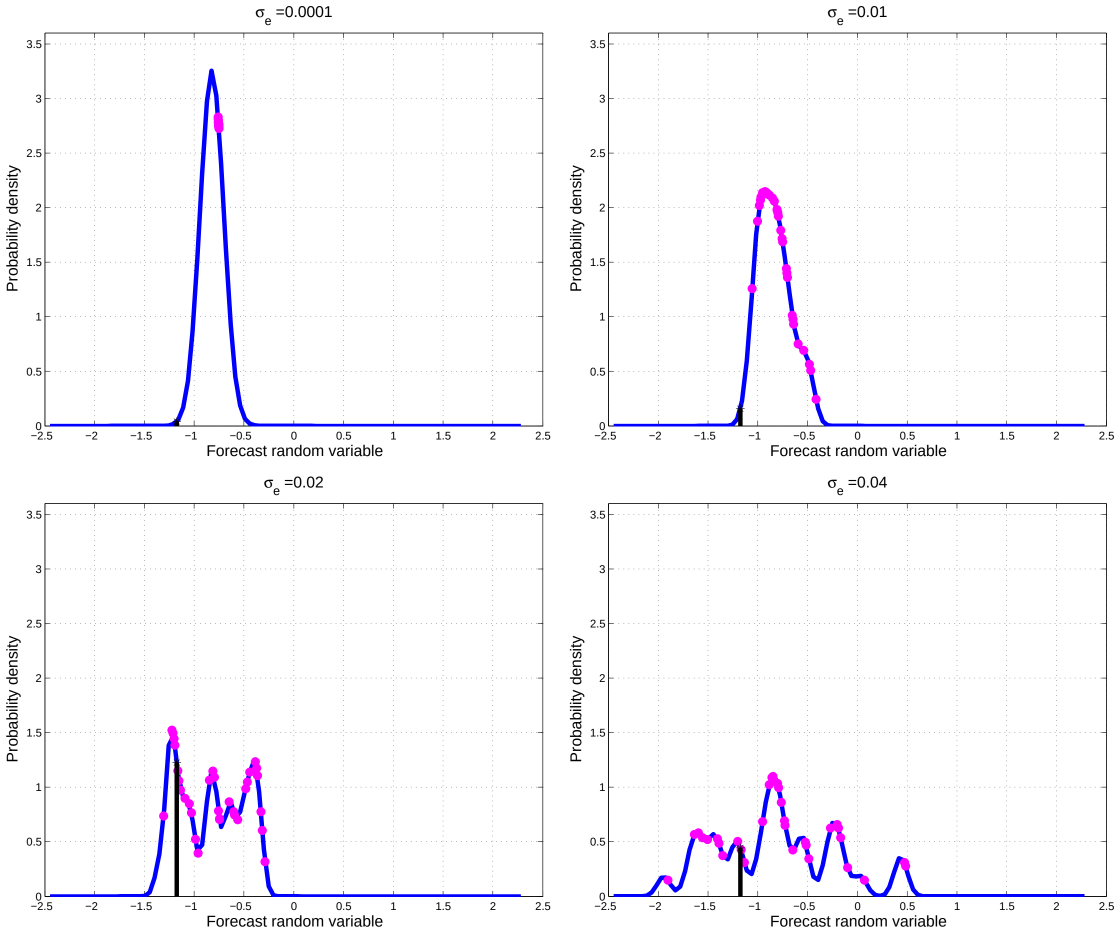


Figure 4: A series of ensemble pdfs at a lead time of 16 time-steps from the initial condition. The magenta dots are the actual ensemble members. Notice that as the initial distribution spread, σ_e, increases, the value of $f_t(x^*)$ (indicated by the black vertical line) increases and then decreases, where x^* is the verification.

there is a clear difference with the PMS on clean data shown in figure 2. In the two former cases, the average Ignorance curves do not show a linear rise. This furnishes us with a simple, heuristic test of whether or not we are in the PMS with clean data. If the noise level dominates model error, we may be in a position to detect that level. Otherwise, in general, we cannot be sure if the problem is due to model error or observational noise.

4.2.2 The Circuit

Before concluding this section, we consider the circuit. The main question we wish to answer for the circuit is: what IDS should we use for a given model? This question is addressed using average Ignorance as we have explained in the preceding paragraphs. A graph of average Ignorance versus IDS for the circuit is shown in figure 9. Notice that the graph of the first lead time begins to rise at $\sigma_e \approx 10^{-3}$, which is quite small. This suggests that the underlying noise level is very low. This is comparable to the standard deviation of the one-step errors of the model (not shown). The graphs look much like those obtained with M-S data without observational noise, but with an imperfect

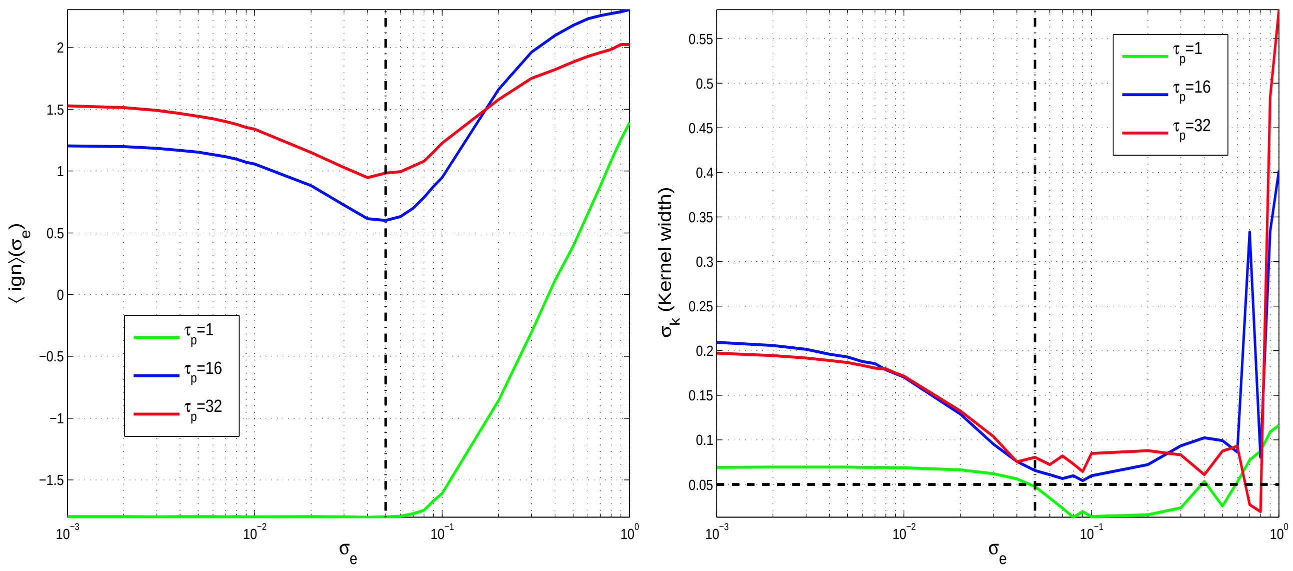


Figure 5: (left) The graph of ignorance versus initial distribution spread, σ_e, for the perfect M-S model with 512 ensembles for three lead times, each ensemble containing 256 members. Observational noise had standard deviation 5×10^{-2}. (right) Graph of kernel width versus initial distribution spread. The vertical and horizontal thick dash-dotted lines correspond to the noise spread.

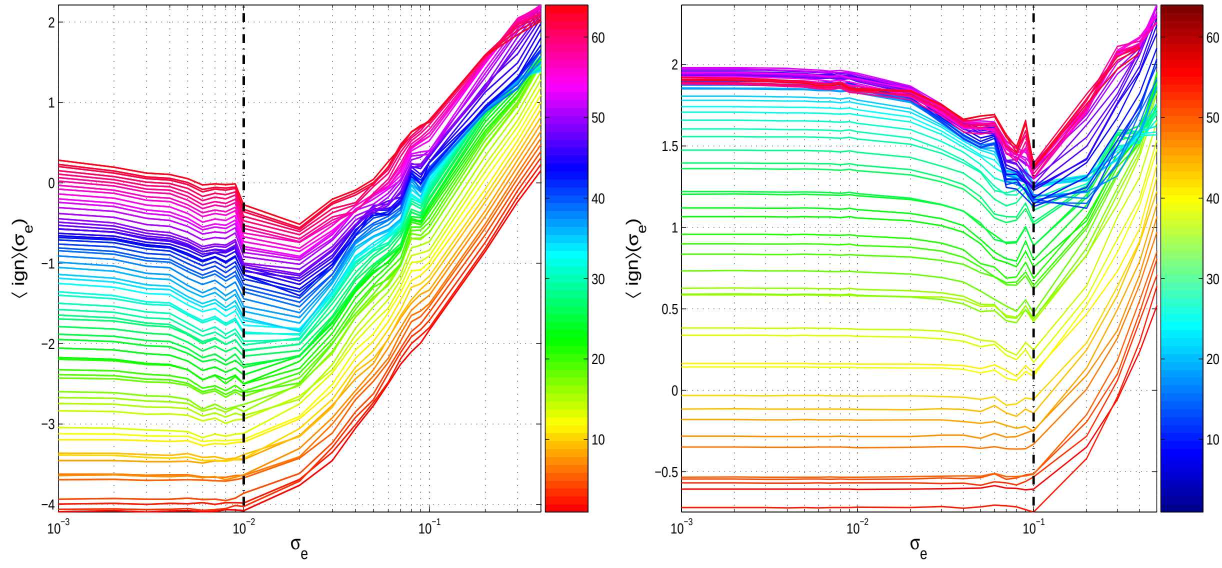


Figure 6: Graphs of average Ignorance versus logarithmically initial distribution spread, σ_e, with observational error of standard deviation $\delta = 10^{-2}$ (left) and 10^{-1} (right) on M-S data with an imperfect model. 128 initial conditions with a time step of 64 between them were used. Each initial ensemble containing 32 members was iterated forward up to 64 time steps. The multiple lines correspond to different lead times. The lowest lines correspond to the lowest lead times but there is a mixing up of higher lead times at the top of each graph. The vertical thick dash-dotted lines correspond to the noise spread.

model (see figure 8), albeit without dipping to such an extent.

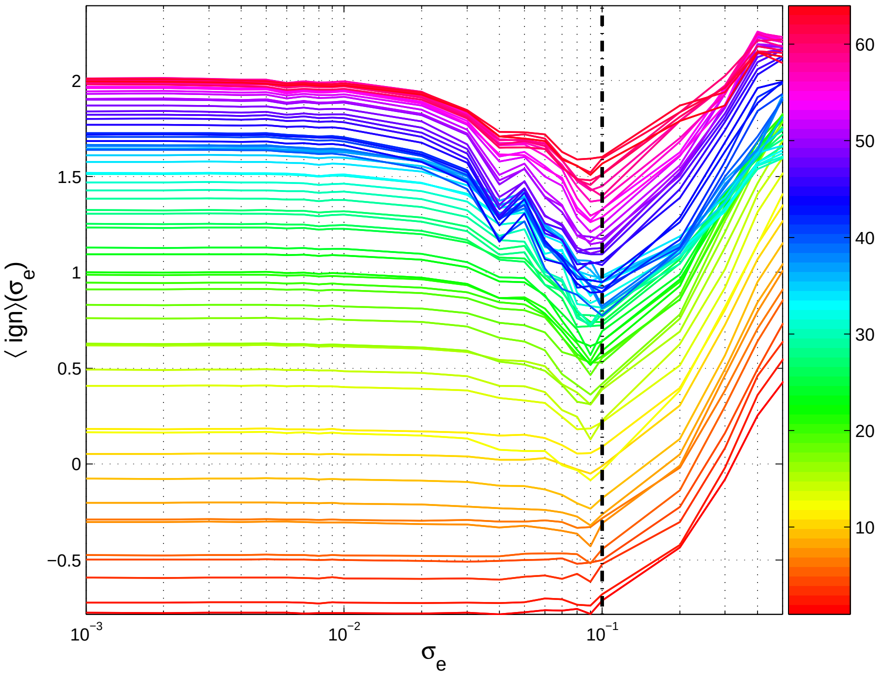


Figure 7: Graphs of Ignorance versus logarithmically varying initial distribution spread, σ_e, of ensemble perturbations with uniformly distributed observational error of standard deviation $\delta = 10^{-1}$ on M-S data with an imperfect model. 128 initial conditions with a time step 64 between them were used. 32 initial ensembles were generated in each and iterated forward up to 64 time steps. The multiple lines correspond to different lead times according to the colour bar on the right. The vertical thick dash-dotted line corresponds to the noise spread.

4.3 Theoretical Considerations

To explain the previous observations, we consider two pdfs of the perfect forecast and the imperfect forecasts: $p_t(x; \sigma_p, \mu_p)$ and $f_t(x; \sigma_f, \mu_f)$, where σ_p (or σ_f) and μ_p (or μ_f) are the standard deviation and mean respectively of p_t (or f_t), assuming that

$$\sigma_p(t) = h_p(\sigma_e, t) \quad \text{and} \quad \sigma_f(t) = h_f(\sigma_e, t).$$

Suppose our forecast, f_t, is Gaussian [2], so that

$$f_t(x; \sigma_f, \mu_f) = \frac{1}{\sigma_f\sqrt{2\pi}} e^{-(x-\mu_f)^2/2\sigma_f^2}.$$

Then the expected skill of f_t is

$$\begin{aligned}\mathbb{E}[\text{ign}(f_t, X)] &= -\int_{-\infty}^{\infty} p_t(x; \sigma_p^2, \mu_p) \log f_t(x; \sigma_f^2, \mu_f)\mathrm{d}x \\ &= \frac{1}{2}\log(2\pi\sigma_f^2) + \frac{\sigma_p^2}{2\sigma_f^2} + \frac{1}{2\sigma_f^2}(\mu_p - \mu_f)^2. \qquad (11)\end{aligned}$$

We assume that the standard deviations, σ_p and σ_f, are monotonic increasing functions of σ_e. If $\sigma_p = \sigma_f$ then (11) reduces to

$$\mathbb{E}[\text{ign}(f_t, X)] = \frac{1}{2}\log(2\pi e\sigma_f^2) + \frac{1}{2\sigma_f^2}(\mu_p - \mu_f)^2 \qquad (12)$$

[2] Operationally, this may not be the case.

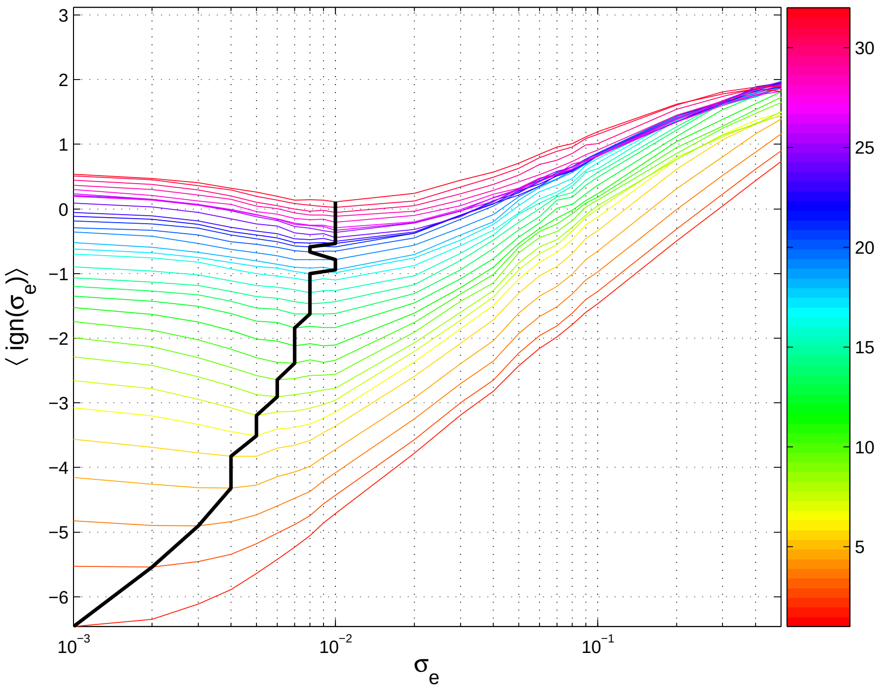


Figure 8: The graphs of average Ignorance versus initial distribution spread with 512 ensembles, each ensemble containing 32 members and using a cubic rbf model on noise free M-S data. The colour bar on the right shows the lead times for the different graphs of Ignorance. The solid black line indicates where the global minimum occurs for each graph. Notice how the optimum spread varies with lead time.

and the expected skill is minimised by

$$\sigma_f = |\mu_p - \mu_f|. \tag{13}$$

If, in addition, $\mu_p = \mu_f$, then

$$\mathbb{E}[\mathrm{ign}(f_t, X)] = \frac{1}{2}\log(2\pi e \sigma_f^2), \tag{14}$$

which is a monotonically increasing function of σ_f. This may explain why straight line graphs were obtained in the noise free PMS. They arise when the perfect and the imperfect forecast ensembles have equal means and variances.

If $\mu_p \neq \mu_f$, then the expected skill has a global minimum given by

$$\min_{\sigma_f > 0} \mathbb{E}[\mathrm{ign}(f_t, X)] = \frac{1}{2}\log\left[2\pi e^2(\mu_p - \mu_f)^2\right]. \tag{15}$$

In particular,

$$\min_{\sigma_e > 0} \mathbb{E}[\mathrm{ign}(f_0, X)] = \frac{1}{2}\log\left[2\pi e^2 \xi_0^2\right], \tag{16}$$

where $\xi_0 = \mu_p - \mu_f \sim N(0, \delta^2)$. Here, μ_p is the mean of the initial distribution from which the initial condition, x_0, was drawn and μ_f is the mean of the forecast pdf. The initial condition lies on the attractor. For a more general case, at lead time t, we define $\xi_t = \mu_p(t) - \mu_f(t)$. If $\sigma_e > \xi_0$, then the minimum in (16) will not be attained by

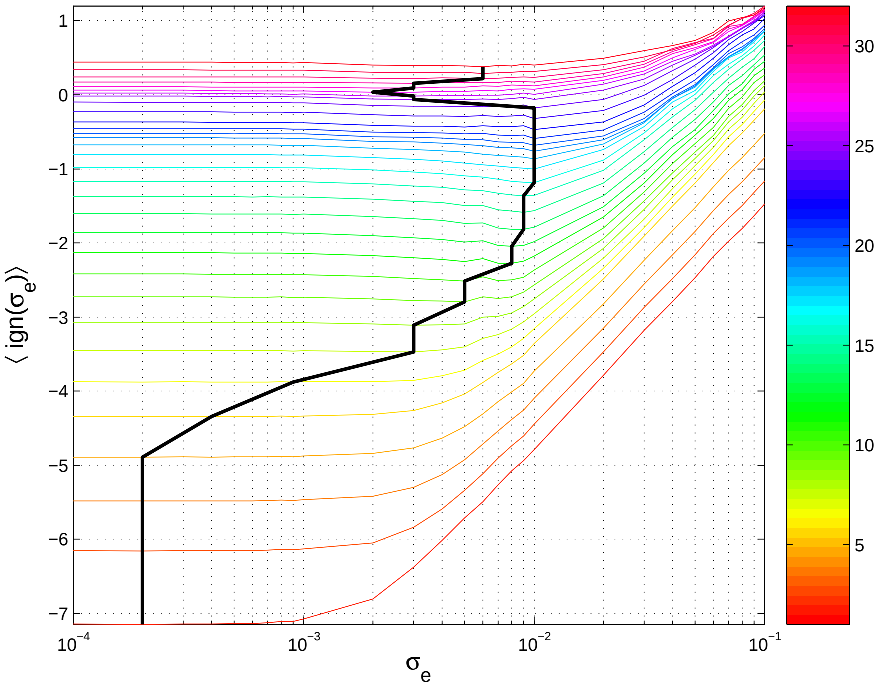


Figure 9: Graphs of average Ignorance versus logarithmically varying initial distribution spread, σ_e, on circuit data. 512 initial conditions with a time step of 64 between them were used. Each initial ensemble containing 32 members was iterated forward up to 32 time steps. The multiple lines correspond to different lead times according to the colour bar on the right.

increasing σ_e because it can only be attained when $\sigma_f = \sigma_e = |\xi_0|$. However, over a window of time series, the average may be constant for a while as witnessed in figure 3. We assume that for t close to zero, the distribution of ξ_t is approximately that of ξ_0. For higher lead times, the minima of the average skill are attained at $\sigma_e = \delta$.

5 Discussion

The computational results presented in this paper demonstrated a way to select the spread of the distribution from which to sample an initial ensemble of points. The goal was to obtain an initial ensemble that would minimise uncertainty in the forecast distributions. The forecasting model need not be perfect for the method to be applied. Information theoretic approaches were used to obtain the computational results and justify them. The methodology is a departure from traditional data assimilation and ensemble forecasting techniques in a number of ways. We recognise that the ultimate goal of any method that estimates an initial distribution is to obtain more accurate forecasts.

Data assimilation techniques either focus on estimating the true state of the system or a set of such estimates. To this end, a model trajectory may be sought that is consistent with observations [2, 25]. It is believed that forecasts made from an ensemble that lies along such trajectories would provide good forecasts. An ensemble of trajectories is obtained by making perturbations of some initial observation. When there is model error, there is no model trajectory that is consistent with observations. Therefore, Judd and Smith [5] talk of pseudo-orbits instead. Notwithstanding these difficulties, the method presented here could be used to determine the spread of this distribution, regardless of the data assimilation technique. For a given structure of the correlation matrix, we would seek the scalar multiple that yields the most informative forecast distributions.

Other techniques for producing the initial ensemble aim at selectively sampling those points that are dynamically the most relevant. In particular, the ECMWF ensemble prediction system seeks perturbations of the initial state based on the leading singular vectors of the linear propagator [1]. This approach can lead to over-confidence when there is model error. One falls into the trap of confusing the dynamics of the model with those of the underlying system as highlighted in [26]. To safeguard this problem, our methodology may be used to select the IDS.

The results also suggest that the method may be useful in nonlinear noise reduction. For nonlinear noise reduction, the quality of the model would have to be very good, at least in the sense of forecasting. However, the primary value of the method is to find the spread of the initial distribution. It is also interesting that even when there is no observational noise, sampling the initial distribution could still help mitigate model inadequacy.

Finally, possible areas of application go beyond meteorology and the Geo-sciences. For instance, evidences of nonlinear dynamics have already been reported in economics and finance [27]. In some cases these dynamics are fairly low dimensional (e.g. [28]), thus reducing the computational costs that may arise from generating an initial ensemble. We envision the method being of great value in these disciplines to tackle density forecasting.

6 Conclusions

This paper argued for combining the task of choosing the initial ensemble with density forecasting. The point is that, when faced with model error, a knowledge of the true state of the system is irrelevant because it cannot provide one with a perfect forecast. Moreover, using the true state with an imperfect model can provide forecasts that are further from the truth than forecasts obtained with imperfect initial states. Therefore, it has been argued that the task of the forecaster should be to choose initial distributions that yield the most informative forecast distributions. Whereas this approach may be incorporated into traditional ensemble forecasting techniques, it can also stand independently as a forecasting method.

To recap, it was demonstrated that the logarithmic scoring rule can be used to estimate an optimum IDS for a given system and model. At the optimal spread, higher lead time graphs of the logarithmic scoring rule versus IDS tend to dip. Although it is critical that we use Gaussian initial distributions, the distribution of the underlying observational uncertainty or model error seems not to play a crucial role. It turns out that we can also diagnose the fictitious case of a perfect model with perfect initial states. A theoretical explanation for the empirical observations regarding the dipping of the graphs has been presented. Also, by appealing to ergodicity, the theoretical score has also been related to the empirical score. It is noteworthy that Bernardo's theorem [29] states that the logarithmic score is the only proper, local scoring rule. This point should dispel concern about the generality of the results.

Acknowledgements

The authors would like to thank Leonard A. Smith, Devin Kilminster and members of the Applied Dynamical Systems Group at Oxford for fruitful discussions and useful insights. This work was supported by the Commonwealth scholarship, the LSE's Higher Education Infrastructure fund and the RCUK Digital Economy Programme fund at the University of Reading.

References

[1] M. Leutbecher, T. N. Palmer, Ensemble Forecasting, Journal of Computational Physics 227 (2008) 3515–3539.

[2] K. Judd, L. A. Smith, Indistinguishable states I: Perfect model scenario, Physica D 151 (2001) 125–141.

[3] L. M. Stewart, Correlated observation errors, Ph.D. thesis, University of Reading (2010).

[4] A. Hollingsworth, P. Lonnberg, The statistical structure of short-range forecast errors as determined from radiosonde data. part 1: The wind field, Tellus 38A (1986) 111–136.

[5] K. Judd, L. A. Smith, Indistinguishable states II: The imperfect model scenario, Physica D 196 (2004) 224–242.

[6] L. A. Smith, C. Ziehman, K. Fraedrich, Uncertainty in dynamics and predictability in chaotic systems, Q. J. R. Meteorol. Soc. 125 (1999) 2855–2886.

[7] M. S. Roulston, L. A. Smith, Combining dynamical and statistical ensembles, Tellus 55A (2003) 16–30.

[8] J. Broecker, L. A. Smith, From Ensemble Forecasting to Predictive Distribution Functions, Tellus A 60 (2008) 663.

[9] E. Parzen, On the Estimation of a Probability Density Function and Mode, The Annals of Mathematical Statistics 33 (1962) 1065–1076.

[10] B. W. Silverman, Density Estimation for Statistics and Data Analysis, 1st Edition, Chapman and Hall, 1986.

[11] I. J. Good, Rational decisions, Journal of the Royal Statistical Society. Series B (Methodological) 14 (1952) 107–114.

[12] M. S. Roulston, L. A. Smith, Evaluating Probabilistic Forecasts Using Information Theory, Monthly Weather Review 130 (2002) 1653–1660.

[13] A. I. Khinchin, Mathematical Foundations of Information Theory, 1st Edition, Dover Publications, Inc., 1957.

[14] C. E. Shannon, A Mathematical theory of communication, Bell Systems Technology Journal 27 (1948) 379–423,623–656.

[15] J. Broecker, U. Parlitz, M. Ogorzalek, Nonlinear noise reduction, Proc. of the IEEE 90.

[16] W. D. Moore, E. A. Spiegel, A thermally excited nonlinear oscillator, The Astrophysical Journal 143 (1966) 871–887.

[17] L. O. Chua, Experimental chaos synchronisation in Chua's circuit, International Journal of Bifurcation and Chaos 2 (1992) 704.

[18] M. I. Gorlov, A. V. Strogonov, ARIMA Models Used to Predict the Time to Degradation Failure of TTL IC's, Russian Microelectronics 36 (2007) 261–270.

[19] E. A. Coddington, N. Levinson, Theory of Ordinary Differential Equations, New York:McGraw-Hill, 1955.

[20] E. Ott, T. Sauer, J. A. Yorke (Eds.), Coping With Chaos: Analysis of chaotic data and the exploitation of chaotic systems, John Wiley and Sons Inc., 1994.

[21] J. Broecker, L. A. Smith, Scoring Probabilistic Forecasts: The importance of being proper, Weather and Forecasting 22 (2007) 382–388.

[22] T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction and estimation, J. Amer. Math. Soc. 102 (2007) 359–378.

[23] K. Judd, A. Mees, On selecting models for nonlinear time series, Physica D. 82 (1995) 426–444.

[24] R. L. Machete, Modelling a Moore-Spiegel Electronic Circuit: the imperfect model scenario, DPhil Thesis, University of Oxford (2008).

[25] G. Burgers, P. J. van Leeuwen, G. Evensen, Analysis scheme in the ensemble kalman filter, Monthly Weather Review 126 (1998) 1719–1724.

[26] L. A. Smith, Chaos: A Very Short Introduction, 1st Edition, Oxford University Press, 2007.

[27] T. Terasvirta, Forecasting Economic Variables with Nonlinear Models, in: G. Elliott, C. W. J. Granger, A. Timmermann (Eds.), Handbook of Economic Forecasting, Vol. 1, North-Holland, 2006.

[28] N. Linden, S. Satchell, Y. Yoon, Predicting British Financial Indices: An approach based on chaos theory, Structural Change and Economic Dynamics 4.

[29] J. M. Bernardo, Expected information as expected utility, Annals of Statistics 7 (1979) 686–690.

[30] J.-P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, Rev. Mod. Phys. 57 (1985) 617–653.

A Invariant Density

Associated with the attractor A is some *invariant measure* [30], ϱ, such that

$$\varrho[\boldsymbol{\varphi}_{-t}(E)] = \varrho(E), \tag{17}$$

where $E \subset \mathbb{R}^m$ is a measurable set and $\boldsymbol{\varphi}_{-t}(E)$ is the set obtained by evolving each point in E backwards in time. A probability measure on E may be defined as [30]

$$\varrho(E) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{1}_E(\boldsymbol{\varphi}_t(\boldsymbol{x}_0))\mathrm{d}t, \tag{18}$$

where $\mathbf{1}_E$ is an indicator function [3]. Provided the attractor A is *ergodic* [4],

$$\varrho(E) = \int_E \varrho(\mathrm{d}\boldsymbol{x}). \tag{19}$$

Associated with ϱ is some probability density function, ρ, such that (19) may be rewritten as

$$\varrho(E) = \int_E \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{20}$$

We call $\rho(\boldsymbol{x})$ the *invariant density* of the attractor A or the flow $\boldsymbol{\varphi}_t(\boldsymbol{x}_0)$. This invariant density is indeed the climatology [5] mentioned in the introduction.

[3] An indicator is defined by

$$\mathbf{1}_E(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{x} \in E, \\ 0 & \text{if} \quad \boldsymbol{x} \notin E. \end{cases}$$

[4] In an ergodic attractor, state space averages are equal to time averages [30].

[5] Including its marginal densities.