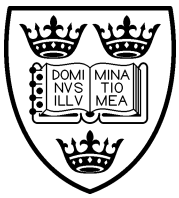
Modelling a Moore-Spiegel Electronic Circuit: the imperfect model scenario



Reason L. Machete

Worcester College

University of Oxford

Thesis submitted for the degree of

Doctor of Philosophy in Applied Mathematics

Trinity, 2007

This theses is dedicated to my mother, who worked hard to ensure that I obtained an education, the only heritage she could afford to give me, and esteemed above material riches, amidst extremely limited resources, but was denied by the chilling hand of death, at a tender age of 46, from seeing her dreams realised.

Acknowledgements

Let me first thank my supervisors: Dr. I. M. Moroz for guiding me through this initially daunting endeavour, giving various suggestions on the way and always available to read my draft documents and give useful feed-back; Dr L. A. Smith, for his invaluable support and ability to engender interest. He also introduced me to the enviable CATS group at LSE. The various train discussions we have had shall remain memorable; Dr. D. Kilminster, whose contribution to this research has been superb and I am highly indebted to his generosity. Thank you all for the patience to bear with inspecting the many "pebbles and shells" that I often brought to your attention.

May I also thank those who assessed me during transfer and confirmation of status, among whom were Dr A. Fowler and Dr. D. A. Allwright. Their comments have had a positive influence on the direction of the research. I am particularly grateful to Dr. D. A. Allwright, whose office was always open for even the simplest of my questions.

I also appreciate comments from Dr. F. Kwasniok and Dr. G. Nicholls. Among other comments, they pointed out that what I thought was interesting phenomenon could have simply been due to rational aliasing.

The ADS group in OCIAM, the CATS group in LSE have all given me valuable support and fruitful discussions. I also thank all those who read draft portions of the thesis, among whom are Dr. L. Clarke at the CATS group and Mr. O. Mogapi in the Statistics department. My heartfelt appreciations also go to the Lancaster University Physics Group for initially providing their laboratory and equipment to build the circuit that is central to this thesis in 2004. This research would not have been carried out without financial support, most of which came from the Commonwealth Scholarship. Other funds, not less appreciated, have been provided by the University of Botswana, University of Oxford, London School of Economics and Political Science and Worcester College (Oxford).

Let me also thank the various friends and family members. Although I will not name any, some of them have performed such kindly deeds that shall never cease to occupy my mind.

I also wish to express appreciation to the support of my wife, Oabona, whose courage and sacrifice saw us through often enduring strait circumstances. She has often had to bear with taking the most menial jobs, which were not always available, to help us take care of our monthly expenses.

Now unto the King eternal, immortal, invisible, the only wise God, be

honour and glory for ever and ever. Amen. (I Timothy 1:17)

Abstract

The goal of this thesis is to investigate model imperfection in the context of forecasting. We focus on an electronic circuit built in a laboratory and then enclosed to reduce environmental effects. The non-dimensionalised model equations, obtained by applying Kirchhoff's current and voltage laws, are the Moore-Spiegel Equations [47], but they exhibit a large disparity with the circuit. At parameter values used in the circuit, they yield a periodic trajectory whilst the circuit exhibits chaotic behaviour. Therefore, alternative models for the circuit are sought.

The models we consider are local and global prediction models constructed from data. We acknowledge that all our models have errors and then seek to quantify how these errors are distributed across the circuit attractor. To this end, *q-pling* times of initial uncertainties are computed for the various models. A q-pling time is the time for an initial uncertainty to increase by a factor of q [67], where q is a real number. Whereas it is expected that different models should have different q-pling time distributions, it is found that the diversity in our models can be increased by constructing them in different coordinate spaces.

To forecast the future dynamics of the circuit using any of the models,

we make probabilistic forecasts [8]. The question of how to choose the spread of the initial ensemble is addressed by the use of *skill scores* [8, 9]. Finally, the diversity in our models is exploited by combining probabilistic forecasts from them so as to minimise some skill score. It is found that the skill of combined not-so-good models can be increased by combining them as discussed in this thesis.

Contents

1 Introduction **1**

1.1 Dynamical Systems 3

1.2 Predictability 5

1.3 Uncertainty Directions 9

1.4 Recurrence 11

1.5 Organisation 12

2 The Moore-Spiegel Circuit **15**

2.1 Design Issues 16

2.2 The Circuit 22

2.3 Experiment and Data Acquisition 25

2.4 Circuit Data Exploration 30

2.5 Conclusions 35

3 Dynamical Models **38**

3.1 Prediction Embeddings 40

3.1.1 Embedding Theorems 40

3.1.2 Time Delay 42

3.1.3 Box Counting Dimension 43

3.2 The Prediction Problem 46
3.2.1 Local Models 46
3.2.2 Kwasniok-Smith Algorithm 50
3.2.3 Global Models 55
3.3 Conclusions 58

4 Q-pling Time Distributions **60**
4.1 The Perfect Model Scenario 62
4.2 The Imperfect Model Scenario 66
4.2.1 Local Dynamics 67
4.2.2 Delay Space Models 70
4.2.3 State space model 73
4.3 Similarity Measure 75
4.4 Takens and Predictability 77
4.4.1 Q-pling times 79
4.4.2 Q-pling regions 83
4.4.3 Conclusions 83
4.5 One-step-error q-pling times 85
4.6 Lyapunov q-pling times 88
4.7 Conclusions 96

5 Ensemble Prediction **99**
5.1 Perturbed MS system 101
5.2 Skill Scores 107
5.3 Ignorance 109

5.4 Dressing 112
5.5 Climatology 119
5.6 Initial-Perturbation Spread 125
5.6.1 The Perfect Model Scenario (PMS) 126
5.6.2 The Imperfect Model Scenario (IMS) 130
5.7 Conclusions 135

6 Multiple Models **137**
6.1 Multiple Model Ensembles 138
6.2 Combining Probabilistic Forecasts 140
6.3 Application to the Circuit 142
6.4 Conclusions 147

7 Conclusions and Further Work **149**

A Lorenz system **151**

B Circuit Modules **152**

C Least Squares Solution **154**

List of Figures

2.1 Graphs showing how resistivity varies with temperature in both a semiconductor and a conductor. 17

2.2 (a) Pictorial view of a physical chip and (b) circuit symbol of an OpAmp. 18

2.3 OpAmp negative feedback network 20

2.4 A functional block diagram of the multiplier 21

2.5 Moore Spiegel circuit diagram. 23

2.6 A periodic orbit of the MS-system at parameter values $\gamma = 3.2$ and $\Gamma = 10$. 25

2.7 Bifurcation transition sequence of the modified MS system obtained by plotting the extrema of z against γ with $\Gamma = 10$ fixed and the coefficient of y in the first equation of (2.10) being 10. 26

2.8 (a) The square wave that is input to an integrator from a signal generator and (b) the sawtooth wave expected at the output of the integrator if components are functional at optimum conditions. 27

2.9 Voltage divider network used to monitor circuit ambient temperature changes. 30

2.10 Smoothed (using the same filter) temperature proxies for the long circuit data sets. T_7 (for Set7), T_8 (for Set8) and T_9 (for Set9) are the temperature proxies, with $T_i = V_T^{(i)}$. 32

2.11 Projections onto the (V_3, V_1)-plane of the dynamics visited by circuit (From Set7). (a) Typical orbit mostly manifested in the first 3 minutes and (b) Typical attractor that the circuit finally settles in. The points plotted were uniformly spaced in time. 33

2.12 Projections of the (a) circuit (Set7) and (b) MS attractor onto the (V_1, V_2) and (x, y)-space respectively. The MS-system was integrated at parameter values $\gamma = 36, \Gamma = 100$. Each plot contains 4000 data points. 34

2.13 First return map of successive maxima of (a) V_3 (for Set7 with the periodic orbit removed) and (b) z. In each case, $V_3^{(n)}$ and $z^{(n)}$ are each the nth successive local maximum. Notice the overall similarities between the two pictures. There are also marked differences between the two maps, with the left arm of return map for MS system folding upward. The circuit return map appears to have some extra portions. 35

2.14 Circuit time series for Set7 in four consecutive time epochs with $R_1 = 10k\Omega$ and $R_7 = 3.85k\Omega$. 36

3.1 A graph of the mutual information, $I(\tau)$ versus the time lag, τ, for the voltage signal V_3 from Set7. The first local minimum occurs at $\tau = 6$ time steps, which is the best time delay to use in the delay vectors. 44

3.2 (a) Graphs of the correlation integral versus the box size. The different lines correspond to different dimensions used in the order $m = 2 - 6$ downwards. (b) Graph of gradient estimates with colours corresponding to dimensions on (a). None of the lines shows any convergence, but the m=2 line (blue) exhibits the minimum variation, assuming a minimum of 1.28 and maximum of 2.2. We used 5×10^5 data points from SET9 of the circuit. 45

3.3 The graph of RMS error (of forecasts made using local linear models) versus the number of nearest neighbours for Set7. The different lines correspond to different samples of points from the learning set for which we make in-sample predictions and compute the RMS error. Notice that the general pattern of the lines is that they start off decreasing, reach a local minimum, and then rise up. The optimum number of nearest of neighbours is the mean or median of the global minima. We used $m = 3$, $\tau = 4$ time steps, 10^4 points in the learning set from the circuit, sampling (8 times) 200 points to perform cross validation, and found the mean to be 15.25 and median 16. 49

3.4 The (a) initial and (b) final learning sets of the circuit projected onto the 3D delay space of the processing 1.2×10^5 with KS algorithm. Notice that the refined learning set (in (b)) has a different distribution of points from the initial learning set. 53

3.5 Each graph in (a), (b) and (e) shows the running mean over 1024 points when implementing the Kwasniok-Smith algorithm. (a) Shows the running mean of absolute out-of-sample errors. (b) Running root mean square error. (c) Absolute out-of-sample error. (d) Red dots are points in the time series where the out-of-sample error exceeds 0.006 and the blue area is the time series processed by KS algorithm. (e) An estimate of the exchange probability at each stage of KS algorithm. 54

4.1 A view of the Lorenz attractor showing the doubling times of initial perturbations in the (a) Lyapunov direction, (b) maximal vector direction and (c) random direction. Red indicates $\tau_2 < 0.12$, yellow $0.15 \leq \tau_2 < 0.25$, green $0.25 \leq \tau_2 < 0.52$, cyan $0.52 \leq \tau_2 < 0.72$, blue $0.72 \leq \tau_2 < 1.22$, magenta $1.22 \leq \tau_2 < 2.02$, and black $\tau_2 \geq 2.02$. After Smith et al. [67] 64

4.2 (left) Histograms of doubling times for the Lorenz attractor in the uncertainty directions of the diagrams shown in figure 4.1 after Smith [67] and (right) the corresponding cumulative distributions. On the vertical axis of left picture is the relative frequency of the bins used (or the proportion of sample points within a given bin). 65

4.3 A view showing where the uncertainty doubling occurs in the three uncertainty directions described in figure 4.1. The starting points (green) reflect the natural measure on the attractor while the points where doubling occurs (red) do not. 66

4.4 A view of the Moore-Spiegel attractor at $\gamma = 36$ showing the doubling times of initial perturbations in the (a) Lyapunov directions, (b) maximal direction and (c) random direction. Red indicates $\tau_2 < 0.05$, yellow $0.05 \leq \tau_2 < 0.1$, green $0.1 \leq \tau_2 < 0.5$, cyan $0.5 \leq \tau_2 < 0.8$, blue $0.8 \leq \tau_2 < 1$, magenta $1 \leq \tau_2 < 2$, and black $\tau_2 \geq 2$. 67

4.5 A view of the Moore-Spiegel attractor in embedding space at $\gamma = 36$ showing the doubling times for 10000 initial conditions for (a) the Lyapunov direction, (b) maximal direction and (c) random direction. Red indicates $\tau_2 < 0.05$, yellow $0.05 \leq \tau_2 < 0.1$, green $0.1 \leq \tau_2 < 0.5$, cyan $0.5 \leq \tau_2 < 0.8$, blue $0.8 \leq \tau_2 < 1$, magenta $1 \leq \tau_2 < 2$, and black $\tau_2 \geq 2$. The embedding dimension was $m = 5$ and the length of trajectories over which the linear propagator was computed with $p = 4$ iterations. 69

4.6 MS attractor doubling time distributions based on (a) local dynamics approximations and (b) the perfect model with time partitions as in figure 4.5. 70

4.7 A view of the Moore-Spiegel circuit in delay space showing distributions of doubling times for a delay space model with initial errors in (a) Lyapunov, (b) maximal, and (c) random directions using data Set7. Red indicates $\tau_2 < 0.4$, yellow indicates $0.4 < \tau_2 < 0.9$, green indicates $0.9 < \tau_2 < 1.3$, cyan indicates $1.3 < \tau_2 < 3$, blue indicates $3 < \tau_2 < 8$, magenta indicates $8 < \tau_2 < 20$ and black indicates $\tau_2 > 20$. 73

4.8 A view of the Moore-Spiegel circuit in delay space showing the doubling times of initial perturbations in the (a) Lyapunov direction (b), singular vector direction (c) and random direction with the Jacobian approximated from the local dynamics according to section 4.2.1 using data Set7. Red indicates $\tau_2 < 0.15$, yellow $0.15 \leq \tau_2 < 0.2$, green $0.2 \leq \tau_2 < 0.4$, cyan $0.4 \leq \tau_2 < 0.6$, blue $0.6 \leq \tau_2 < 0.9$, magenta $0.9 \leq \tau_2 < 2$, and black $\tau_2 \geq 2$. The length of trajectories over which the linear propagator was computed was $p = 4$. 74

4.9 Distributions of doubling times for models (a) M_1, (b) M_2, (c) M_4 and (d) M_5 of the circuit. Models M_1 and M_5 were built in 3D delay space, model M_2 in 4D delay space and model M_4 in measurement space. From measurement space, the distribution of doubling times for model M_4 were then mapped into delay space. Red indicates $F(\tau_2) < 0.2$, yellow indicates $0.2 < F(\tau_2) < 0.4$, green indicates $0.4 < F(\tau_2) < 0.6$, cyan indicates $0.6 < F(\tau_2) < 0.8$ and blue indicates $F(\tau_2) > 0.8$. $F(\cdot)$ is the cumulative distribution function of the doubling times. In each case we considered the doubling of the prediction variable with the initial vector in the Lyapunov direction. 80

4.10 View of the MS circuit showing distributions of octapling times for models (a) M_1, (b) M_2, (c) M_4 and (d) M_5 of the circuit. Models M_1 and M_5 were built in 3D delay space, model M_2 in 4D delay space and model M_4 in measurement space. From measurement space, the distribution of octapling times for model M_4 were then mapped into delay space. Red indicates $F(\tau_2) < 0.2$, yellow indicates $0.2 < F(\tau_2) < 0.4$, green indicates $0.4 < F(\tau_2) < 0.6$, cyan indicates $0.6 < F(\tau_2) < 0.8$ and blue indicates $F(\tau_2) > 0.8$. $F(\cdot)$ is the cumulative distribution function of the doubling times. In each case we considered the octa-pling of the prediction variable with the initial error on the axis of the prediction variable. 81

4.11 Views of the MS circuit showing distributions of doubling times for model 4 of the circuit on data sets 7 and 9. Red indicates $F(\tau_2) < 0.2$, yellow indicates $0.2 < F(\tau_2) < 0.4$, green indicates $0.4 < F(\tau_2) < 0.6$, cyan indicates $0.6 < F(\tau_2) < 0.8$ and blue indicates $F(\tau_2) > 0.8$. $F(\cdot)$ is the cumulative distribution function of the doubling times. We used $||\boldsymbol{\epsilon_0}||=10^{-6}$. The similarity between these two pictures suggests that the circuit dynamics are not altered by the ambient temperature fluctuations. 82

4.12 Regions where uncertainty octapling occurs for diagrams shown in figure 4.10. Green reflects the underlying circuit attractor and red indicates regions where octapling occurs. 84

4.13 The region where uncertainty octapling occurs for the diagram shown in figure 4.8(a). Green dots indicate the initial conditions and red dots indicate where doubling occurs. Notice that the regions where octapling time occurs reflect the underlying natural measure of the circuit. 84

4.14 A view of the MS attractor showing OSEQ times for model M_{1e} with (a) $q = 2$, (b) $q = 64$ and (c) $q = 128$. The distributions of the OSEQ times are as follows: red reflects $F(\tau_q) < 0.2$, yellow $0.2 < F(\tau_q) < 0.4$, green $0.4 < F(\tau_q) < 0.6$, cyan $0.6 < F(\tau_q) < 0.8$ and blue $F(\tau_q) > 0.8$, where $F(\cdot)$ is the cumulative distribution function of τ_q. 86

4.15 Views of the circuit attractor showing variations of OSE-quadrupling times of models (a) M_1, (b) M_2, (c) M_4 and (d) M_5. 87

4.16 Cumulative distributions of OSE quadrupling times, of the circuit models, corresponding to figure 4.15. 89

4.17 Regions where OSE-quadrupling occurs (red) and on the circuit attractor (green) for the models with quadrupling times shown in figure 4.15. 90

4.18 Histogram of Lyapunov doubling times of the Lorenz attractor. Notice the oscillatory nature, a signature of self-similarity. The successive extrema have been denoted ν_i, $i = 1, \ldots, 9$. 91

4.19 (a) View of the Lorenz attractor showing distributions of doubling times in the Lyapunov with the colour scheme dictated by the extrema of figure 4.18: Red reflects $\tau_2 < 0.31 = \nu_2$, yellow indicates $\nu_2 = 0.31 < \tau_2 < 0.71 = \nu_3$, green reflects $\nu_3 = 0.71 < \tau_2 < 1.04 = \nu_4$, cyan reflects $\nu_4 = 1.04 < \tau_2 < 1.39 = \nu_5$, blue reflects $\nu_5 = 1.39 < \tau_2 < 1.58 = \nu_6$, and black reflects $\tau_2 > 1.58 = \nu_6$ and (b) points on the attractor whose doubling times correspond to extrema on the PDF of figure 4.18. 92

4.20 Histogram of Lyapunov doubling times of the MS attractor. Notice the oscillatory nature, a signature of self-similarity. The successive extrema have been denoted μ_i, $i = 1, \ldots, 11$. 93

4.21 Points on the MS attractor whose doubling times correspond to successive extrema on the PDF of figure 4.20. The arrows indicate the direction of the flow. 94

4.22 Histogram of Lyapunov doubling times of the MS circuit using model M_4 on data sets 7 and 9. It is oscillatory just like those obtained in the perfect model scenario. We used $||\epsilon_0|| = 10^{-6}$. 95

4.23 OSE quadrupling times for the models M_1 (green), M_2 (blue) and M_4 (red) of the circuit. 96

5.1 Diagramatical representation of equation (5.5). 103

5.2 Projections of the perturbed Moore Spiegel attractor with $\mu = 0, 10^{-3}$, $\lambda = 0.64$ and $\varepsilon = 0.1$ onto the (x, y)-plane. Notice that the two attractors cannot be distinguished by eye, yet modelling the $\mu = 10^{-3}$ system with the $\mu = 0$ system fails in some regions of state space (see figure 5.3). 104

5.3 Ensemble predictions of the perturbed Moore-Spiegel system ($\mu = 10^{-3}$) using original Moore-Spiegel system as the model. Each ensemble had 32 members obtained by perturbing the initial conditions with Gaussian perturbations of standard deviation $\epsilon_x = 10^{-3}$ for the x-variable and ϵ_y and ϵ_z for the other variables according to equation (5.7). In each figure, the cyan trajectories are the ensemble members, the black trajectory (control) corresponds to unperturbed initial conditions and the red one is the truth. Notice that, in (a)-(c), the control trajectory stays close to the truth for a while before going astray. In (a) and (b), the ensembles do pick the verification when the control goes astray and in (c) both the ensembles and the control go astray after a while. However, most regions like the one in (d) are well tracked by the control. 105

5.4 Integrated mean square error as a function of n (number of observations) using A) histogram estimates and B) kernel estimates. The circles (or upward and downward triangles) correspond to three random number selections for each n. For downward triangles, Gaussian kernels were used and rectangular kernels were used for upward triangles. The kernels were used with optimum widths. From Glanovic [15]. 116

5.5 PDFs obtained by dressing the ensemble forecasts (At lead lead time of 6.4ms) of the circuit without (left) and with (right) the climatology. The magenta circles are the forecasts and the black asterisk is the verification. Clearly the right distributions are sharper than the left ones and this fact is reflected by the respective entropies, yet the variances of the left ones are lower. 122

5.6 Shows the climatology of the circuit made from 8192 data points and then used to compute the PDFs of figures 5.5-5.7. 123

5.7 PDFs obtained by dressing the ensemble forecasts (At lead time of 6.4ms) of the circuit without (left) and with (right) climatology. The magenta circles are the forecasts and the black asterisk is the verification. In this case the diagnosis of both variance and entropy are in agreement. 123

5.8 Graphs of average Ignorance in the training (top left) and testing (top right) period. The lines graphs correspond to dressing with climatology (blue), without climatology (red) and black dashed line is Ignorance of the climatology. On the bottom is the graph of the blending parameter versus lead time. 124

5.9 The graphs of average Ignorance versus ensemble perturbation with 512 ensembles for various lead times (according to the right colour bar), each ensemble containing 32 members. The MS data was noise free and we used a perfect M-S model. The perturbations were Gaussian with standard deviation ε. 126

5.10 The graph of ignorance versus ensemble perturbation with 512 ensembles for various lead times (according to the right colour bar), each ensemble containing 32 members. The MS data was corrupted with observational noise of standard deviation 5×10^{-2} and we used perfect M-S model. . . . 127

5.11 A series of ensembles for at lead time 16 of the circuit obtained with an imperfect model. Notice that as ϵ increases, the value of $\rho(x)$ increases and then decreases, where x is indicated by the asterisk (verification). The corresponding graph of ignorance versus ϵ is shown in figure 5.15. 128

5.12 Graphs of average Ignorance versus logarithmically varying standard deviation, ϵ, of ensemble perturbations with initial observational error of standard deviation $\delta = 10^{-2}$ in (a) and 10^{-1} in (b) on MS 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. 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. 131

5.13 Graphs of ignorance versus logarithmically varying standard deviation, ϵ, of ensemble perturbations with uniformly distributed observational error of standard deviation $\delta = 10^{-1}$ on MS 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 colorbar on the right. 132

5.14 The graphs of ignorance versus ensemble perturbation with 256 ensembles, each ensemble containing 32 members and using a cubic global model with dynamical noise of standard deviation 3×10^{-3} (left) and a without dynamical noise (right) on noise free M-S data. The colour bar on the right shows the lead times for the different graphs of Ignorance. 133

5.15 Graph of ignorance versus logarithmically standard deviation, ϵ, of Gaussian perturbations on circuit data. 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 colorbar on the right. 134

6.1 The graphs show ensemble predictions of a voltage signal, V_3 (magenta), by two models, M_1(green) and M_2 (blue), from two initial conditions x_0^1 (left) and x_0^2 (right).The vertical line at $t = 0$ indicates the time from which predictions are made. Notice that in the left pictures, M_1 stays close to the verification for a longer time than M_2 and vice versa in the right pictures. 139

6.2 Graphs of average Ignorance against lead time on 1024 training data points for dressed ensemble predictions by four models of the circuit M_1 (blue line), M_2 (green), M_3 (cyan), M_4 (red) and the combined model forecasts (of M_1, M_3 and M_4), M_c (black). In each case the solid lines correspond to ensembles blended with climatology and the corresponding dashed lines to not blending with it. The black dashed line is the climatology. 143

6.3 Graphs of average Ignorance over 1024 testing data against lead time on testing data for dressed ensemble predictions by four models of the circuit M_1 (blue line), M_2 (green), M_3 (cyan), M_4 (red) and the combined model forecasts (of M_1, M_3 and M_4), M_c (black) in the testing data. The black dashed line is the climatology. 144

6.4 Graphs of the weights of the constituent models of the combined model, M_c, against lead time: M_1 (blue line), M_3 (cyan), M_4 (red) and the climatology (magenta). 145

6.5 Graphs of the out-of-sample mean absolute errors of the deterministic forecasts of models M_1 (blue), M_2 (green), M_3 (red), M_4 (cyan) and M_c (black) obtained using equations (6.11) for the individual models and (6.12) for the combined model. 146

B.1 Schematic diagram of an adder module. 152

B.2 Schematic diagram of an integrator module. 153

B.3 Schematic diagram of a multiplier module. 153

Chapter 1

Introduction

Although the last three decades have witnessed a plethora of contributions to the study of nonlinear dynamics and chaos [1], operational challenges still remain to be confronted when one has to apply mathematical results to real physical systems; mainly because we do not have exact mathematical representations of these systems. Often, even the data measured from these systems may suffer from various limitations such as noise, few observation variables, sparsity or a relatively short observation period. Dool [3] highlighted the importance of studying physical analogue experiments whose observational period is long relative to their *recurrence time*[2]. In this thesis, we focus our attention on an electronic circuit for a number of reasons. Firstly, the circuit we consider provides long data sets collected over an approximately 14 hours duration. The data exhibit low noise and are sufficiently dense [3], which affords us high quality models that provide good forecasts over a few cycles.

Since the circuit equations obtained by a straight forward application of Kirchhoff's current and voltage laws are those originally proposed by Moore-Spiegel [47] as a

[1] A few examples are [12, 22, 63, 67].

[2] To be defined in § 1.4.

[3] Dense in the sense of Nyquist-Shannon [61] sampling theorem, which says the sampling frequency has to be greater than the Nyquist rate of the signal.

model for the height of an ionised gas in the atmosphere of a star, we call it the Moore-Spiegel (MS) circuit. The MS equations are

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -y + \Gamma x - \gamma(x + z) - \Gamma x z^2, \\ \dot{z} &= x, \end{aligned} \tag{1.1}$$

with classical parameters are $\gamma \in [0, 50]$ and $\Gamma = 100$. However, there is a huge disparity (see § 2.4) between the behaviour of the circuit and the numerical solution of the MS equations. This should not be thought to be due to the circuit having been built poorly. It may be largely due to bifurcations within the circuit. Therefore, we depart from using the MS system as a model for the circuit and consider other models constructed from the circuit data. The data we measure are voltages from three points on the circuit sampled at a high frequency.

We need to admit that all models have error [66] and *the purpose of this thesis is to explore the roles played by model error in the forecasting of the circuit.* Simply put, model error is the fundamental part addressed by this thesis. In effect, we explore the diversity in our models and then find a way of combining these models to utilise model diversity. To this end, we also consider a mathematical system (integration of a coupled set of MS equations) which we know perfectly. By pretending that we do not know much about the mathematical system, we can then try to understand it in the same way we do for the circuit.

In our modelling approach, we shall consider the circuit to be a *chaotic dynamical system*, which aspect is discussed in § 1.1. Since we acknowledge that our models have error, which inherently places a limit on how long our forecasts (or predictions) will remain relevant, the concept of *predictability* shall be discussed in this context in

§ 1.2. In § 1.5, the general organisation of the theses is outlined. Key issues of the thesis are given in the conclusions section.

1.1 Dynamical Systems

We envision a deterministic dynamical system to be one that can be modelled by a set of coupled ordinary deferential equations of the form

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t); \boldsymbol{\lambda}), \tag{1.2}$$

with the initial conditions $\boldsymbol{x}(0) = \boldsymbol{x}_0$, where $\boldsymbol{x}, \boldsymbol{F} \in \mathbb{R}^m$, $\boldsymbol{\lambda}$ is a vector of parameters, $\boldsymbol{F}$ is smooth and t denotes time. By Picard's theorem, (1.2) will have a solution (or flow) $\boldsymbol{\varphi}_t(\boldsymbol{x_0}; \boldsymbol{\lambda})$ for all $t > 0$. Whenever the parametric dependence is not relevant to the discussion, it will be suppressed and we will write the flow as $\boldsymbol{\varphi}_t(\boldsymbol{x_0})$. If the underlying dynamical system is *dissipative*[4] (i.e. $\nabla \cdot \boldsymbol{F} < 0$ for all $\boldsymbol{x}$) then phase space volumes are contracted [48]. The set of all possible values $\boldsymbol{x}(t)$ is called the *state space* and the path traced out by $\boldsymbol{x}(t)$ from a given $\boldsymbol{x}_0$ is called the *state space path* or *state space trajectory.*

A subset $A \subset \mathbb{R}^m$ is an *invariant set* for the flow $\boldsymbol{\varphi}_t$ if

$$\boldsymbol{\varphi}_t(\boldsymbol{x}) \in A \quad \text{for} \quad \boldsymbol{x} \in A \quad \text{for all} \quad t \in \mathbb{R}. \tag{1.3}$$

A closed invariant set $A \subset \mathbb{R}^m$ is called an *attracting set* if there is some neighbourhood U of A such that $\boldsymbol{\varphi}_t(\boldsymbol{x}) \in U$ for $t \geq 0$ and $\boldsymbol{\varphi}_t(\boldsymbol{x}) \to A$ as $t \to \infty$, for all $\boldsymbol{x} \in U$ [22]. The set $\bigcup_{t \leq 0} \boldsymbol{\varphi}_t(U)$ is called the *basin of attraction.* This framework allows us to define an *attractor* as an *attracting set which contains a dense orbit* [22]. An orbit of

[4]Does not necessarily correspond to any physical dissipation.

$\boldsymbol{x}_0 \in A$,

$$\mathcal{O}(\boldsymbol{x}_0) = \bigcup_{t \geq 0} \boldsymbol{\varphi}_t(\boldsymbol{x}_0), \tag{1.4}$$

is dense[5] in A if

$$\mathcal{L}(\boldsymbol{x}_0) = A, \tag{1.5}$$

where $\mathcal{L}(\boldsymbol{x}_0)$ is the set of limit points of $\mathcal{O}(\boldsymbol{x}_0)$ [43]. This set of limit points is defined to be [22]

$$\mathcal{L}(\boldsymbol{x}_0) = \left\{ \boldsymbol{y} : \text{for some } \boldsymbol{x} \in \mathcal{O}(\boldsymbol{x}_0) \; \exists \; \{\boldsymbol{\varphi}_{t_i}(\boldsymbol{x})\}_{i \geq 1} \; s.t. \; \lim_{t_i \to \infty} \boldsymbol{\varphi}_{t_i}(\boldsymbol{x}) = \boldsymbol{y} \right\}. \tag{1.6}$$

The basin of attraction is the set of initial conditions that are trapped onto A, i.e.[22]

$$A = \bigcap_{t \geq 0} \boldsymbol{\varphi}_t(U). \tag{1.7}$$

For some systems, the geometry of the set A can be very interesting, with different length scales exhibiting self-similar structures [42]. We will discuss this for the Lorenz system [37] in chapter 4. For classical parameters, integrations of the MS system yield solutions that occupy a bounded region in phase space.

One way of disentangling the structure of an attractor A is to look at the Poincare map of the underlying flow. Consider a periodic orbit of the flow $\boldsymbol{\varphi}_t$ and take a local cross section $\Sigma \subset \Re^m$ of dimension $m-1$. The hyper-surface Σ must be chosen such that the flow is everywhere transverse to it and it is local in the sense that the flow should always cross it in the same direction. If we denote the point where the orbit intersects Σ by p and let U be some neighbourhood of p, then the first return or Poincare map $P : U \longrightarrow \Sigma$ is defined for a point $q \in U$ by [22]

$$P(q) = \boldsymbol{\varphi}_\tau(q), \tag{1.8}$$

[5]This is not the same as the topological definition of one set being dense in another.

where $\tau = \tau(q)$ is the time for the orbit $\boldsymbol{\varphi}_\tau(q)$ based at q to first return to U.

Although the models considered in this section are flows, the models of the circuit will be maps of the form

$$\boldsymbol{x}_{n+1} = \boldsymbol{\phi}(\boldsymbol{x}_n), \tag{1.9}$$

where $n = t/\delta t$ and δt is some uniform time discretesation. This should not worry us because the circuit dynamics are actually a flow. Even the numerical solution of (1.2) is obtained using a finite difference scheme, which is a map. We expect that in the limit of the discretisation time step $\delta t \to 0$, $\boldsymbol{\phi}^n(\boldsymbol{x}) \to \boldsymbol{\phi}_t(\boldsymbol{x})$. For a *perfect model*, defined below, $\boldsymbol{\phi}_t = \boldsymbol{\varphi}_t$. Otherwise the model is imperfect.

1.2 Predictability

Consider a system whose long term dynamics converge onto an attractor for some basin of attraction. Suppose the *perfect model* for this system is known exactly. By *perfect model* we mean that the model used is isomorphic[6] to the system we are modelling and let $\boldsymbol{\varphi}_t(\boldsymbol{x})$ be the underlying flow. For any initial condition $\boldsymbol{x}_0 \in A$, denote an ϵ ball centred at $\boldsymbol{x}_0$ by $B_{\boldsymbol{x}_0}(\epsilon)$. We say that $\boldsymbol{\varphi}_t(\boldsymbol{x})$ has *sensitive dependence on initial conditions* if there exits $\delta > 0$ such that for every initial condition $\boldsymbol{x}_0 \in A$ and any $\epsilon > 0$, we can find a state $\boldsymbol{y}_0 \in B_{\boldsymbol{x}_0}(\epsilon)$ such that for some $T > 0$, $t \in (0, T)$ implies $||\boldsymbol{\varphi}_t(\boldsymbol{y}_0) - \boldsymbol{\varphi}_t(\boldsymbol{x}_0)|| > \delta$ [43]. If $\boldsymbol{\varphi}_t(\boldsymbol{x})$ has sensitive dependence on uncertainties in the initial conditions, then the underlying system is said to be *chaotic*. Sensitivity on the initial conditions was experimentally observed by Shaw [16] on an analogue electronic circuit in 1977. Since one could not precisely set the initial conditions on

[6]Two mappings ϕ and φ are isomorphic if there is a one-to-one, invertible mapping h such that $\phi = h^{-1}\varphi h$.

the electronic circuit, each time they set the initial conditions as closely as possible to the previous one, they noticed that the resulting trajectories diverged from each other. One of the papers that gave an impetus to the study of chaotic dynamics was by Lord May [44] in 1976.

In chaotic systems, infinitesimally small initial uncertainties, $\boldsymbol{\epsilon}_0$, develop exponentially according to the relation

$$\dot{\boldsymbol{\epsilon}} = D\boldsymbol{F}(\boldsymbol{x})\boldsymbol{\epsilon}, \tag{1.10}$$

where $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(t)$, $\boldsymbol{\epsilon}_0 = \boldsymbol{\epsilon}(0)$ and D is the differential operator. However, as long as the uncertainties are infinitesimal, they place no finite limit on our ability to forecast the future evolution of the system.

This brings us to the notion of *predictability*. Suppose we have an initial observation of some $\boldsymbol{x}_0 \in A$ by some observation function $\boldsymbol{h}$, such that

$$\boldsymbol{h}(\boldsymbol{x}_0) = \boldsymbol{y}_0, \tag{1.11}$$

$$\boldsymbol{y}_0 \in B_{\boldsymbol{x}_0}(\epsilon), \tag{1.12}$$

where $\boldsymbol{\epsilon}$ is the observational uncertainty. According to Smith et al [67], in the perfect model scenario, predictability is lost either (i) when an initial uncertainty increases by a factor of q or (ii) when the forecast adds no new information to the *climatology*[7]. The factor q is a natural number.

Let us explain (ii) a little more. Consider an *invariant measure* [13], ϱ, so that

$$\varrho[\boldsymbol{\varphi}_{-t}(E)] = \varrho(E), \quad t > 0, \tag{1.13}$$

[7]Climatology is the long term distribution of the dynamics.

where $E \subset \mathbb{R}^m$ is a measurable set, $\boldsymbol{\varphi}_{-t}(E)$ is the set obtained by evolving each point in E backward. We define a probability measure ϱ on E by [13]

$$\varrho(E) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_E(\boldsymbol{\varphi}_t(\boldsymbol{x}_0))\mathrm{d}t. \tag{1.14}$$

and $\mathbf{1}_E$ is an indicator[8]. As long as A is ergodic[9],

$$\varrho(E) = \int_E \varrho(\mathrm{d}\boldsymbol{x}). \tag{1.15}$$

Associated with ϱ is a probability density function ρ such that (1.15) may be rewritten as [13]

$$\varrho(E) = \int_E \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1.16}$$

For $\boldsymbol{x}_0 \in A$ we can define (new) a probability measure associated with $B_{\boldsymbol{x}_0}(\epsilon)$ by

$$\varrho_0(E) = \lim_{T\to\infty} \frac{1}{T\varrho(B_{\boldsymbol{x}_0}(\epsilon))} \int_0^T \mathbf{1}_{E\cap B_{\boldsymbol{x}_0}(\epsilon)}(\boldsymbol{x}(t))\mathrm{d}t \tag{1.17}$$

This new measure induces some probability density function, $p_0(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$. We will call the set of points distributed according to the density $p_0(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$ a *perfect initial ensemble*. At any time t the forecast of the perfect initial ensemble will be distributed according to some probability density $p_t(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$. We will call the distribution $p_t(\boldsymbol{x}; \boldsymbol{x}_0, \epsilon)$ a *perfect forecast* at time t.

Predictability is lost when

$$\int p_t(\boldsymbol{x}; \boldsymbol{x}_0) \log p_t(\boldsymbol{x}; \boldsymbol{x}_0)\mathrm{d}\boldsymbol{x} \approx \int p_t(\boldsymbol{x}; \boldsymbol{x}_0) \log \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{1.18}$$

[8] An indicator is defined by

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

[9] In ergodic attractor, space averages are equal to time averages.

where the integrations in (1.18) are over $\mathbb{R}^m$ and it will be the case whenever the region of integration is not indicated. Notice here that we have used an approximation to give ourselves some freedom to choose the tolerance. In fact,

$$\int p_t(\boldsymbol{x};\boldsymbol{x}_0)\log p_t(\boldsymbol{x};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x} \geq \int p_t(\boldsymbol{x};\boldsymbol{x}_0)\log \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{1.19}$$

which is the Kullback-Leiber inequality[10] [34]. To prove (1.19), we need to first note that $\log\zeta \leq \zeta - 1$ for $\zeta \geq 0$ with equality only when $\zeta = 1$. We then use this inequality to get

$$\int p_t \log(p_t/\rho)\mathrm{d}\boldsymbol{x} = -\int p_t \log(\rho/p_t)\mathrm{d}\boldsymbol{x} \geq -\int p_t(\rho/p_t - 1)\mathrm{d}\boldsymbol{x} = 0.$$

The perfect model scenario has practical limitations because we often do not have the correct functional form of the model of the underlying system, in which case we are confronted with *model inadequacy*. Even we if we had the correct functional form, effectively meaning we knew the flow $\boldsymbol{\varphi}_t(\boldsymbol{x};\boldsymbol{\lambda})$, we would not know parameter values $\boldsymbol{\lambda}$ exactly, and we call this case *parametric uncertainty*. In literature [26], *parametric uncertainty* is referred to as the perfect model scenario, but we will call both model inadequacy and parametric uncertainty the *imperfect model scenario*.

We need a definition of the loss of predictability in the imperfect model scenario. Definition (i) cannot be used but (ii) can be modified slightly. If our model is imperfect, then the distribution of the forecast of the perfect initial ensemble $p_0(\boldsymbol{x}:\boldsymbol{x}_0)$ by the flow $\boldsymbol{\phi}_t(\boldsymbol{x})$ will not be the perfect forecast $p_t(\boldsymbol{x};\boldsymbol{x}_0)$. For this reason, the initial ensemble need not be perfect as long as our forecasting model is not perfect. If we

[10]This inequality may be deduced from the positive-definiteness of the *mean information for discrimination* [34], $I(1:2) = \int p_1(\boldsymbol{x})\log\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$, often mistaken to be the Kullback-Leiber divergence term, $J(1,2) = \int (p_1(\boldsymbol{x}) - p_2(\boldsymbol{x}))\log\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$. However, $J(1,2) = I(1:2) + I(2:1)$.

denote by $f_t(\boldsymbol{x};\boldsymbol{x}_0)$ the probability density of the forecast, regardless of whether the initial ensemble is perfect or not, then we say that predictability is lost if

$$\int p_t(\boldsymbol{x};\boldsymbol{x}_0)\log_2 f_t(\boldsymbol{x};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x} \leq \int p_t(\boldsymbol{x};\boldsymbol{x}_0)\log_2 \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1.20}$$

The operational evaluation of (1.20) will be addressed in chapter 5. We may also define the time t when predictability is lost on the average to be the minimum time for which

$$\frac{1}{T}\int_0^T \left\{\int p_{\tau+t}(\boldsymbol{x};\boldsymbol{x}_\tau)\log_2 f_{\tau+t}(\boldsymbol{x};\boldsymbol{x}_\tau)\mathrm{d}\boldsymbol{x}\right\}\mathrm{d}\tau \leq \frac{1}{T}\int_0^T \left\{\int p_{\tau+t}(\boldsymbol{x};\boldsymbol{x}_\tau)\log_2 \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\}\mathrm{d}\tau \tag{1.21}$$

holds. In other words, predictability is lost when the model performs worse than the climatology. Reliable estimates of the quantities on either side of inequality (1.21) will be obtained as $T \to \infty$.

1.3 Uncertainty Directions

To investigate the first kind of predictability, we will need to carefully choose the initial uncertainty directions. We shall concern ourselves with three orientations of initial uncertainties. The first two are based on the solution to the coupled system of equations (1.2) and (1.10), called the *linear propagator* [49],

$$\mathcal{M}(\boldsymbol{x}_0,\Delta t) = \exp\left(\int_{t_0}^{t_0+\Delta t} D_{\boldsymbol{x}}\boldsymbol{F}dt\right). \tag{1.22}$$

It maps $\boldsymbol{\epsilon}_0$ to $\boldsymbol{\epsilon}(t)$ via

$$\boldsymbol{\epsilon}(t_0+\Delta t) = \mathcal{M}(\boldsymbol{x}_0,\Delta t)\boldsymbol{\epsilon}_0. \tag{1.23}$$

The linear propagator governs the evolution of an infinitesimal perturbation of the initial conditions. The singular value decomposition of $\mathcal{M}$ is $\mathcal{M} = U\Sigma V^T$ with orthogonal matrices V (U) containing the right (left) singular vectors as columns and

Σ the diagonal matrix of the singular values, σ_i with $\sigma_i \geq \sigma_j$ for $i < j$.

The first uncertainty direction we shall define is called the *Lyapunov direction*. At a fixed location, $\boldsymbol{x}_0$, the *Lyapunov vector* is determined from the singular value decomposition of $\lim_{\Delta t \to -\infty} \mathcal{M}(\boldsymbol{x}_0, \Delta t)$ [67, 77]. Then the Lyaponov direction (Lyapunov vector), $\boldsymbol{u}_1$, is the first column of U. The *right singular vector direction*, $\boldsymbol{v}_1$, is the first column vector of V, the set of right orthogonal vectors in the singular value decomposition of $\mathcal{M}(\boldsymbol{x}_0, \Delta t)$, where Δt is chosen according to the local dynamics [67, 77]. The singular vector direction is also called the *maximal direction* because error growth is fastest in this direction [77]. At each point $\boldsymbol{x}(t)$ on the attractor, the Lyapunov and maximal uncertainty directions will be $\boldsymbol{u}_1(t)$ and $\boldsymbol{v}_1(t)$ respectively. Alternatively, we could choose the initial uncertainty direction randomly, which we call the *random direction*. All these uncertainty directions will be used in chapter 4. The random direction will be drawn from a multi-variate Gaussian distribution with a diagonal covariance matrix.

How do we compute the linear propagator in practice? Eckmann and Ruelle [13] have discussed this question in their seminal 1985 paper. Let $T^t_{\boldsymbol{x}(0)}$ be the Jacobian matrix of $\boldsymbol{x}(t)$ with respect to the initial conditions. That is, its entries are $T^t_{ij} = \frac{\partial x_i(t)}{\partial x_j(0)}$. Thinking of $\boldsymbol{x}$ as a function of t and $\boldsymbol{x}(0)$, differentiate (1.2) with respect to $x_j(0)$ to get

$$\frac{\partial^2 x_i}{\partial t \partial x_j(0)} = \sum_k \frac{\partial F_i}{\partial x_k(t)} \frac{\partial x_k(t)}{\partial x_j(0)},$$

which is equivalent to

$$\frac{\partial}{\partial t}\left(T^t_{ij}\right) = \sum \frac{\partial F_i}{\partial x_k}\Big|_t T_{kj}. \tag{1.24}$$

Equation (1.24) is equivalent to the matrix equation

$$\frac{dT^t_{\boldsymbol{x}(0)}}{dt} = (D_{\boldsymbol{x}(t)}\boldsymbol{F})T^t_{\boldsymbol{x}(0)}. \tag{1.25}$$

Equation (1.25) can be solved subject to the initial condition $T^0_{\boldsymbol{x}(0)} = I$, where I is an $m \times m$ identity matrix. So, the m equations given by (1.2) are supplemented with m^2 other equations (1.25) and the $m + m^2$ may be solved numerically. For large t, the eigenvalues of $(T^t_{\boldsymbol{x}})^* T^t_{\boldsymbol{x}}$ ('*' is transpose) have very different orders of magnitude and this creates numerical problems when $(T^t_{\boldsymbol{x}})^* T^t_{\boldsymbol{x}}$ is diagonalised. We mitigate this problem by choosing a unit of time τ_s so that the eigenvalues, $e^{\lambda_i \tau_s}$, do not differ too much in their orders of magnitude. In practice, one can choose τ_s such that the largest eigenvalue is of the order of 10 and the smallest is of the order of 10^{-1}. The τ_s must not be too small either since we will have to multiply a number of matrices proportional to τ_s^{-1}. The chosen τ is then used to discretise the time, setting $\tilde{\boldsymbol{\varphi}} = \boldsymbol{\varphi}_{\tau_s}$, and then proceeding as in the discrete case. $\tilde{\boldsymbol{\varphi}}$ then corresponds to the time one map. In solving the $(m + m^2)$ equations in the different intervals, one should set the initial conditions of $T^{\tau_s}_{\boldsymbol{x}}$ to I at every time step before performing the integration. Let us denote by $T^{(i)}$ the $T^{\tau_s}_{\boldsymbol{x}}$ computed in the ith time interval. Then the linear propagator, $\mathcal{M}(\boldsymbol{x}_0, \Delta t)$ may be approximated by

$$\mathcal{M}(\boldsymbol{x}_0, \Delta t) = \prod_{i=1}^{n} T^{(i)}, \tag{1.26}$$

where n is the number of time intervals.

1.4 Recurrence

One of the attributes of attractors is encapsulated by Poincaré's recurrence theorem which is one of the precious gems that emerged from his famous Three Body Problem[6]. It is stated as follows:

Theorem 1 *In an attractor, for any region r_0, however small, there will be trajectories which traverse it infinitely often. That is to say, in some future time, the system will return arbitrarily close to its initial situation and will do so infinitely often.*

The return of trajectories to any region r_0 is called recurrence and the time for this to happen is called *recurrence time*. To put this another way, let us consider the system whose attractor is A. The return time of a point $\boldsymbol{x} \in A$ to the ball $B_{\boldsymbol{x}}(\epsilon)$ is defined to be

$$\tau_r(\boldsymbol{x}, \epsilon) = \inf\{T : \varphi_T(\boldsymbol{x}) \in B_{\boldsymbol{x}}(\epsilon) \text{ and } \varphi_t(\boldsymbol{x}) \notin B_{\boldsymbol{x}}(\epsilon) \text{ for some } t \in (0, T)\} \quad (1.27)$$

We can then define the recurrence time of the system to be

$$\tau_r = \lim_{T\to\infty} \frac{1}{T} \int_0^T \lim_{\epsilon\to 0} \tau_r(\boldsymbol{x}(t), \epsilon) \mathrm{d}t. \quad (1.28)$$

Operationally, one needs to choose the ϵ such that it is slightly bigger the noise level and not smaller than the spacing between the data points[11]. We want the observation time, T, to satisfy the condition $\tau_r \ll T$ for us to see the overall behaviour of the system.

1.5 Organisation

Thus far we have set the framework within which we want to study the circuit in the imperfect model scenario. The ultimate aim is to combine the models that exhibit diversity in their behaviour across the attractor. We now give the general organisation of the thesis.

[11] This will be explained more in chapter 3

In chapter 2, the circuit is introduced. Issues that had to be born in mind during the design are highlighted in § 2.1. The model of the circuit based on Kirchhoff's laws is presented in § 2.2 and the values of the components chosen are given. These values correspond to parameter values $0 \leq \gamma \leq 5$ and $\Gamma = 10$ in the MS equations and at these values the numerical solution always settles onto a periodic orbit. Therefore, the MS system at these parameter values is not an appropriate model for the circuit because it manifests chaotic behaviour. Conditions under which the experiment was run and the data was acquired are discussed in § 2.3. A temperature proxy suggests that there was temperature drift during the run, which could have caused the circuit dynamics to drift with time. Nevertheless, in § 2.4 the circuit is seen to yield a next return map very similar to that of the MS system in the chaotic regime of the classical parameters.

An alternative way of modelling the circuit without appealing to the theory of electronics is presented in chapter 3. The approach is to construct models from data based on either the local dynamics or global dynamics. In particular, § 3.2.2 presents the Kwasniok-Smith algorithm and shows how it may be used to detect drift in the dynamics. The models that have been constructed for use in subsequent chapters are tabulated in § 3.3, all of which are global in space.

Chapter 4 discusses computational results of quantifying model behaviour in both the circuit and the MS system in the sense of q-pling times[12]. The perfect and imperfect model scenario are compared and contrasted. In the perfect model scenario,

[12]The time it takes for an initial uncertainty to multiply by a factor of q

the Lorenz system is also considered because it was in its study that the question of why the histogram of doubling times yielded an oscillatory pattern was posed. § 4.4 considers the implications, on distributions of q-pling times, of applying Takens theorem in the modelling of the circuit by delay reconstructions.

Probabilistic forecasting is discussed in chapter 5. The crucial aspects are how to choose the size of the initial uncertainty ball and interpret the resulting ensemble forecasts. Numerical experiments are first performed in the perfect model scenario and then carried over to the imperfect model scenario. The principles applied are mainly drawn from information theory [60] and density estimation [52].

In chapter 6, we present a way of combining probabilistic forecasts to take advantage of diversity in the models. The theory is then applied to circuit models. The final chapter, which is chapter 7, gives a summary of the conclusions and suggestions for further work.

Except the last chapter, each chapter ends with a conclusion. Each of these conclusions contains a list of things that are new in that chapter for ease of identification.

Chapter 2

The Moore-Spiegel Circuit

In this thesis, we deal with the modelling of a laboratory-made electronic circuit. The circuit data exhibit low noise and the observation period is relatively much longer than the recurrence time. Our data comprises voltages measured at the points on the circuit. The crucial design issues of the circuit are highlighted in § 2.1. One of the difficulties often cited in electronic experiments is the possibility of temperature driving the circuit from one state to another. This has caused some to go to great lengths of running the experiments in temperature controlled refrigerators or ovens. While some attention is paid to control and monitor the ambient temperature, we argue that careful design and component choice may suffice to keep the circuit stable over the time scale of the experiment. This section ends with a summary of the key issues to bear in mind when designing and building an electronic circuit.

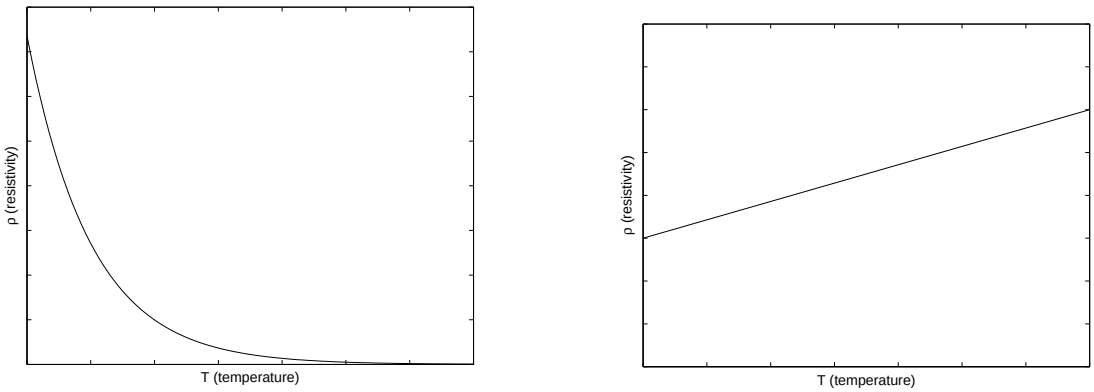
In § 2.2, the circuit is introduced. Since its non-dimensionalised model equations based on Kirchhoff's laws are the Moore-Spiegel (MS) system of equations, albeit with non-classical parameter values, it is called the Moore-Spiegel circuit. In the subsequent discussions, by MS-data we shall mean data obtained by numerically integrating the MS system with a Runge-Kutta 4-5 method. Data from the circuit

will simply be called circuit data or the circuit. Knowledge of circuit symbols and Kirchhoff's laws shall be assumed in the subsequent discussions.

§ 2.3, discusses the key experimental hurdles and how to clear them, together with the conditions surrounding the data collection. In particular, we talk about how the possible sources of noise and parametric drift were controlled. Finally, the circuit data is explored in § 2.4, comparing it with MS data. The MS data corresponding to the same parameter values as used in the circuit yields a periodic orbit, whereas the circuit only manifests a transient periodic orbit, and finally settles onto a chaotic attractor. However, small perturbations of parameters in the MS equations do not yield any chaotic behaviour at all and even the periodic orbits in question are not really similar to the one in the circuit. This should not be surprising because the circuit and the equations are different. For instance, the circuit has many degrees of freedom through which it can explore the parameter space, which could account for the dissimilarities manifested. The final section gives some concluding remarks and a list of things considered new in this chapter.

2.1 Design Issues

The main work-horse of the circuit under consideration is the Operational Amplifier and we shall spend some bit of time on practical implications of using it. Another crucial device used is the *multiplier*. The main concerns are *stability* and *saturation*. The former has implications on the long term dynamics of the circuit and the latter needs to be avoided if we are to have sensible behaviour in the circuit.



(a) Semiconductor (b) Conductor

Figure 2.1: Graphs showing how resistivity varies with temperature in both a semiconductor and a conductor.

Both an Op-Amp and multiplier are *semiconductor* devices. A *semiconductor* material like Silicon or Germanium [2, 4] is neither a good insulator nor a very good conductor of electricity. Its electrical *resistivity*, ρ, decreases strongly with increasing temperature, whilst that of a conductor increases weakly with temperature (see figure 2.1(a) and 2.1(b) respectively). Resistivity is the difficulty with which an electric current can pass through the material under the influence of an electric field. Insulators have very high resistivity, metals have very low resistivity and semiconductors have intermediate resistivity. Resistivity should not be confused with resistance, R, which is the constant of proportionality in Ohm's law[1]. In fact, for a material with cross-sectional area A, length l and resistivity ρ, the resistance is given by [51]

$$R = \frac{\rho l}{A}. \tag{2.1}$$

In the production of a semiconductor device, a semiconductor material is *doped*[2] with impurities, thus drastically modifying its properties.

[1]Ohm's law may be stated as $V \propto I$, where V is the voltage drop across a load and I is the current through the load.

[2]Doping is a process of deliberately adding known impurities in a controlled manner.

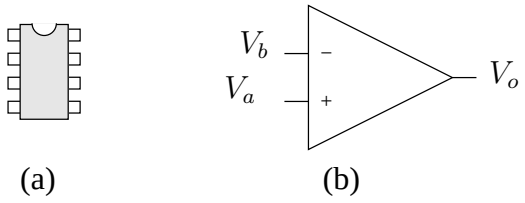


Figure 2.2: (a) Pictorial view of a physical chip and (b) circuit symbol of an OpAmp.

The OpAmps we used was AD712J, which is a 5 stage cascade network of transistors forming an Integrated Circuit (IC) on a chip. The circuit symbols and the diagram of a physical chip are shown in figure 2.2. V_a and V_b are gate inputs of two transistors at the *differential input stage* of the IC [3]. Typically, an OpAmp has very high input impedance, very high voltage gain and low output impedance[4]. For a moment let us see the implications of these properties. Let $V_{ab} = V_a - V_b$ and the open loop gain of the amplifier be A. Then

$$V_o = AV_{ab}. \tag{2.2}$$

However, in practice there is always some input offset voltage, δV_{ab}, whose inclusion in equation (2.2) gives

$$V_o = A(V_{ab} - \delta V_{ab}). \tag{2.3}$$

For the AD712J OpAmp that we used, it is given on the specifications form that $\delta V_{ab} \leq 0.3mV$ and $A = 4 \times 10^5$. This means if $V_{ab} = 0V$, we get the error in the output to be

$$\delta V_o \approx 12V,$$

[3] www.wikipedia.org/wiki/Operational_amplifier has a nice discussion of the different stages of an OpAmp.

[4] Impedance is resistance to an AC signal.

enough to cause *saturation*. Saturation is the situation whereby an increase in the input voltage does not lead to a corresponding increase in the output voltage. It is not a desirable scenario and needs to be avoided. The AD712J has been fitted with an internal trimming mechanism to counter this potential offset problem by the bound just used. There may also be input current offset, δI, which is also trimmed by the use of voltage trimming. Offset trimming is, however, complete at a fixed temperature because offset *drifts* with temperature, according to the relation [10]

$$\frac{d(\delta V_{ab})}{dT} = \pm\alpha, \tag{2.4}$$

where α is a constant and T is temperature. Drift introduces noise, setting a limit on the smallest value we can measure. This kind of noise is called *Flicker noise* [10]. For the AD712J, $\alpha \leq 7\mu V/^oC$. This means, with the given voltage gain, a change by 1^oC could results in the output change by $2.8V$, which is huge.

The situation is not that bleak, provided we use negative feedback. In particular, let us consider the feedback network shown in figure 2.3. As was the case with an open loop, we assume that the potentials at the OpAmp inputs are V_a and V_b and that the current offset is δI. Using the infinite gain approximation (IGA), we get the equation [10]

$$V_o = R_f\delta I - (1 - a_v)\delta V_{ab}, \tag{2.5}$$

where $a_v = -\frac{R_f}{R_b}$ is the closed loop gain and $R_a = \frac{R_f R_b}{R_f + R_b}$ is chosen to cancel the effects of bias currents. Obviously, to reap any rewards from using the feedback, we need to ensure that a_v is not too big. It is then evident that the use of feedback will downgrade the effects of drift and offset. It is still necessary, though, to ensure that voltage limits of the OpAmps are not exceeded to avoid saturation. Furthermore, if

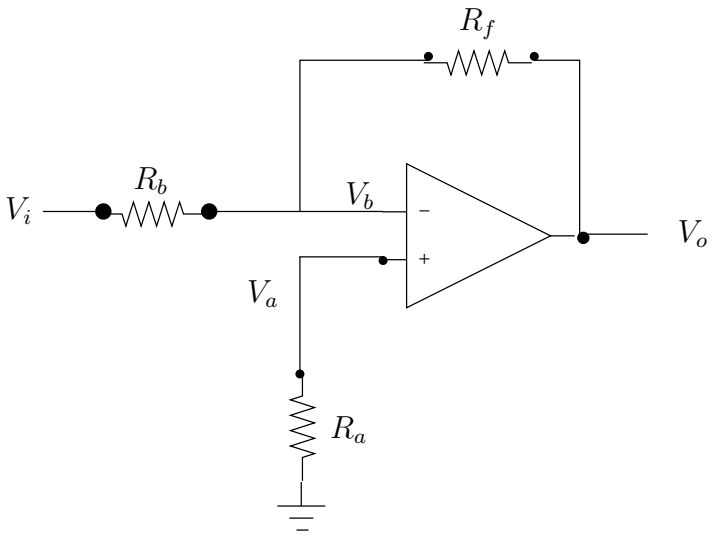


Figure 2.3: OpAmp negative feedback network

the input somehow changes in sign due to temperature, etc, the feedback will cause it to change in the opposite direction and maintain a fairly constant and stable output [2]. The value of using negative feedback cannot be over-emphasised.

We will now turn to the multipliers used, which were AD534J and internally trimmed to reduce offset effects. The functional block diagram is shown in figure 2.4. X_i, Y_i and Z_i are input terminals ($i = 1, 2$), V_o is the output, SF the scale factor, pretrimmed to 10.00V, and A is the open-loop gain. The transfer function is [5]

$$V_o = A\left(\frac{(X_1 - X_2)(Y_1 - Y_2)}{SF} - (Z_1 - Z_2)\right). \tag{2.6}$$

For the multiplier used, $A = 70$dB. This implies $A = 10^{3.5} \approx 3162$. The offset voltage of X_i and Y_i is bounded above by 20mV and that of Z_i by 30mV. This means if we perform multiplication by simply grounding Z_i, output offset could be as large as $\pm$90V, which will obviously cause saturation because the maximum permissible output is $\pm$12V. As was the case with the OpAmp, this problem can be circumvented

[5]Further details can be found in the data sheet at http://www.analog.com/en/prod/0,2877,AD534,00.html.

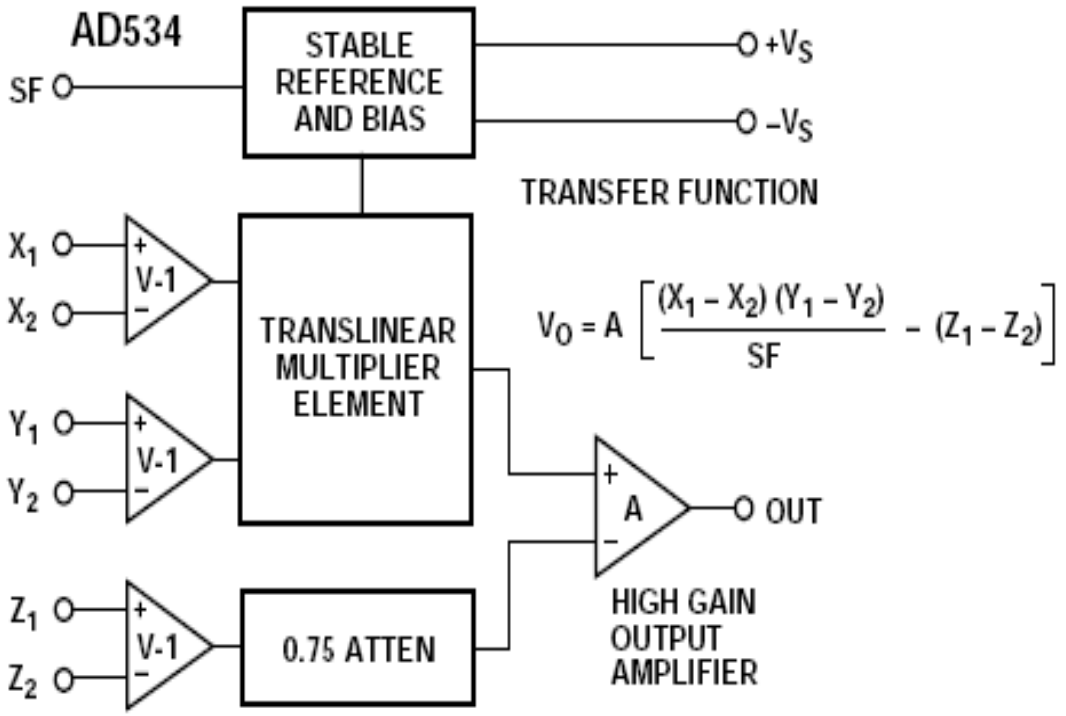


Figure 2.4: A functional block diagram of the multiplier

by the use of feedback. The feedback is performed by connecting Z_1 to V_o, so that (2.6) becomes

$$V_o = \frac{A}{A+1}\left(\frac{(X_1 - X_2)(Y_1 - Y_2)}{SF} + Z_2\right). \tag{2.7}$$

The infinite gain approximation yields $A/(A+1) \rightarrow 1$. Thus, the feedback also counteracts the effects of input voltage drift, which is $500\mu V/^oC$. Without feedback, a temperature change of 1^oC would account for a voltage drift of about $3V$. With feedback, the offset effects are expected to be bounded by about $1mV$.

We will conclude this section by mentioning other issues that have to be borne in mind when building an electronic circuit [39].

1. The number of active components should be kept minimum to minimise the extraneous noise introduced by the circuit itself.

2. The components used should operate under manufacturer specifications and various limits should not be exceeded.

3. At all points in the circuit, the signal should be larger than the background noise and drift produced by the circuit. In our particular case, not less than $5mV$.

4. Connections should be as short and direct as possible to reduce the effects of stray capacitance.

5. To speed up data acquisition, the circuit should run fast. This can be achieved by choosing capacitors such that the time constant is of the order of magnitude small enough to allow data collection.

The last point is ensured by choosing relatively small capacitor values because the time scale of the circuit, $\tau \propto C$. However, we must make sure that we are within the allowed frequencies by not violating the specified *slew rates*. For a given *OpAmp* or multiplier, slew rate is defined as

$$SR = \left.\frac{dV_o}{dt}\right|_{\max}, \tag{2.8}$$

where V_o is the output voltage. In our particular case, $SR = 20V/\mu s$.

In the next section, we shall introduce the MS-circuit and the corresponding model equations.

2.2 The Circuit

In this thesis, the physical system we shall primarily concern ourselves with is an electronic circuit whose diagram is shown in figure 2.5. We shall call it the Moore-

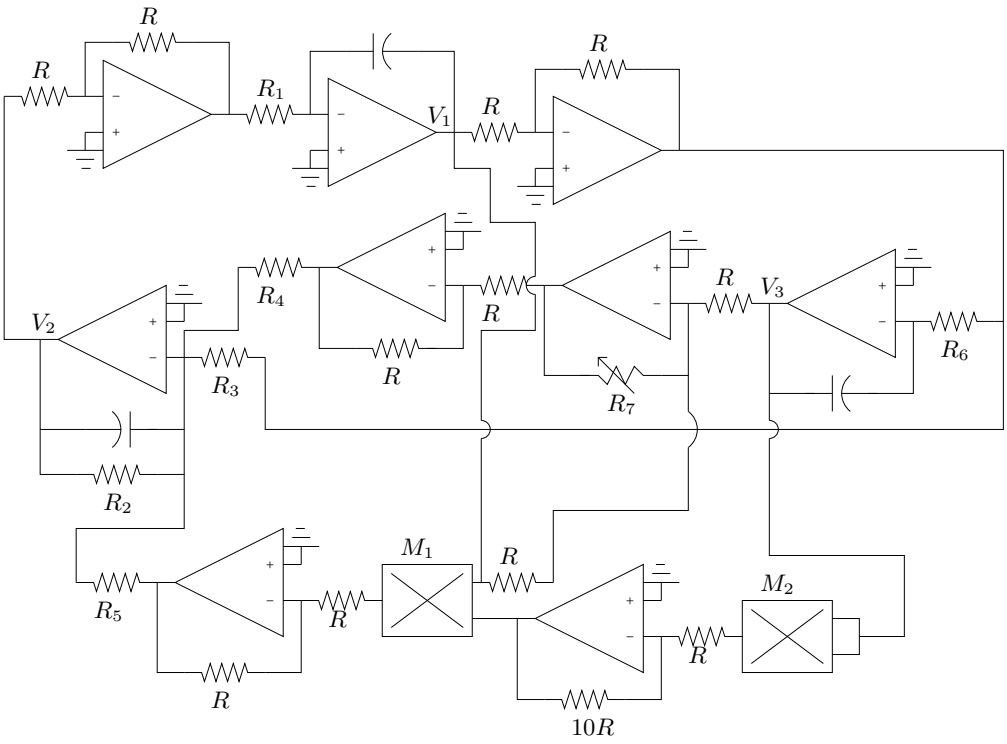


Figure 2.5: Moore Spiegel circuit diagram.

Spiegel circuit by virtue that, under a number of standard engineering assumptions[6], its rescaled Kirchhoff's equations are the Moore-Spiegel equations. It was built on a bread board using capacitors, resistors, operational amplifiers (OpAmp) and multipliers and the component values used are shown in table 2.1. The main work-horse in the design is the OpAmp. In the circuit shown in figure 2.5, R, R_i, $i = 1, .., 7$ are

$R = 10k\Omega$	$0 \leq R_1 \leq 10k\Omega$	$R_2 = 100k\Omega$
$R_3 = 100k\Omega$	$R_4 = 100k\Omega$	$R_5 = 10k\Omega$
$R_6 = 100k\Omega$	$0 \leq R_7 \leq 5k\Omega$	$C = 10nF$

Table 2.1: Table of component values used in the Moore-Spiegel circuit shown in figure 2.5.

resistors, C's are capacitors, and V_i, $i = 1, \ldots, 3$ are voltages. By applying Kirchhoff's

[6]Assumptions such as IGA, perfect multipliers, perfect integrators, etc. These assumptions could account for very different long term dynamics.

Current laws, we obtain the following system of equations:

$$\begin{aligned} R_1C\tfrac{dV_1}{dt'} &= V_2, \\ R_2C\tfrac{dV_2}{dt'} &= -V_2 + \tfrac{R_2}{R_3}V_1 - \tfrac{R_7R_2}{R_4R}(V_1+V_3) - \tfrac{R_2}{10R_5}V_1V_3^2, \\ R_6C\tfrac{dV_3}{dt'} &= V_1. \end{aligned} \tag{2.9}$$

For some scalar, σ, and time scale τ, we let $\sigma R_1C = R_2C = R_6C = \tau$. These are isomorphic to the Moore-Spiegel equations [47]:

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -y + \Gamma x - \gamma(x+z) - \Gamma x z^2, \\ \dot{z} &= x. \end{aligned} \tag{2.10}$$

To see this, let us re-scale the variables x, y, z, t by k_1, k_2, k_3, τ: $\hat{x} = x/k_1$, $\hat{y} = y/k_2$, $\hat{z} = z/k_3$ and $t' = t\tau$ to obtain

$$\begin{aligned} \tau\tfrac{d\hat{x}}{dt'} &= \tfrac{k_2}{k_1}\hat{y} \\ \tau\tfrac{d\hat{y}}{dt'} &= -\hat{y} + \tfrac{k_1}{k_2}\Gamma\hat{x} - \tfrac{\gamma}{k_2}(\hat{x}k_1 + \hat{z}k_3) - \tfrac{\Gamma k_1 k_3^2}{k_2}\hat{x}\hat{z}^2, \\ \tau\tfrac{d\hat{z}}{dt'} &= \tfrac{k_1}{k_3}\hat{x}. \end{aligned} \tag{2.11}$$

For the standard Moore-Spiegel parameter values, $\Gamma = 100$ and $0 < \gamma < 50$, we can choose parameter values $k_1 = k_3 = 1V^{-1}$, $k_2 = 100V^{-1}$ and $\tau = 10^{-3}s$ and perform the transformations $\hat{x} \to V_1$, $\hat{y} \to V_2$, $\hat{z} \to V_3$. $0 < R_7 < 5k\Omega$ corresponds to $0 < \gamma < 50$ and $R_1 = 1k\Omega$ corresponds to $\Gamma = 100 = k_2/k_1$. However, $R_1 = 10k\Omega$ corresponds to the Moore-Spiegel equations with parameters $\Gamma = 10$ and $0 \leq \gamma \leq 5$. This effectively means that the usual parameters are divided by 10 and the y variable is rescaled to $y' = y/10$. In particular, the first equation in (2.9) may, after applying the transformation $V_1 \to \hat{x}$, $V_2 \to \hat{y}$, be rewritten as

$$\frac{\tau}{10}\frac{d\hat{x}}{dt'} = \hat{y}. \tag{2.12}$$

Using $\hat{x} = x$, $\hat{y} = y/100$ and $t' = t\tau$ yields

$$\frac{dx}{dt} = \frac{y}{10}. \tag{2.13}$$

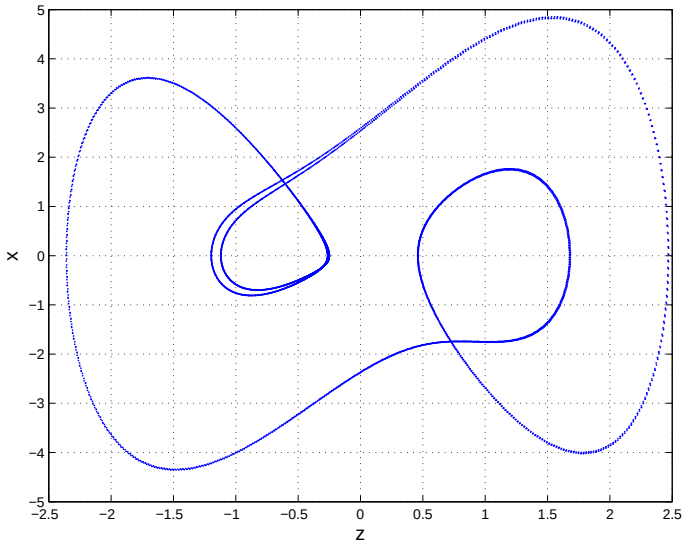


Figure 2.6: A periodic orbit of the MS-system at parameter values $\gamma = 3.2$ and $\Gamma = 10$.

For this set of parameter values, the MS-system settles on a periodic orbit, whereas the circuit yields chaotic behaviour. A typical periodic orbit of the MS system at parameters $\Gamma = 10$ and $\gamma = 3.2$ is given in figure 2.6. A bifurcation transition sequence for the modified MS-system is given in figure 2.7. It was obtained by integrating the modified MS system with the coefficient of y in the first equation of (2.10) being 10. However, we do not get such a transition sequence in the MS system at parameter values $0 < \gamma < 5$ and $\Gamma = 10$. Perturbations of $\Gamma = 10$ do not yield any chaotic behaviour either. This indicates that the MS-system and the circuit are different.

In the next section, we discuss the circuit build up and conditions surrounding the data acquisition.

2.3 Experiment and Data Acquisition

The procedure for building and testing the circuit requires a lot of care. One has to break up the circuit into modules that separately can be tested and then joined

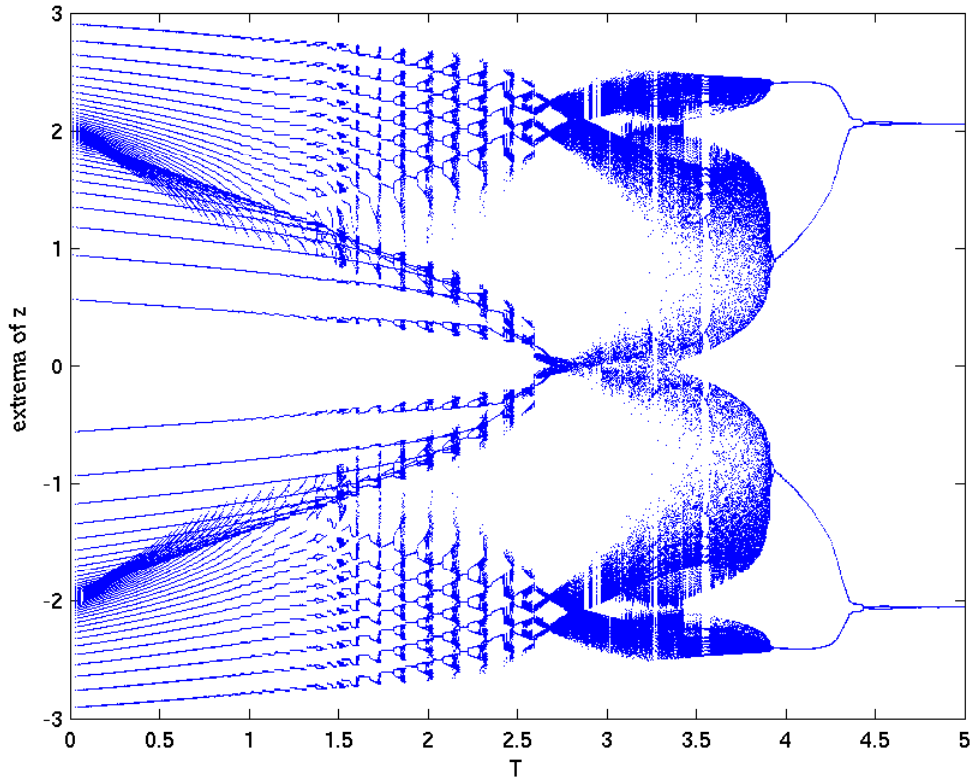


Figure 2.7: Bifurcation transition sequence of the modified MS system obtained by plotting the extrema of z against γ with $\Gamma = 10$ fixed and the coefficient of y in the first equation of (2.10) being 10.

together. Each of the integrators, adders, and multipliers[7] were tested to ensure that they were functional.

To test if an integrator works properly, we input a square wave with a signal generator and used an oscilloscope to check if the output was a saw-tooth wave. A square wave is a periodic function given by[8]

$$w_1(t) = \begin{cases} -1, & -\pi < t \leq 0, \\ 1, & 0 < t \leq \pi, \end{cases} \tag{2.14}$$

[7]See appendix B for what constitute each of these modules.

[8]A square wave of any other frequency can be rescaled to be written in this form. The units of time are dimensionless.

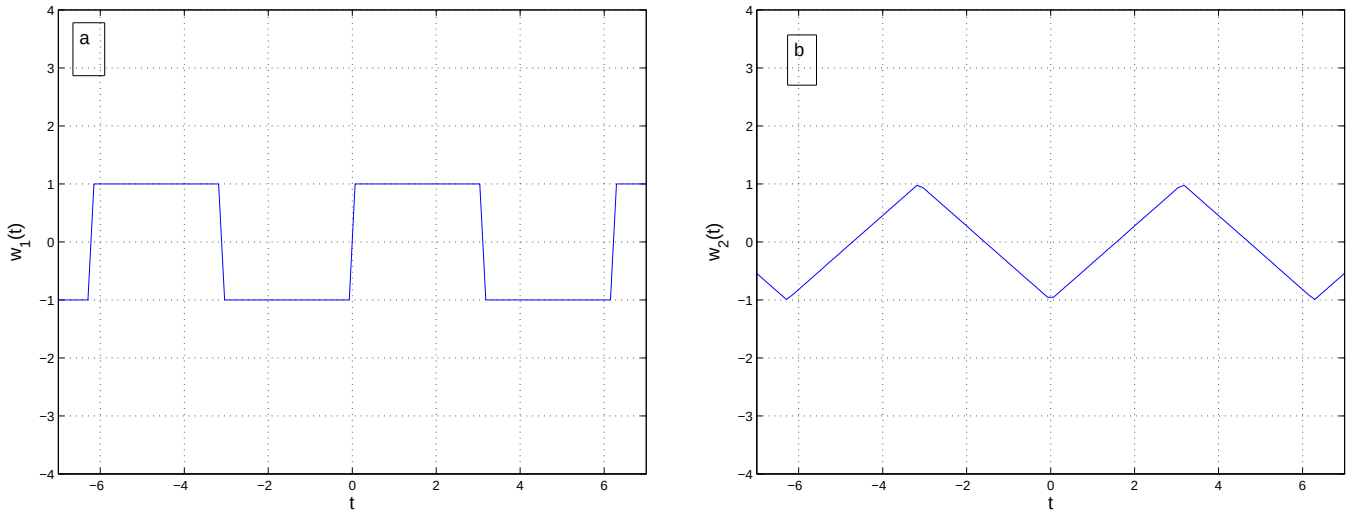


Figure 2.8: (a) The square wave that is input to an integrator from a signal generator and (b) the sawtooth wave expected at the output of the integrator if components are functional at optimum conditions.

with Fourier series expansion

$$w_1(t) = \frac{4}{\pi}\sum_{n=0}^{\infty}\frac{1}{(2n+1)}\sin(2n+1)t. \tag{2.15}$$

On the other hand, a saw-tooth wave is given by

$$w_2(t) = \begin{cases} -(t+\frac{\pi}{2}), & -\pi < t \leq 0, \\ t-\frac{\pi}{2}, & 0 < t \leq \pi, \end{cases} \tag{2.16}$$

with Fourier series expansion

$$w_2(t) = -\frac{4}{\pi}\sum_{n=0}^{\infty}\frac{1}{(2n+1)^2}\cos(2n+1)t. \tag{2.17}$$

It is clear that $w_2(t) = \int w_1(t)dt$. In figure 2.8, we show typical signals of $w_1(t)$ and $w_2(t)$ that one expects to see on the oscilloscope[9]. To check an adder, we use an oscilloscope to see if the amplitude of the output is the sum of the amplitudes of the constituent input signals, provided they are in phase. Finally, each multiplier module was checked by inputing one signal to both inputs of the multiplier and then plotting

[9]Not actual oscilloscope realisations.

the phase portrait of the input versus the output. If the multiplier works well, the oscilloscope plot should be a quadratic curve[10].

After testing each of the modules, one can then bring them together to build the circuit in figure 2.5. The circuit did not straight away yield behaviour that does not have saturation until re-adjusting resistor values for a while. It is crucial that one measures voltage outputs of the OpAmps to ensure that they do not saturate. If any of the output voltages remains flat (at $\pm 10V$) or clipped, then saturation has occurred and we need to re-adjust the resistor values of that OpAmp network to remove the saturation. In our case, we noticed that adjusting R_1 removed saturation. At $R_1 = 1k\Omega$, the typical Moore-Spiegel system, the circuit saturated and $R_1 = 10k\Omega$ removed the saturation. We, therefore, considered the circuit with parameter values shown in table 2.1 and $R_1 = 10k\Omega$ fixed.

For data acquisition, we used an instrument called a Microlink 770 [11] which is a 16-bit analogue to digital converter, able to sample data at frequencies of up to 100 kHz. Data was collected at a frequency of 10 kHz. To choose a sampling frequency, we sought one that ensured a smooth signal without over-sampling. To this end we applied Nyquist-Shannon sampling theorem [61]. The theorem says that if a signal is band-limited, with the upper bound of frequencies exhibited being B, then the condition for reconstructability from samples at a rate of f_s is that

$$f_s > 2B. \tag{2.18}$$

[10]It has been pointed out that this could fail if the transfer function yields x^3/y for and any two inputs x and y.

[11]Supplied by Microlink Engineering Solutions (http://www.microlink.co.uk/770.html).

$2B$ is called the Nyquist rate. For the circuit, an estimate of $B \approx 0.3kHz$.

It has already been mentioned that the behaviour of semiconductor devices, of which OpAmps and multipliers are examples, depend strongly on temperature changes [4]. For this reason, we used a temperature dependent resistor to measure voltage changes as proxy for the ambient temperature of the circuit. We measured V_1, V_2, V_3 corresponding to x, y, z and the temperature proxy, T, was also measured in volts. The data was collected while the circuit was encased in a metallic box which was then placed in a bigger and insulated box. By boxing the circuit we wanted to control the offset drift. Data sets saved in files, Setj.txt, $j = 1, \ldots, 6$, and Setj.imx, $j = 7, \ldots, 9$, were collected with R_7 set to $3.85k\Omega$. From here on, we shall refer to these data sets simply as setj, $j = 1, \ldots, 9$. Data setj, $j = 7, \ldots, 9$, were collected over a duration of about 14 hours while the shorter data sets corresponding to $j = 1, \ldots, 6$ were collected over approximately one hour.

Set7

Collection of Set7 started soon after the circuit was switched on. During the first three hours, the air-conditioner was on and the room was open until my colleagues had left, at which time the door was closed and the air-conditioner switched off.

Set8

The circuit was allowed to run for several minutes before data collection and the air-conditioner was switched on about 45 minutes before collection to allow the room temperature to stabilise. Collection ran over-night. The following morning, it was

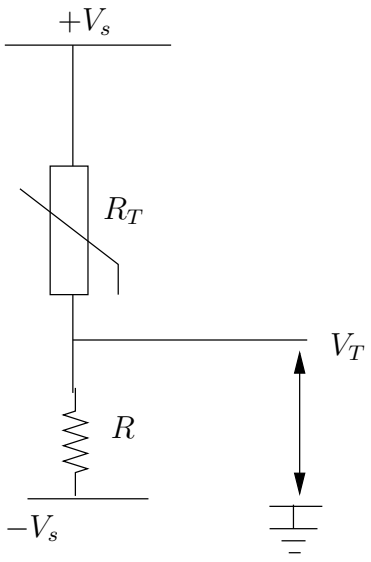


Figure 2.9: Voltage divider network used to monitor circuit ambient temperature changes.

noticed that the box containing the encased circuit was not entirely sealed.

Set9

The conditions surrounding the collection of this data set were made to be like those of Set8, but ensuring that the insulating box was entirely sealed.

In our subsequent discussions our attention shall focus more on the three sets just explained. In the next section, we look at these data sets and compare them with the MS-system.

2.4 Circuit Data Exploration

The circuit ambient temperature was monitored using a voltage divider network whose simplified version is shown in figure 2.9. R_T is a thermistor, a temperature dependent resistor which we attached to the body of the metal encasing (making sure that a good thermal contact was made), R is a fixed colour coded resistor, V_s is the supply

DC voltage, and V_T is the voltage potential (relative to ground) at a point between R and R_T. It follows from Kirchhoff's laws that

$$V_T = \frac{(R - R_T)}{R + R_T} V_s. \tag{2.19}$$

Changes in the ambient temperature were then monitored by monitoring changes in V_T because V_s and R are fairly stable. In fact, to first order approximation,

$$\Delta R_T = -2R \frac{V_s \Delta V_T}{(V_T + V_s)^2}. \tag{2.20}$$

Over a small range of temperatures, the thermistor may be assumed to vary linearly with temperature according to

$$\Delta R_T = k \Delta T, \tag{2.21}$$

where k is a constant, called *temperature coefficient*. If k is positive, then the thermistor is said to have a positive temperature coefficient, PTC, and if it is negative, it is said to have a negative temperature coefficient, NTC. Over a wide range of temperatures, the Stein-hart equation [69] is more appropriate and a special form of it is given by

$$R_T = R_{T_0} \exp\left[\frac{\beta(T_0 - T)}{T_0 T}\right], \tag{2.22}$$

where T_0 is some standard temperature, R_{T_0} is the resistance of the thermistor at T_0. The temperature is in Kelvin and T_0 is usually $298.15K$. The thermistor used was an NTC with $R_{T_0} = 100k\Omega$, $T_0 = 298.15K$ and $\beta = 4450$. The temperature proxies for the data sets 7 to 9 are shown in figure 2.10. In each case,

$$T_i = V_T^{(i)}, \tag{2.23}$$

where $V_T^{(i)}$ is the V_T measured during the collection of Set i. Its rises and falls reflect the increases or decreases in the ambient temperature. T_9 seems to be still noisy

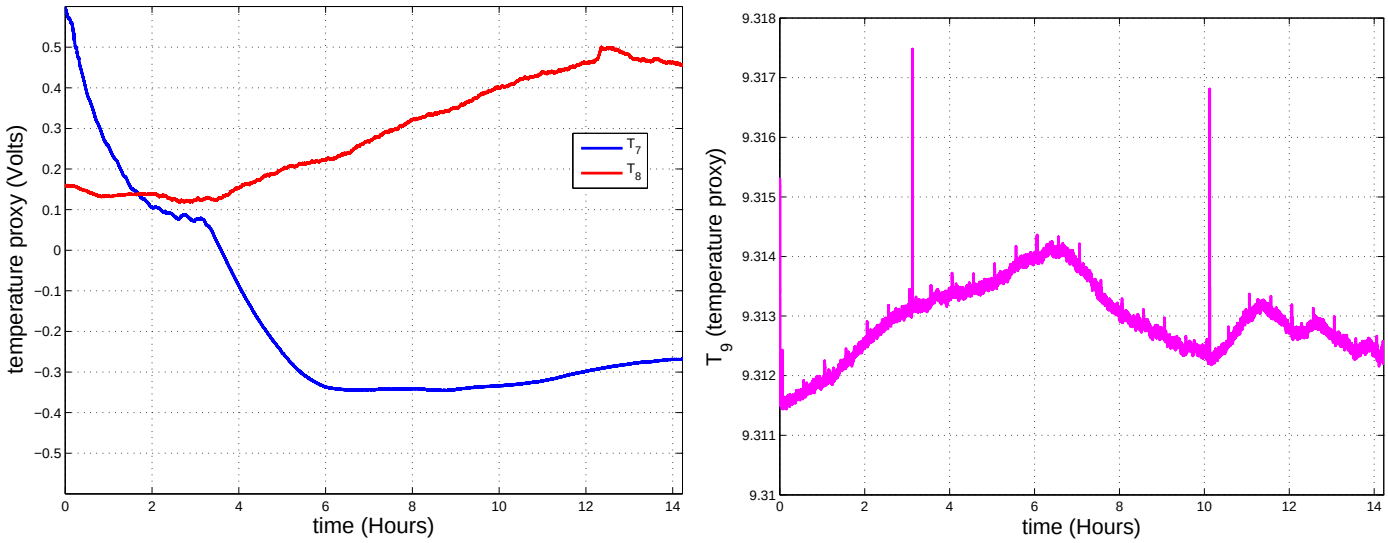


Figure 2.10: Smoothed (using the same filter) temperature proxies for the long circuit data sets. T_7 (for Set7), T_8 (for Set8) and T_9 (for Set9) are the temperature proxies, with $T_i = V_T^{(i)}$.

relative to T_7 and T_8. In fact, there was a loose connection leading to wrong values of V_T. To get an idea of the underlying temperature fluctuations, we need to transform a given voltage change. Equation (2.22) may be rearranged into

$$T = \left[\frac{1}{\beta}\ln\left[\frac{R_T}{R_{T_0}}\right] + \frac{1}{T_0}\right]^{-1}, \tag{2.24}$$

from which we get

$$\begin{aligned} \Delta T &\approx \frac{\partial T}{\partial R_T}\Delta R_T \\ &= -\left[\frac{1}{\beta}\ln\left[\frac{R_T}{R_{T_0}}\right] + \frac{1}{T_0}\right]^{-2}\frac{\Delta R_T}{\beta R_T}. \end{aligned} \tag{2.25}$$

Notice that the maximum voltage swing of T_7, say, is $\Delta V_T^{(7)} = -1V$. Therefore, we can use (2.20) to get $\Delta R_T = 20k\Omega$, taking $R = 100k\Omega$ and $V_s = 10V$. Substituting these values into (2.25) yields

$$\Delta T \approx 3.54^oC,$$

where we used $T = 296K$, the temperature at which the air-conditioner was set. This is the bound on the temperature change during the collection of each data set.

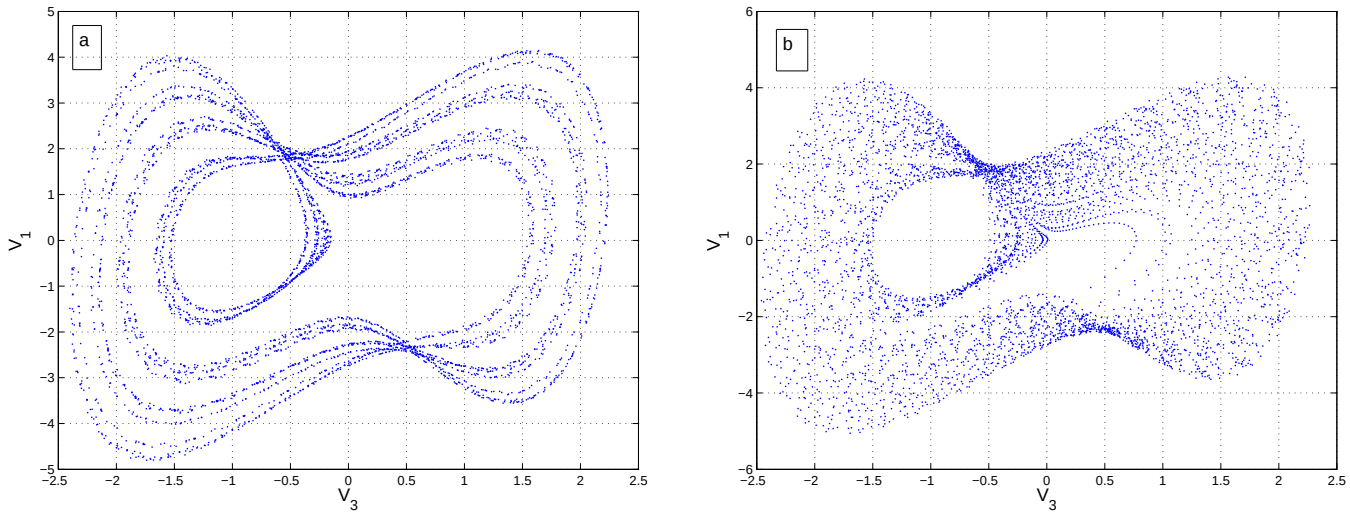


Figure 2.11: Projections onto the (V_3, V_1)-plane of the dynamics visited by circuit (From Set7). (a) Typical orbit mostly manifested in the first 3 minutes and (b) Typical attractor that the circuit finally settles in. The points plotted were uniformly spaced in time.

Let us now turn to projections of the phase portraits of the voltages onto the (V_1, V_3)-plane in the different temperature regimes. In the first 3 minutes of Set7, the dynamics are dominated by the object shown in figure 2.11(a). The MS-system at parameter values corresponding to the circuit never exhibited chaotic behaviour [12]. The periodic orbit that is manifest in the circuit is only transient for about 3 minutes, after which the dynamics become chaotic (typically like figure 2.11(b)). Transient dynamics do not have to be a result of parametric drift, but can also happen when the initial conditions do not lie on the attractor. In fact, for Set8 and Set9, which were collected after the circuit had run for a while, the dynamics are typically chaotic. A visual inspection of the phase portraits in the different temperature regimes suggests that circuit attractors were qualitatively similar. The projections of the chaotic attractor exhibited by the circuit onto the (V_1, V_2)-plane bears a lot of resemblance to the

[12] We integrated the system using various, randomly chosen initial conditions and parameter perturbations.

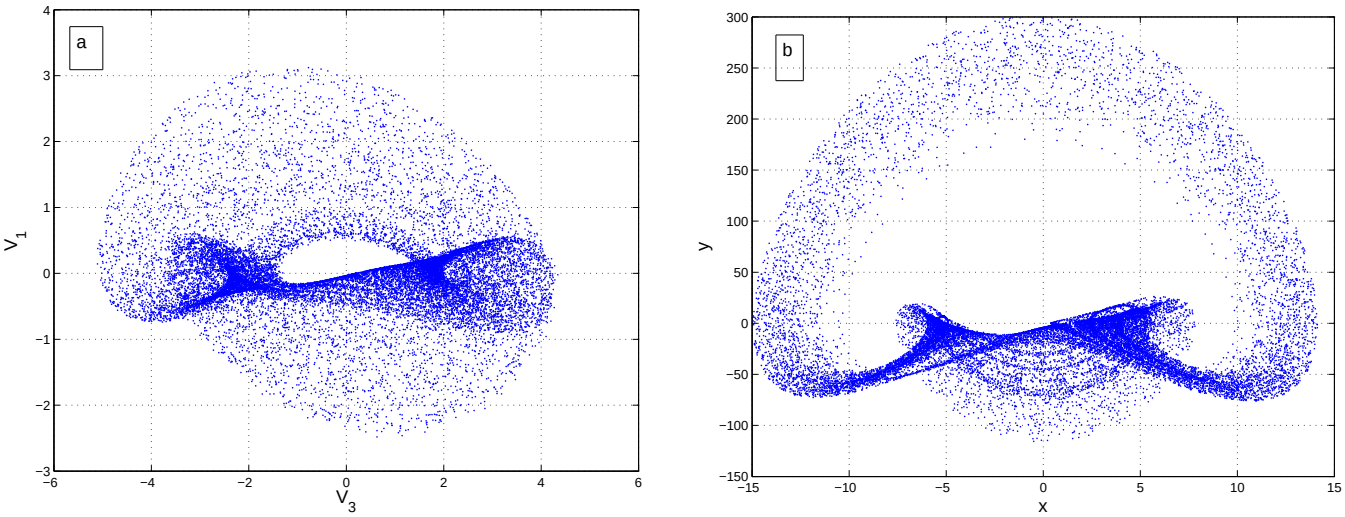


Figure 2.12: Projections of the (a) circuit (Set7) and (b) MS attractor onto the (V_1, V_2) and (x, y)-space respectively. The MS-system was integrated at parameter values $\gamma = 36, \Gamma = 100$. Each plot contains 4000 data points.

MS-system at parameter values, say $\gamma = 36, \Gamma = 100$ projected onto the (x, y)-plane. These projections of the circuit and MS system onto the (V_1, V_2) and (x, y)-plane are shown in figures 2.12. More similarities are manifest when we look at the Poincaré return maps of successive maxima of V_3 (resp. z), which were plotted in figure 2.13. There are also marked differences between the two return maps, with the left arm of MS-return map folding upwards. The return map of the circuit shows more points above the general U-shape and extra arm missing some portion on the left. The circuit time series in four consecutive time epochs is shown in figure 2.14.

The main points of this section could be summed up as follows: The model equations of the circuit obtained by applying Kirchhoff's laws yield dynamics that differ widely from the observed circuit dynamics. Even the transient, seemingly periodic orbit observed in the circuit does not look like that obtained from the model equations. Except for the transient orbit, the phase space trajectories over the different

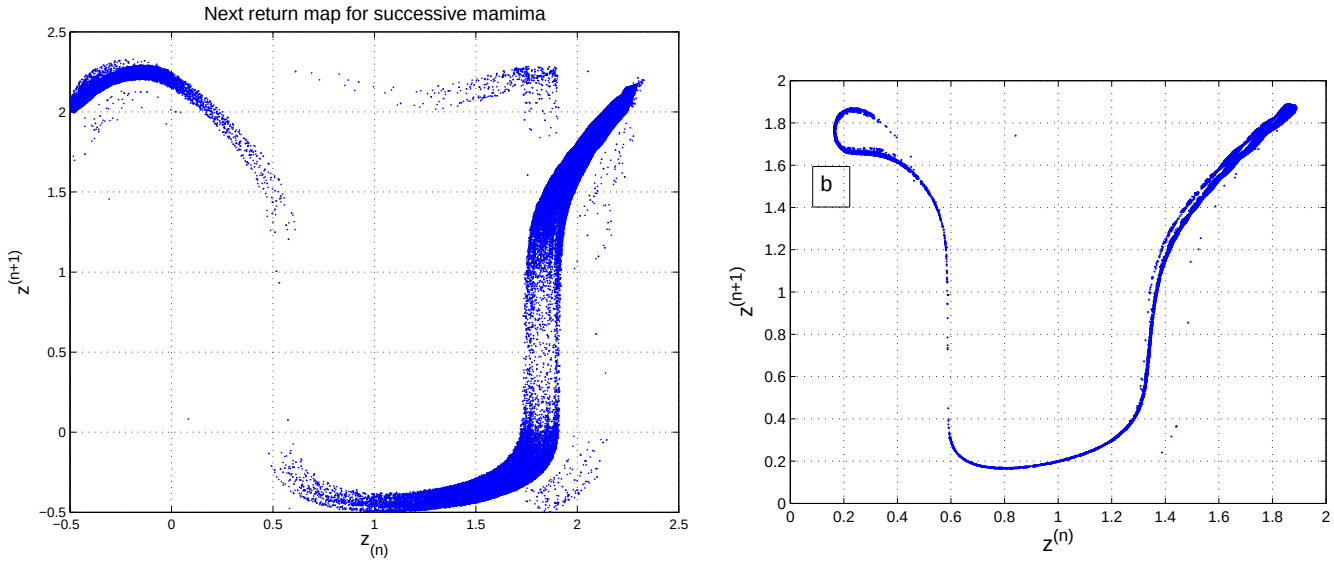


Figure 2.13: First return map of successive maxima of (a) V_3 (for Set7 with the periodic orbit removed) and (b) z. In each case, $V_3^{(n)}$ and $z^{(n)}$ are each the nth successive local maximum. Notice the overall similarities between the two pictures. There are also marked differences between the two maps, with the left arm of return map for MS system folding upward. The circuit return map appears to have some extra portions.

temperature regimes are strikingly similar. We shall revisit this aspect in chapter 4.

2.5 Conclusions

This chapter introduced the circuit that is central to this thesis. By the use of negative feedback in the integrators, adders and multipliers, it was argued that the circuit's stability properties are enhanced by reducing the effects of drift. For the same parameter values, the circuit manifests chaotic behaviour whereas the MS system only yields a periodic orbit. Therefore, in order to use the MS-system as a model for the circuit, effort would first have to be spent in finding "good"[13] parameter values. Small perturbations of the parameters in the MS-system still persist to yield periodic behaviour, which urges us to take a step back and rethink a way to model the circuit. The next chapter addresses the modelling question by a data based approach.

[13]What constitutes good will depend on what the modelling aim is.

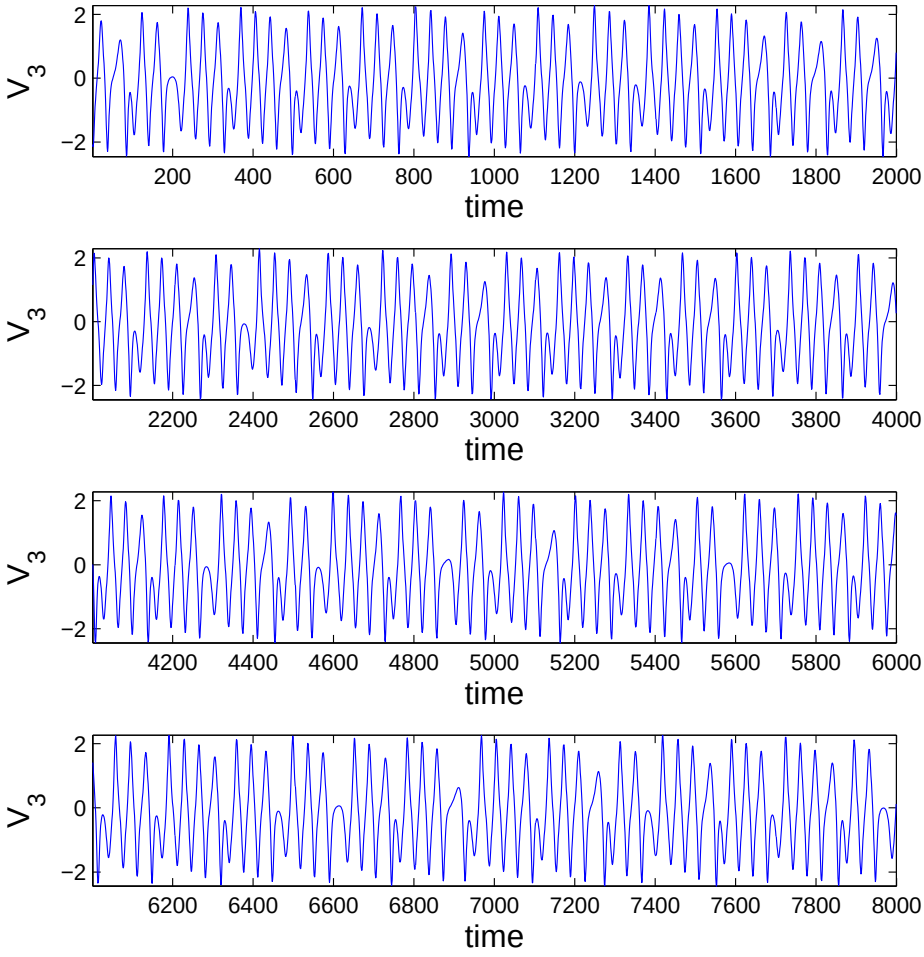


Figure 2.14: Circuit time series for Set7 in four consecutive time epochs with $R_1 = 10k\Omega$ and $R_7 = 3.85k\Omega$.

Notwithstanding the model inadequacies of the MS system, there are still some resemblances to the circuit that were found at the classical parameter values. In particular, the next return maps of the circuit and the MS-system at parameter values $\gamma = 36$ and $\Gamma = 100$ are strikingly similar, hinting that the circuit has some features of the MS-system.

The temperature was monitored by a thermistor, and it was estimated that, over the duration of acquisition of data sets Set7 and Set8, it changed by approximately 3.54^oC. There was no visually discenable change in the circuit dynamics over the duration of the data collection, except for the initial transient orbit. This point shall be revisited in chapter 4.

The following things are new in this chapter:

- The circuit.
- The circuit exhibiting chaotic behaviour whereas the MS system for the corresponding parameter values does not.
- The next return maps of both the MS-system and the circuit.
- Data sets from the circuit.

Chapter 3

Dynamical Models

In the previous chapter, we saw that the Moore-Spiegel equations are not a faithful model of the circuit. Although there are many similarities in the underlying attractors, the regions they each occupy and the amplitudes of oscillation are different, to say the least. While acknowledging that the MS-system is an imperfect model of the circuit, we could seek the coefficients $\alpha_i, \beta_j > 0$, $i = 1, \ldots, 4$ and $j = 1, 2$ such that the system

$$\begin{aligned} \frac{dV_1}{dt} &= \alpha_1 V_2 \\ \frac{dV_2}{dt} &= -\alpha_2 V_2 + \alpha_3 V_1 - \beta_1(V_1 + V_3) - \beta_2 V_1 V_3^2 \\ \frac{dV_3}{dt} &= \alpha_4 V_1 \end{aligned} \tag{3.1}$$

is the *best* model of the circuit. To do this, one would have to search a 5-dimensional parameter space according to some notion of best. If the *best* model is one that gives the longest shadowing times[1], finding the parameters might require weeks of computer time. Having such parameters for (3.1) might be useful to the design of hardware because of the potential to cast light on the intricacies that creep in when ideal component behaviour is employed when the underlying circuit is non-linear or

[1]Shadowing time is the amount that a model trajectory that is initially close to the observations remains close [64].

chaotic.

In this thesis, we shall not concern ourselves with finding optimum parameters to equation (3.1), although it is a useful problem which we shall revisit in a future study. Our main focus is to explore limits to predictability when the underlying models are fraught with errors, or even inadequate, which we believe is a feature shared by models used in various fields of the applied sciences, economics and finance. The modelling approach we adopt affords us high quality models that out-perform each other across the circuit attractor, a feature that we ultimately seek to exploit.

Before discussing the models of interest, we shall first introduce *embedings* in § 3.1. Quite often, scientists can measure only one variable at a time while the underlying system possibly lies in a higher dimensional space. In that case the scalar data has to be embedded into higher dimensional space. Models obtained from such a space will be called *delay models* and those from multi-dimensional data, *state space models*[2].

This chapter is organised as follows: In § 3.1, we discuss the theory of embeddings and then move on to local models in § 3.2.1. Local models are based on the local dynamics of the circuit. § 3.2.2 discusses the Kwasniok-Smith algorithm, which seeks to improve the learning set without increasing its size. § 3.2.3 discusses models that are based on the global dynamics, and they are called *global models* [64]. Throughout this report, the only *radial basis functions* we shall employ are the cubics, preferred for their simplicity and efficiency. In § 3.3 the conclusion is given together with a

[2]The assumption is that the measurement space contains the state space of the circuit.

table of the models that shall be used in the subsequent chapters.

3.1 Prediction Embeddings

Recall that we took three simultaneous voltage measurements from the circuit. These were V_i, $i = 1, 2, 3$, corresponding to the non-dimensional variables x, y, z respectively, in the MS system. A knowledge that the origin is an unstable point of the MS system was used to set the initial conditions, $(V_1, V_2, V_3) \approx (0, 0, 0)$. Each time the circuit dynamics would then evolve into some bounded region in state space. It is possible that the circuit lies in higher dimensional space than the observation space and observing it in a lower dimensional space may result in determinism being lost because of projection effects. For forecasting purposes, it is necessary to preserve determinism by at least making sure that there is no self intersection [3] in the system trajectory.

3.1.1 Embedding Theorems

Theoretically, we could use the Fractal Whitney's Embedding Prevalence Theorem [58] to decide whether the (V_1, V_2, V_3) space sufficiently embeds the underlying circuit attractor. An *embedding* is a one-to-one map that is an *immersion*. A smooth map $\boldsymbol{H}$ on an attractor A is an immersion if the Jacobian matrix $D\boldsymbol{H}(\boldsymbol{x})$ has full rank. In such a case, the differential structure of A is preserved in $\boldsymbol{H}(A)$ [48, 58]. Such an embedding is called a differentiable embedding [48]. The embedding theorem from Sauer [58] may be stated in verbatim as follows:

Theorem 2 *(Fractal Whitney Embedding Prevalence Theorem). Let A be a compact subset of $\mathbb{R}^k$ of box counting dimension d and let m be an integer greater than $2d$.*

[3] In our case, this will be true with probability one.

For almost every smooth map $\boldsymbol{H} : \mathbb{R}^k \to \mathbb{R}^m$,

1. $\boldsymbol{H}$ *is one-to-one on* A

2. $\boldsymbol{H}$ *is an immersion on each compact subset* $\mathcal{C}$ *of a smooth manifold contained in* A.

This theorem is a modification of the one originally given by Whitney [75]. Theorem 2 does not tell us how to find the map, $\boldsymbol{H}$. Takens' [70] theorem provides us with a way of finding the map. Before stating the theorem, let us first suppose that the flow of the system under study is $\boldsymbol{\varphi}_t$ on a manifold Ω and let $\tau > 0$ be some time delay and $h : \Omega \to \mathbb{R}$ is smooth function. The delay coordinate map, $\boldsymbol{H}(h, \boldsymbol{\varphi}_t, \tau) : \Omega \to \mathbb{R}^m$ is defined by:

$$\boldsymbol{H}(h, \boldsymbol{\varphi}_t, \tau)(\boldsymbol{x}) = (h(\boldsymbol{x}), h(\boldsymbol{\varphi}_{-\tau}(\boldsymbol{x})), \ldots, h(\boldsymbol{\varphi}_{-(m-1)\tau}(\boldsymbol{x}))). \tag{3.2}$$

We will call this the **delay vector** and the associated space, the **delay space**. The theorem is the given by [58]:

Theorem 3 *(Fractal Delay Embedding Prevalence Theorem). Let* $\boldsymbol{\varphi}_t$ *be a flow on an open subset* U *of* $\mathbb{R}^k$*, and let* A *be a compact subset of* U *of box-counting dimension* d*. Let* $m > 2d$ *be an integer, and let* $\tau > 0$*. Assume that* A *contains at most a finite number of equilibria, no periodic orbits of* $\boldsymbol{\varphi}_t$ *of period* τ *or* 2τ*, at most finitely many periodic orbits of period* $3\tau, 4\tau, \ldots, m\tau$*, and the linearisations of those periodic orbits have distinct eigenvalues. Then for almost every smooth function* h *on* U*, the delay coordinate map* $\boldsymbol{H}(h, \boldsymbol{\varphi}_t, \tau) : U \to \mathbb{R}^m$ *is:*

1. *One-to-one on* A.

2. *An immersion on each compact subset* C *of a smooth manifold contained in* A.

3.1.2 Time Delay

The preceding theorem does not tell us how to choose the time delay, τ. Fraser and Swinney [14] suggested employing a *mutual information* technique. In sequel, we shall explain how to use the mutual information to find the time delay, τ. We keep the problem simple and do not delve into the intricacies[4] addressed by Judd and Mees [28]. Although we shall here explain what mutual information is, further aspects of information theory may be found in § 5.3.

Let us first suppose that we have a scalar time series $\{s_n\}_{n=1}^N$ where

$$s_n = s(n\tau_s), \quad n = 1, \ldots, N, \tag{3.3}$$

and τ_s is the sampling time. In Fraser and Swinney [14], the argument put forth is that the best time delay for reconstruction and prediction is the minimum one for which the coordinates $s(t)$ and $s(t+\tau)$ exhibit minimum dependence. Suppose that $(x, y) = (s(t), s(t+\tau))$, so that the associated random variables are (X, Y) and H denotes *entropy* (or uncertainty). The question then posed is, "Given that s has been measured at time t, what is the average uncertainty in a measurement of s at time $t+\tau$? Kantz & Shreiber [30]" The answer is [14, 24, 57, 60, 74]

$$H(Y|X) = H(X, Y) - H(X), \tag{3.4}$$

where

$$H(X, Y) = -\sum_{ij} P_{xy}(x_i, y_j) \log P_{xy}(x_i, y_j), \tag{3.5}$$

$$H(X) = -\sum_{i} P_x(x_i) \log P_x(x_i) \tag{3.6}$$

[4]Since an attractor may exhibit widely varying frequencies across different regions, the choice of the time delay may have to be adaptive.

and P_{xy} is the joint probability function of X and Y, P_x is is the probability function of X and H is called entropy. The *mutual information* is then given

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{3.7}$$

Since, in our case, X and Y are delay coordinates of each other, I is a function of the time delay so that [30]

$$I(\tau) = \sum_{i,j} p_{ij}(\tau) \log p_{ij}(\tau) - 2 \sum_{i} p_i \log p_i. \tag{3.8}$$

The probabilities p_{ij} are obtained by partitioning the plane with squares of width ϵ. Although Fraser and Swinney [14] argued for a partition into equi-probable squares, we found that this did not change our results on the circuit. Also using more and more data has been found not to change the value of τ [73]. The time lag they suggested is the one that yields the first local minimum of the mutual information. In figure 3.1, a graph of the mutual information versus the time delay is given for the voltage signal V_3 of the circuit computed from Set7. Notice that first local minimum of the mutual information occurs at the time lag, $\tau = 6$ time steps. In the models of § 3.2.3, we found $\tau = 4$ to yield models that were more stable under iteration.

3.1.3 Box Counting Dimension

To find the box counting dimension d of an attractor A, suppose we partition the $\mathbb{R}^m$ space with a grid of cubes each of width ε and let $M(\varepsilon)$ be the number cubes that intersect A. Then

$$d = \lim_{\varepsilon \to 0^+} \frac{\log M(\varepsilon)}{-\log \varepsilon}. \tag{3.9}$$

It was Mandelbrot [42] who originally suggested the scaling law

$$M(\varepsilon) \sim \varepsilon^{-d}. \tag{3.10}$$

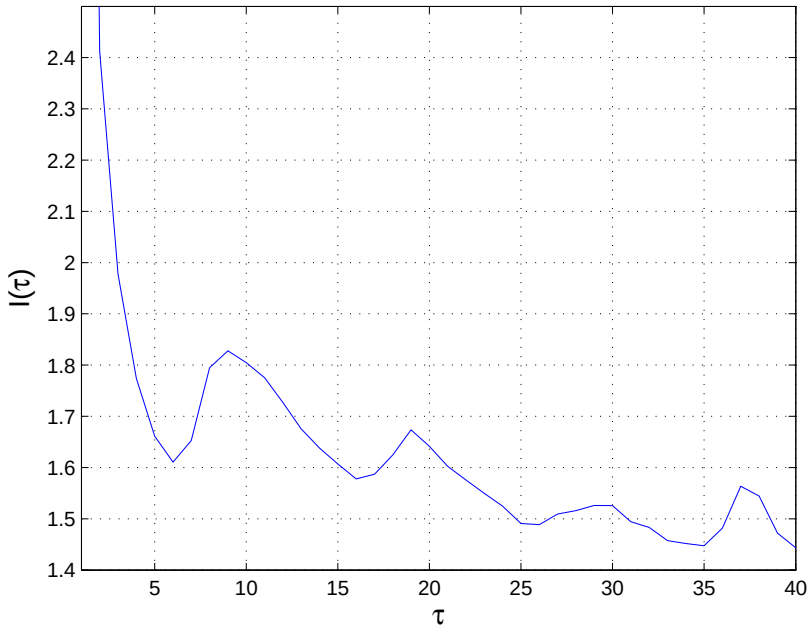


Figure 3.1: A graph of the mutual information, $I(\tau)$ versus the time lag, τ, for the voltage signal V_3 from Set7. The first local minimum occurs at $\tau = 6$ time steps, which is the best time delay to use in the delay vectors.

It has been noted that d can be intractable [20]. Grassberger and Procaccia (G-P) [20] later proposed an algorithm for computing an alternate quantity, ν, which is called the *correlation dimension*. It is based upon the *correlation integral*

$$C(\varepsilon) = \lim_{\varepsilon \to 0} \frac{1}{N^2} \sum_{i,j=1}^{N} \theta(\varepsilon - ||\boldsymbol{x}_i - \boldsymbol{x}_j||), \tag{3.11}$$

where $\theta(\cdot)$ is the Heaviside step function and $\{\boldsymbol{x}_i\}_{i=1}^{N}$ is some time series of observations. The correlation integral follows the scaling law [20]

$$C(\varepsilon) \sim \varepsilon^{\nu}. \tag{3.12}$$

Up to a factor of $O(1)$, it can be shown that [20]

$$C(\varepsilon) \geq \frac{1}{M(\varepsilon)} \sim \ \varepsilon^{d}, \tag{3.13}$$

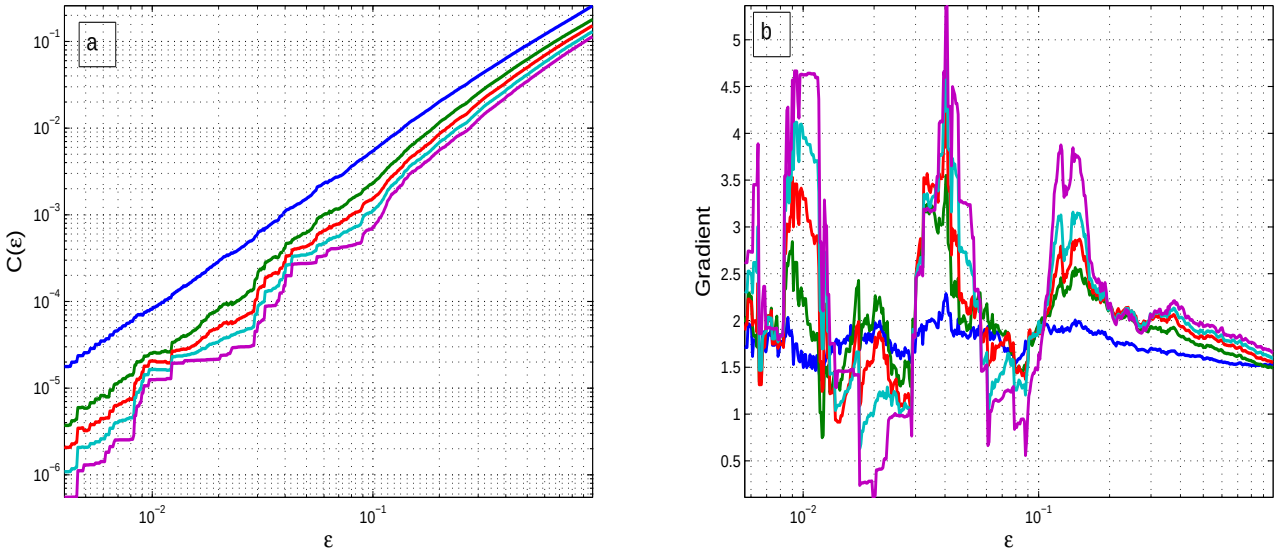


Figure 3.2: (a) Graphs of the correlation integral versus the box size. The different lines correspond to different dimensions used in the order $m = 2 - 6$ downwards. (b) Graph of gradient estimates with colours corresponding to dimensions on (a). None of the lines shows any convergence, but the m=2 line (blue) exhibits the minimum variation, assuming a minimum of 1.28 and maximum of 2.2. We used 5×10^5 data points from SET9 of the circuit.

From which it may be deduced that [20]

$$\nu \leq d. \tag{3.14}$$

The correlation dimension may be used as an estimate of the box counting dimension [20]. In sequel, we perform calculations of the correlation dimension.

Using 5×10^5 data points from Set9 of the circuit, we plotted the graphs of the Correlation integral versus ε and their corresponding gradients in figure 3.2. The gradients were obtained by sliding a logarithmically constant window of ε. Convergence in the gradient would be manifest by any of the lines in figure 3.2 (b) being constant somewhere as we slide the window and that value would be a good candidate for an estimate of d. The gradient line for $m = 2$ exhibits the least variation and the worst

is exhibited by the $m = 6$ line. For $m = 2$, the gradient assumes a minimum 1.28 and a maximum of 2.2. The lack of convergence means we have seek another choosing the embedding dimension. In this case we may choose an embedding dimension informed by the model, which is $m \geq 3$.

3.2 The Prediction Problem

From a realisation s_n at time $t = n\tau_s$, where τ_s is the sampling time, the prediction problem is to forecast $s_{n+\tau_p}$ from the m-dimensional delay vector, $\boldsymbol{x}_n = (s_{n-\tau(m-1)}, \ldots, s_n)$, where τ_p is the prediction time. If multivariate observations, $\boldsymbol{x}_n = (s_n^{(1)}, s_n^{(2)}, \ldots, s_n^{(m)})$, were made, then the prediction problem is to forecast $\boldsymbol{x}_{n+\tau_p}$.

In this sections, we shall explain how to construct prediction models from data. The data set from which we construct our models is called the *learning set* and the one on which we test the performance of our models shall be called the *testing set*. Casdagli suggested this kind of approach to chaotic time series in his seminal 1989 paper [11].

3.2.1 Local Models

One way to forecast a dynamical system from some current observation is to construct a sequence of models of the local dynamics. These could be polynomial interpolations, the simplest of which are *local linear models*. Operationally, what constitutes the local dynamics is constrained by the density of measurements in state space. This density may not be uniform as a function of position because the velocity of the system may vary with position, so that the density will be low in high velocity regions and vice versa. Therefore, the size of the neighbourhood may be chosen as a function of position. Whatever the degree of the polynomials that one opts for, the problem

is reduced to finding the best estimates of the coefficients, and the model is then iterated forward. After each iteration, a new local model is made.

We shall consider making a τ_p step ahead prediction with a local linear model, and the prediction from a given point of a time series is based on finding k nearest neighbours. Finding k nearest neighbours is equivalent to choosing the neighbourhood size as a function of position. Let us suppose that the set of measurements is $\{s_n\}_{n=1}^{N}$ with corresponding delay vectors $\{\boldsymbol{x}_n\}_{n\geq 1}$. For each observation s_n from which we want to make predictions, we denote the k nearest neighbours to the delay vector $\boldsymbol{x}_n$ by $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, \ldots, \boldsymbol{x}_{j_k}$, $j_r \in \{1, \ldots, N-(m-1)\tau\}$, where τ is the time delay. A linear model with prediction coefficients $\{a_l\}_{l=0}^{m}$ is fitted by minimising [11, 30]

$$\sum_{r=1}^{k} \left(s_{j_r+\tau_p} - a_0 - \sum_{l=1}^{m} a_l s_{j_r-\tau(l-1)} \right)^2 . \tag{3.15}$$

with respect to $\{a_l\}_{l=0}^{m}$. Minimising (3.15) is equivalent to solving the least squares problem

$$\min_{a\in\Re^{m+1}} ||Ca - b||, \tag{3.16}$$

where

$$C = \begin{bmatrix} 1 & \boldsymbol{x}_{j1} \\ 1 & \boldsymbol{x}_{j2} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_{jk} \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} s_{j1+\tau_p} \\ s_{j2+\tau_p} \\ \vdots \\ s_{jk+\tau_p} \end{bmatrix} . \tag{3.17}$$

This means $C \in \Re^{k\times(m+1)}$ and $b \in \Re^{k}$. The problem can be solved in Matlab using the "backslash" which uses the QR factorisation. If the problem is numerically unstable due to the matrix C being near rank deficient, then SVD is the best algorithm to use [19, 71] (See appendix C). With the coefficients determined from the local

dynamics, the prediction $\hat{s}_{n+\tau_p}$ for $s_{n+\tau_p}$ is then defined by

$$\hat{s}_{n+\tau_p} = a_0 + \sum_{l=1}^{m} a_l s_{n-\tau(l-1)}. \tag{3.18}$$

In solving equation (3.15), we determine $(m+1)$ coefficients. If we use a locally quadratic model, the number of parameters we then have to determine is $(m^2/2 + m + 1)$, and to iterate such a model becomes computationally expensive since the parameters have to be updated at each step. For a local approximation to be useful, one needs to sample the underlying attractor sufficiently. At least $(1 + (m-1)\tau)$ observations have to be made before making a prediction of the future evolution of the system under consideration, since our local models are based on a scalar time series.

How can we estimate the optimum number of nearest neighbours from a given data set, given the embedding dimension m and time delay τ? Let us consider the learning set comprising 10^4 points from Set 1 of the circuit. Predictions of trajectories whose verifications are within this set are called *in-sample* predictions and those for verifications outside the *learning set* are called *out-of-sample* predictions [30, 64]. To choose an optimum neighbourhood size, k, we randomly select a set of points, $\{s_{n_j}\}_{j=1}^{M} \subset \mathcal{L}$, where $\mathcal{L}$ is the learning set.For various values of k, predictions $\{\hat{s}_{n_j}^{(k)}\}_{j=1}^{M}$ are made using *cross validation* which is: For a given value of k, we make a prediction of s_{n_j}, $\hat{s}_{n_j}^{(k)}$, based on the learning set $\mathcal{L} \setminus \{s_{n_j}\}$. This procedure, is repeated for all s_{n_j}. The performance of the models is evaluated using the root mean square (RMS) error, $\varepsilon_{\mathrm{RMS}}(k)$, where

$$\varepsilon_{\mathrm{RMS}}(k) = \left[\frac{1}{M}\sum_{j=1}^{M}\left(\hat{s}_{n_j}^{(k)} - s_{n_j}\right)^2\right]^{1/2}. \tag{3.19}$$

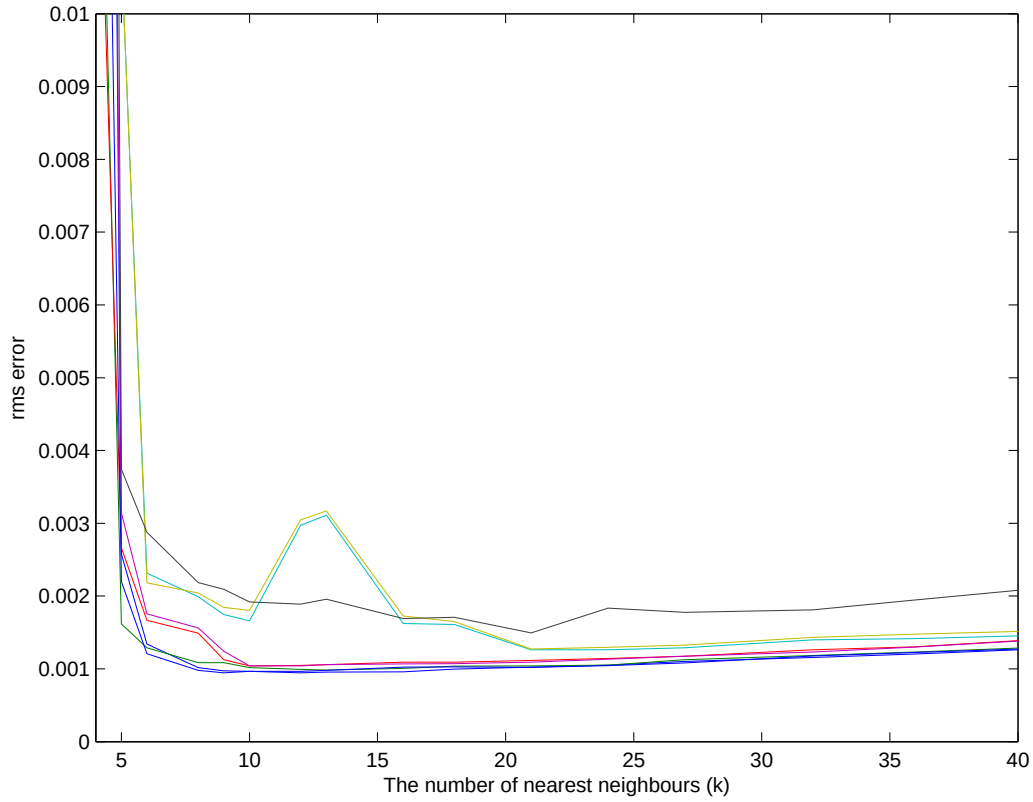


Figure 3.3: The graph of RMS error (of forecasts made using local linear models) versus the number of nearest neighbours for Set7. The different lines correspond to different samples of points from the learning set for which we make in-sample predictions and compute the RMS error. Notice that the general pattern of the lines is that they start off decreasing, reach a local minimum, and then rise up. The optimum number of nearest of neighbours is the mean or median of the global minima. We used $m = 3$, $\tau = 4$ time steps, 10^4 points in the learning set from the circuit, sampling (8 times) 200 points to perform cross validation, and found the mean to be 15.25 and median 16.

We then identify the value of $k = k_m$ at which the graph of ε_{RMS} versus k assumes the minimum [65]. The above procedure is repeated for different samples $\{s_{n_j}\}_{j=1}^{M}$ and we then take the mean or median of the k-values at which minima are attained. The nearest whole number to this is taken as an optimum number of nearest neighbours. Graphs of ε_{RMS} versus k for Set7 are shown on figure (3.3). They were obtained with $m = 3$, $\tau = 4$ time steps, $M = 200$, with 8 different choices of the set $\{s_{n_j}\}_{j=1}^{M}$ and we found $k \sim 16$.

3.2.2 Kwasniok-Smith Algorithm

Traditionally, the learning set is uniformly distributed with respect to the invariant measure of the underlying dynamical system. One way to improve model performance is to increase the size of this learning set. The main draw back of this approach is that increasing the size of the data set increases the computational time of the models, especially when we want to iterate the model to make multiple-step predictions. Memory problems place an ultimate limit on how far one can keep on increasing the size of the learning set. If the underlying system undergoes parametric drift with time, which in turn downgrades the performance of any models based on a fixed learning set, it may be better to update the learning set with time rather increase its size. Confronted with these issues, Kwasniok and Smith [35] developed an algorithm for refining the learning set to include points in the learning set from regions in the delay space with large prediction errors and remove points which are relatively redundant. We shall denote their algorithm, KSA, for Kwasniok-Smith algorithm.

To give the KSA, let us denote the learning set by $\mathcal{L} = \{(\boldsymbol{x}_n, s_{n+\tau_p})\}_{n=1}^{N}$, where $\boldsymbol{x}_n$ is a point in delay space and τ_p is the prediction time (or lead time). Initially, $\mathcal{L} = \mathcal{L}_0$, where $\mathcal{L}_0$ is the traditional learning set. Thus $\mathcal{L}_0$ is taken as the starting point of the learning set, keeping N and the number of nearest neighbour k fixed. The KSA algorithm is the following:

1. Read the next new point $(\boldsymbol{x}, y)$ from the data stream, where $\boldsymbol{x} = (s_{n'-(m-1)\tau,\ldots,s_{n'}})$ and $y = s_{n'+\tau_p}$.

2. Calculate the prediction $\hat{s}_{n'+\tau_p}$ based on the current refined learning set $\mathcal{L}$, and

compute the absolute prediction error, $\varepsilon' = |\hat{s}_{n'+\tau_p} - s_{n'+\tau_p}|$.

3. Draw a point at random from $\mathcal{L}$, each point being equally likely, and denote it by $(\boldsymbol{x}_{n^*}, s_{n^*+\tau_p})$.

4. Calculate the prediction $\hat{s}_{n^*+\tau_p}$ with the learning set $\mathcal{L}^* = \mathcal{L} \cup \{(\boldsymbol{x}, y)\} \setminus \{(\boldsymbol{x}_{n^*}, s_{n^*+\tau_p})\}$ and the corresponding error $\varepsilon^* = |\hat{s}_{n^*+\tau_p} - s_{n^*+\tau_p}|$.

5. If $\varepsilon^* < \varepsilon'$, then exchange $(\boldsymbol{x}_{n^*}, s_{n^*+\tau_p})$ for $(\boldsymbol{x}, y)$, which effectively means one takes $\mathcal{L}^*$ as the refined learning set; otherwise do not alter $\mathcal{L}$. Proceed to (1).

In this algorithm, ε' is a τ_p ahead out-of-sample error and ε^* is a τ_p ahead in-sample error. If we let $\rho(\varepsilon)$ be the probability density function of the out-of-sample prediction errors with the corresponding cumulative distribution function, $\Psi(\varepsilon)$, and $\rho^*(\varepsilon)$ and $\Psi^*(\varepsilon)$ the corresponding in-sample distributions. The exchange probability is then given by [35]

$$p = \int_0^\infty \rho(\varepsilon)\Psi^*(\varepsilon)d\varepsilon. \tag{3.20}$$

The following (new) theorem then holds:

Theorem 4 *(Exchange Probability Theorem) Consider two out-of-sample error cumulative distributions $\Psi_1(\varepsilon)$ and $\Psi_2(\varepsilon)$ with $\Psi_1(\varepsilon) \leq \Psi_2(\varepsilon)$ for all ε, and corresponding exchange probabilities, p_1 and p_2. If $\Psi^*(\varepsilon)$ is the cumulative distribution of in-sample errors with $\rho^*(\varepsilon) = \mathrm{d}\Psi^*(\varepsilon)/\mathrm{d}\varepsilon$, then*

$$p_1 \geq p_2,$$

where

$$p_i = \int_0^\infty \rho_i(\varepsilon)\Psi^*(\varepsilon)d\varepsilon$$

and $\rho_i(\varepsilon) = \Psi_i'(\varepsilon)$, $i = 1, 2$.

Proof: Firstly, the exchange probabilities may be written as[5]

$$\begin{aligned} p_i &= \int_0^\infty \rho_i(\varepsilon)\Psi^*(\varepsilon)d\varepsilon = \int_0^\infty \Psi_i'(\varepsilon)\Psi^*(\varepsilon)d\varepsilon \\ &= \Psi_i(\varepsilon)\Psi^*(\varepsilon)\Big|_0^\infty - \int_0^\infty \Psi_i(\varepsilon)\rho^*(\varepsilon)d\varepsilon \\ &= 1 - \int_0^\infty \Psi_i(\varepsilon)\rho^*(\varepsilon)d\varepsilon. \end{aligned}$$

The proof then proceeds as follows:

$$\begin{aligned} \Psi_1(\varepsilon) \leq \Psi_2(\varepsilon) &\Rightarrow \int_0^\infty \Psi_1(\varepsilon)\rho^*(\varepsilon)d\varepsilon \leq \int_0^\infty \Psi_2(\varepsilon)\rho^*(\varepsilon)d\varepsilon \\ &\Rightarrow 1 - \int_0^\infty \Psi_1(\varepsilon)\rho^*(\varepsilon)d\varepsilon \geq 1 - \int_0^\infty \Psi_2(\varepsilon)\rho^*(\varepsilon)d\varepsilon \\ &\Rightarrow p_1 \geq p_2. \end{aligned}$$

This theorem says that when the out-of-sample errors get bigger, the exchange probability increases, provided the in-sample error distribution remains the same. On the other hand, given two in-sample error distributions, $\Psi_i^*(\varepsilon)$, $i = 1, 2$ with $\Psi_1^*(\varepsilon) \geq \Psi_2^*(\varepsilon)$, it then follows that $\int_0^\infty \rho(\varepsilon)\Psi_1^*(\varepsilon)\mathrm{d}\varepsilon \geq \int_0^\infty \rho(\varepsilon)\Psi_2^*(\varepsilon)\mathrm{d}\varepsilon$. Therefore, smaller in-sample errors will yield bigger exchange probabilities, provided the out-of-sample errors remain unchanged. To sum up, a time series of the exchange probability would reflect relative changes in both the in-sample and out-of-sample error distributions.

The KSA was applied to data Set1 of the circuit. The initial learning set, $\mathcal{L}_0$ contained 10^4 data points of V_3 embedded into $3D$ delay space of V_3, with time delay $\tau = 5$. We then proceeded to process 50 seconds worth of data. At every stage of the processing, we computed the running mean of absolute errors, root mean square errors, and exchange probability, all of which are shown in figure 3.5. Projections of the initial ($\mathcal{L}_0$) and final ($\mathcal{L}_f$) learning sets onto 2D delay space are shown in figure 3.4.

[5]Applying integration by parts.

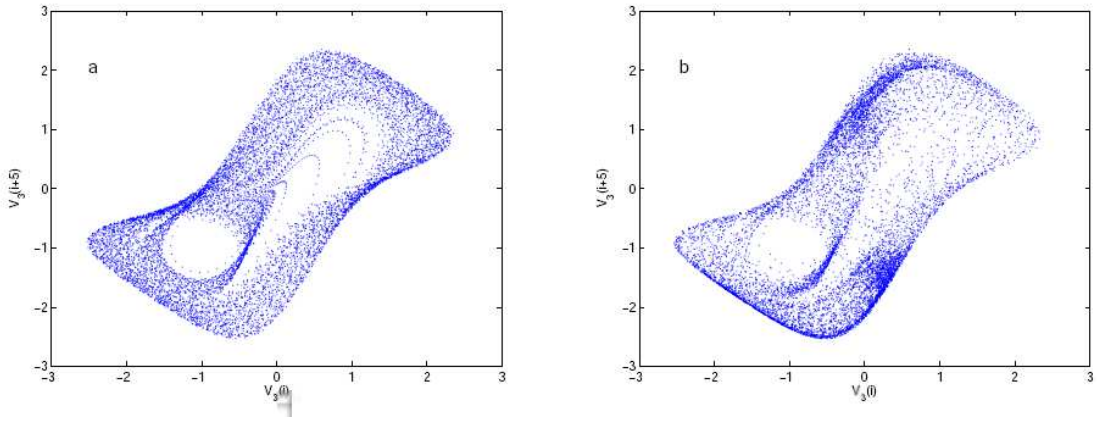


Figure 3.4: The (a) initial and (b) final learning sets of the circuit projected onto the 3D delay space of the processing 1.2×10^5 with KS algorithm. Notice that the refined learning set (in (b)) has a different distribution of points from the initial learning set.

$\mathcal{L}_f$ is what we would use for our forecasts with local models.

Notice in figure 3.5(a), that the out-of-sample mean absolute error is increasing with time, in contradistinction to the exchange probability, which tends to decrease with time (See figure 3.5 (e)). This suggests that the errors within the learning set are getting relatively bigger than the out-of-sample errors, even though the out-of-sample errors are increasing. It is, nevertheless, worrying that the running mean of absolute out-of-sample errors are growing. This may be due to noise in the time series because noisy observations can yield high prediction errors even when they lie in regions that are relatively predictable.

The next section looks at another way of building models from data. Unlike local models, the technique constructs models that can be used globally (In space and time) without the need to build a new model at every stage of the prediction process.

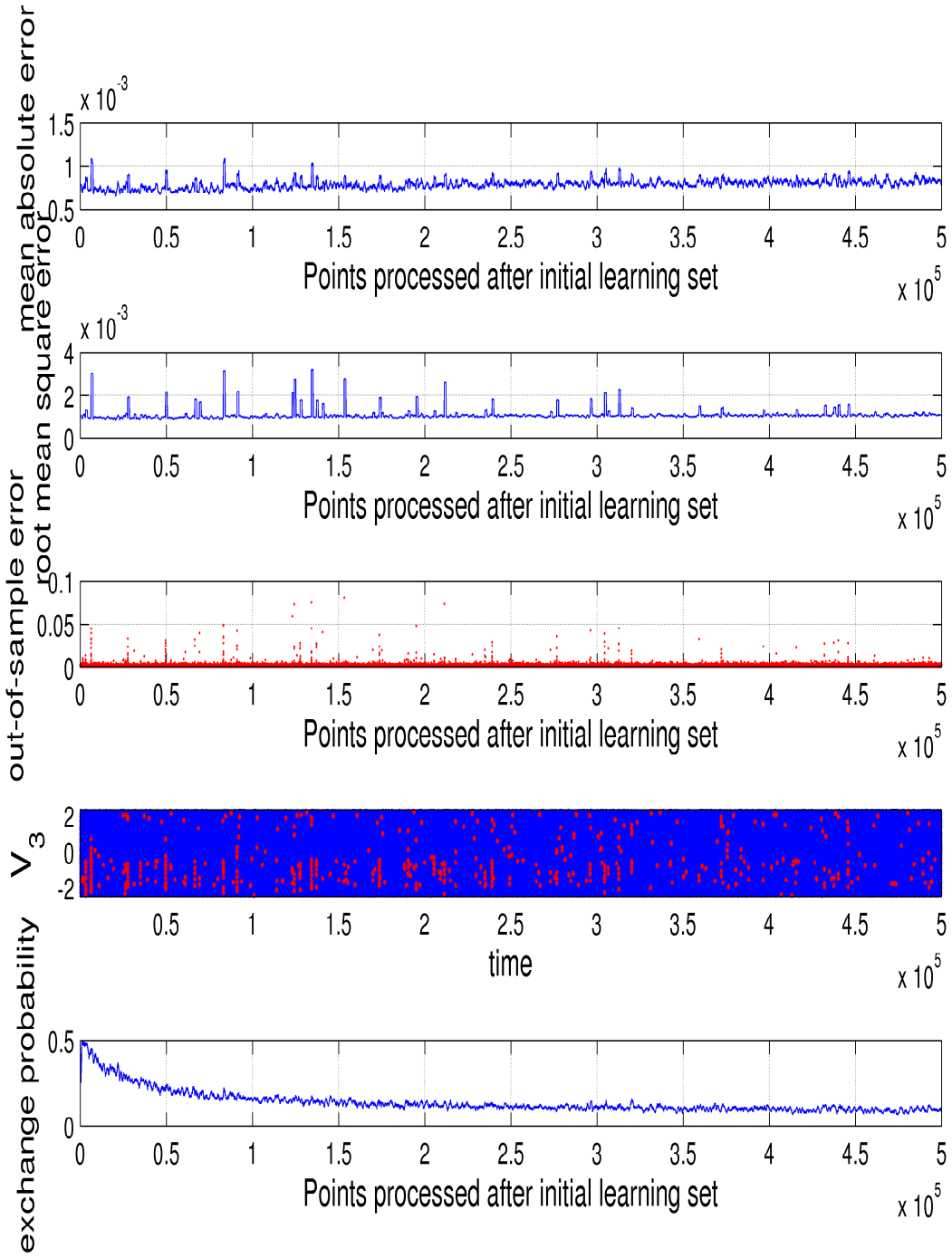


Figure 3.5: Each graph in (a), (b) and (e) shows the running mean over 1024 points when implementing the Kwasniok-Smith algorithm. (a) Shows the running mean of absolute out-of-sample errors. (b) Running root mean square error. (c) Absolute out-of-sample error. (d) Red dots are points in the time series where the out-of-sample error exceeds 0.006 and the blue area is the time series processed by KS algorithm. (e) An estimate of the exchange probability at each stage of KS algorithm.

3.2.3 Global Models

In contradistinction to *local models* global models are constructed once over the whole attractor. In fact the differential equation model given by equation (3.1) is a global model. In this section, global models constructed from *radial basis functions* are the only ones discussed. Such models are relatively inexpensive to run. They may be used either to capture the attractor of the underlying dynamics or merely to predict the future evolution of the state space. Constructing global models with an attempt to capture the attractor is harder and has been partly addressed in [33] using ellipsoidal basis functions and in more detail in [32]. The prediction problem is simpler and has been discussed in many papers such as [11, 27].

The modelling stance we shall adopt will be that of prediction, and we shall treat the issue as an interpolation problem [25, 27] to construct a model $\phi(\boldsymbol{x}) : \mathbb{R}^m \rightarrow \mathbb{R}$ which estimates s for any $\boldsymbol{x}$. Let us consider $F(\boldsymbol{x})$ of the form

$$\phi(\boldsymbol{x}) = \sum_{j=1}^{n_c} \lambda_j \psi(||\boldsymbol{x} - \boldsymbol{c}_j||) + L(\boldsymbol{x}), \tag{3.21}$$

where $\psi(r)$ are radial basis functions, λ_j are constants determined by observations in the learning set so that

$$\phi(\boldsymbol{x_i}) = s_{i+\tau_p}, \tag{3.22}$$

and $L(\boldsymbol{x})$ are linear terms. τ_p is the prediction time and $\boldsymbol{c}_j$ are the associated centres. We shall use cubic radial basis functions, which are of the form

$$\psi(r) = r^3.$$

Notice here that there are two levels in the problem. At one level, we want optimum parameters and another level we have to select the set of radial basis functions to enter

the model. Earlier papers [25, 46] treated the cases where the number of centres was equal to the number of data points. Solvability and convergence properties were then addressed. The problem with that is that the size of the matrix that arises can be intractable when the data set gets large. This also ignores *parsimony*[6], which may loosely be stated as the need to use as few parameters as necessary. Parsimony has been addressed in [27, 28] as a *minimum description length* problem (MDL). It is not our intention to wander along the well trodden path of RBF models at the expense of brevity. We shall rather briefly outline the procedure of model building given in [27] without concerning ourselves with the MDL criterion[7]. In this discussion, $\boldsymbol{x}$ need not be a point in delay space. If $\boldsymbol{x}$ is in measurement space, then we have to build m models of the form (3.22) for each coordinate (see § 4.2.3). The number of centres to use in the model shall be fixed beforehand. To determine the λ_j's we shall solve the least squares problem

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{n_c}} ||V\boldsymbol{\lambda} - \boldsymbol{b}||, \tag{3.23}$$

$b_i = s_{i+\tau_p}$ and V is a matrix whose columns are a subset of the columns of the matrix $\mathcal{A}$ with entries

$$\mathcal{A}_{ij} = \begin{cases} \psi(||\boldsymbol{x}_i - \boldsymbol{c}_j||) & \text{if} \quad j \leq n_d, \\ \boldsymbol{x}_{j-n_c} & \text{if} \quad n_d < j \leq (n_d + m), \\ 1 & \text{if} \quad j = (n_d + m + 1), \end{cases}$$

where n_d is the total number of centres from which we make our selection of n_c optimum centres. Our aim is to apply an algorithm that minimises the number of basis functions over an alphabet while keeping the prediction error to a minimum. This involves choosing a dictionary of basis functions and then selecting an optimum subset for the model. A practical algorithm for solving this problems is: For $k = 0, 1, \ldots,$

[6]Parsimony is desirable.

[7]Since we have lot of data, we need not worry about the MDL.

minimise $e = ||V\boldsymbol{\lambda} - \boldsymbol{b}||$ subject to $\mathcal{N}(\boldsymbol{\lambda}) = k$, where $\mathcal{N}(\boldsymbol{\lambda})$ is the number of nonzero components of $\boldsymbol{\lambda}$. Choosing a dictionary involves placing centres for the radial basis functions in phase space. The method of *chaperons* [27] places the centres near the data regions such that they are perturbations of some sample of the data points. The perturbations may be taken to have mean zero and standard deviation being 1/3 that of the learning set. Let $B = \{i_j\}_{j=1}^{k}$ be the indices for the columns of $\mathcal{A}$ that have been placed in the basis matrix, V. Below is the algorithm for building a global model [27, 28]:

1. Compute the standard deviation, σ_L, of the learning set.

2. Randomly choose centres

$$\boldsymbol{c}_j = \boldsymbol{x}_{l_j} + \boldsymbol{\eta}_j$$

   where $\boldsymbol{\eta}_j \in \boldsymbol{u}(a, b)$ and $\{\boldsymbol{x}_{l_j}\}_{j=1}^{n_d} \subset \{\boldsymbol{x_i}\}_{i=1}^{n_L}$.

3. For $k = 1$ to n_b

   (i) Add a column of A, A_j with $j \notin B$, that minimises $e = ||V\lambda - \boldsymbol{b}||$, into the matrix V and set $\text{in}(V) = A_j$. If $k = 1$, set $\text{out}(V) = \emptyset$, else set $\text{out}(V) = A_{i_j}$, the column of V whose removal least affects the prediction error.

   (ii) While $\text{in}(A) \neq \text{out}(V)$ and $k > 1$

   - Evaluate $\text{out}(V)$,
   - Evaluate $\text{in}(V)$; update V.

In the next section, we shall give the conclusion of the chapter and tabulate the models that will be used in the rest of the thesis.

Model	No. of RBF's	Dimension (m)	Coordinate space
M_1	40	3	Delay
M_2	40	4	Delay
M_3	25 per var	3	Multi-measurement
M_4	25 per var	3	Multi-measurement
M_5	40	3	Delay

Table 3.1: The table models of the circuit that were built using cubic radial basis functions (RBFs). In the second column, per var stands for per measurement variable. Models with the same coordinate space are different because they have different centres.

3.3 Conclusions

In this chapter, instead of estimating the best parameters in the model of the circuit given by equation (3.1) according to some criteria, we have acknowledged that model inadequacy could be the reason for model failure. Since there is no perfect model, we have discussed ways of constructing models from data. The data based models we opted for are local linear models and RBF models. Under local models, we also discussed the Kwasniok-Smith Algorithm, which improves the learning set without increasing its size.

We applied the KSA to the circuit, and features of the time series of the exchange probability and absolute prediction errors were explained.

A table of the RBF models of the circuit is given in table 3.1 and those for the MS system at parameter values $\gamma = 36$ and $\Gamma = 100$ are given in table 3.2. These models are used in the subsequent chapters.

Original work reported in this chapter consisted of:

- The exchange probability theorem, Theorem 4.

Model	No. of RBF's	Dimension (m)	Coordinate space
M_{e1}	40	3	Delay
M_{e2}	40	4	Delay
M_{e3}	25 per var	3	state space
M_{e4}	25 per var	3	state space
M_{e5}	40	3	Delay

Table 3.2: The table models of the MS system at parameter values $\gamma = 36$ and $\Gamma = 100$ that were built using cubic radial basis functions (RBFs). In the second column, per var stands for per state variable. Models with the same coordinate space are different because they have different centres.

- The local, RBF and KSA models constructed for the circuit.

- Time delay and neighbourhood estimates.

- Box counting dimension calculations of the circuit.

Chapter 4

Q-pling Time Distributions

In the previous chapter, we discussed ways of building models from data. These models will inevitably be imperfect. How can we assess the performance of the models across the circuit attractor? Since some chaotic systems are known [38, 67] to have some regions being more predictable than others, would the performance (across the system's attractor) of an imperfect model of good quality reflect this? In the light of statements such as, "The regional loss of predictability is an indication of the instability of the underlying flow, where small errors in the initial conditions (or imperfections in the model) grow to large amplitudes in finite times" (Kalnay et al. [29]), the answer to this question may indeed prove interesting. This statement ignores the fact that an imperfect model may exhibit regional losses in predictability even when the predictability of the underlying flow is reasonably uniform.

To investigate model behaviour across a system's attractor, we will use distributions of *q-pling* times [1] introduced by Smith et al [67] and one-step-error q-pling (OSEQ) times [2] introduced in this chapter. There are other measures of predictability, of which the most well known are the Lyapunov exponents [13]. These are average ex-

[1] q-pling time is the time taken for an initially small error to exceed q-times its original size.
[2] OSEQ-time is the time taken for the forecast error to exceed q times some pre-assigned size.

ponential growth rates, and do not tell us what happens across the different regions of a system's attractor. Lorenz [38]suggested using finite time Lyapunov exponents. However, Smith et al [67] have demonstrated their weakness against q-pling times. Since we are interested in how predictability varies across an attractor, we shall use the q-pling and OSEQ-times. Whereas q-pling times provide us with the predictability of underlying flow in the perfect model scenario, q-pling times with respect to an imperfect model on the underlying attractor do not. The q-pling times of an imperfect model on a systems attractor can, however, be used to explore variations in the behaviour of the model. In this chapter, we use q-pling times to demonstrate that we can obtain more diversity in our models by building them in various coordinate spaces. Model diversity may be exploited to improve the forecasts of the individual models. OSEQ-times can be used to assess variations in loss to predictability of the underlying flow by an imperfect model.

This chapter is organised as follows: § 4.1 discusses the concept of q-pling times in the perfect model scenario, which is then applied to the classical Lorenz and the MS system. In section 4.2, we turn to the imperfect model scenario and discuss ways of obtaining q-pling times for various cases. Q-pling times are then applied in the PMS and IMS to the MS system and the circuit. § 4.2.2 considers computing the q-pling times from delay space models. Measurement space models are discussed in § 4.2.3. The models of § 4.2 are not compared until § 4.4. OSEQ-times are introduced in § 4.5 and then applied to MS and circuit data. Whereas q-pling times assess variation of predictability at a micro scale, OSEQ-times assess it at macro scale. Prior to comparing q-pling time distributions for the various models, a similarity measure

that is useful for that purpose is introduced in § 4.3. In § 4.4, the implications of Takens theorem [70] for the predictability of dynamical systems are highlighted. We then go on to compare the models via q-pling and OSEQ-times and then look at the regions where q-pling occurs.

Smith et al [67] noted that the histogram of the q-pling times for the Lorenz and MS system are oscillatory. We unravel the reasons for this in § 4.6 for the Lyapunov q-pling times. Finally, we give the conclusions in § 4.7 and a list of original work in this chapter.

4.1 The Perfect Model Scenario

Following Smith et al. [67] first review the perfect model scenario, considering a dynamical system given by

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}(t)), \tag{4.1}$$

where $\boldsymbol{x}, \boldsymbol{F} \in \mathbb{R}^m$. The dynamics of an infinitesimally small uncertainty are governed by

$$\dot{\boldsymbol{\epsilon}} = J(\boldsymbol{x})\boldsymbol{\epsilon} \tag{4.2}$$

where $J(\boldsymbol{x})$ is the Jacobian of $\boldsymbol{F}$ at $\boldsymbol{x}$. Given an initial uncertainty, $\boldsymbol{\epsilon}_0$ at $\boldsymbol{x}_0$, a q-pling time $\tau_q(\boldsymbol{\epsilon}_0, x_0)$, is defined as the smallest time for which $||\boldsymbol{\epsilon}(t)|| > q||\boldsymbol{\epsilon}_0||$. The norm may be chosen at will. If $\boldsymbol{\varphi}_t(\boldsymbol{x}_0)$ is the solution to (4.1), then following Smith et al. [67], we can, in capsule form, write

$$\tau_q(\boldsymbol{\epsilon}_0, \boldsymbol{x}_0) = \inf_{t>0}\{t \Big| ||\boldsymbol{\varphi}_{\boldsymbol{t}}(\boldsymbol{x} + \boldsymbol{\epsilon}_0) - \boldsymbol{\varphi}_t(x)|| \geq q||\boldsymbol{\epsilon}_0||\}. \tag{4.3}$$

When $q = 2$, this gives doubling time.

Our aim is to compute the distribution of q-pling times as a function of initial condition on an attractor. To compute q-pling times, we sample points on the attractor that have minimum time separation of the order of the characteristic time of the system. The initial uncertainty directions we consider are those defined in § 1.3, which are the *Lyapunov*, *maximal* and *random*[3] directions. The optimal time over which the linear propagator used to compute the maximal direction was taken to be a quarter of the corresponding Lyapunov doubling time. At each of the points, say $\boldsymbol{x}(t_0)$, we choose

$$\boldsymbol{\epsilon}(t_0) = \frac{\boldsymbol{u}(t_0)}{||\boldsymbol{u}(t_0)||} \times \epsilon \tag{4.4}$$

where $\boldsymbol{u}$ is the uncertainty direction (not necessarily Lyapunov) and $\epsilon \in \mathbb{R}$ is to be chosen to be small relative to the size of the attractor. The ϵ should be small enough to ensure that a linear approximation of the dynamics is valid. The q-pling time is then obtained by integrating the augmented system, (4.1) and (4.2) and then evaluating the norm of $\boldsymbol{\epsilon}(t)$ until it just exceeds $q||\boldsymbol{\epsilon}_0||$.

As a preliminary example, we shall consider the Lorenz system [37, 68] In figure 4.1, we show a picture of the doubling times for the Lorenz attractor in the Lyapunov direction, maximum singular direction and random direction. To obtain these graphs, we integrated equations (4.1) and (1.25) with a time step of 0.01 to obtain a time series of linear propagators. The doubling times at 19900 initial conditions were then obtained by applying (1.23) to the corresponding initial errors in the Lyapunov directions, maximal direction and random directions. From comparing figures 4.1 (a), (b) and (c), it is evident that the right singular vector directions show less predictability

[3]For each point, we chose an independent random vector.

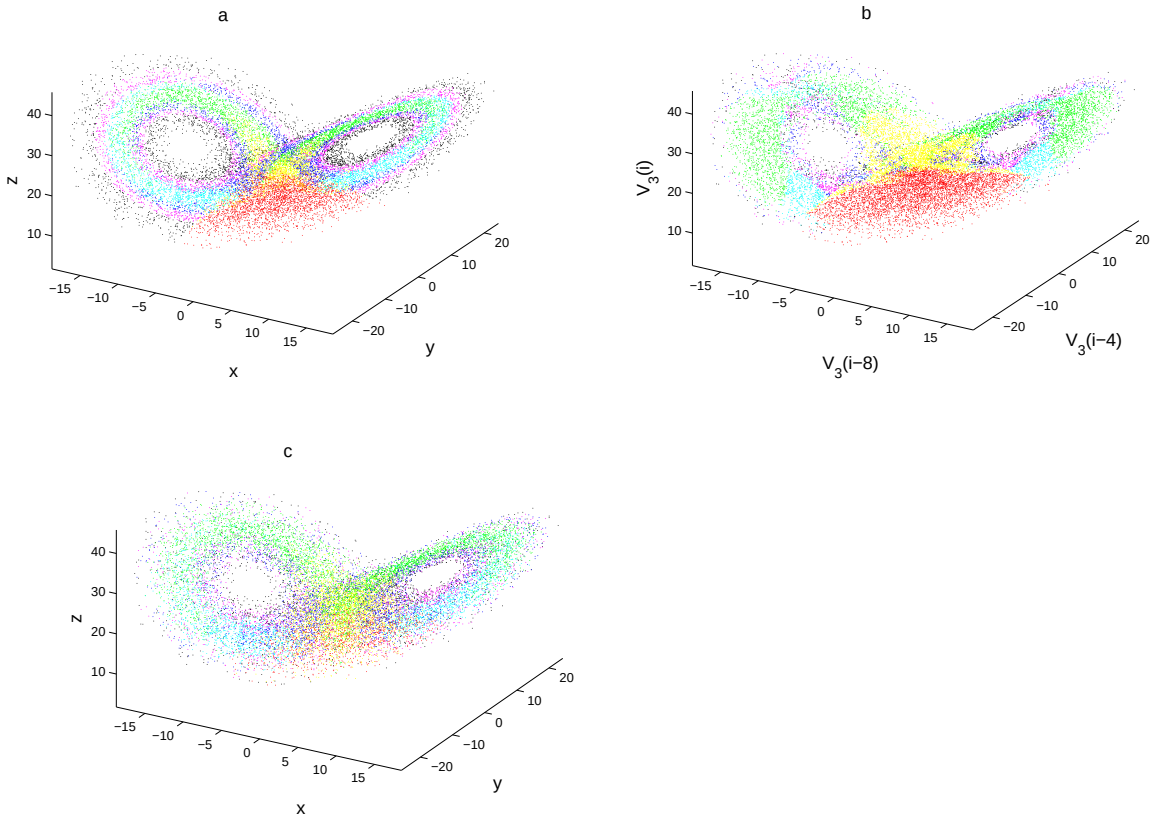


Figure 4.1: A view of the Lorenz attractor showing the doubling times of initial perturbations in the (a) Lyapunov direction, (b) maximal vector direction and (c) random direction. Red indicates $\tau_2 < 0.12$, yellow $0.15 \leq \tau_2 < 0.25$, green $0.25 \leq \tau_2 < 0.52$, cyan $0.52 \leq \tau_2 < 0.72$, blue $0.72 \leq \tau_2 < 1.22$, magenta $1.22 \leq \tau_2 < 2.02$, and black $\tau_2 \geq 2.02$. After Smith et al. [67]

than the Lyapunov directions and random directions. This is more[4] evident from the distributions of doubling times shown in figure 4.2. The Lyapunov direction yields predictability variation with a well organised tapestry. We shall see later that these features are also witnessed in the Moore-Spiegel system and circuit, regardless of whether we are looking at the perfect model scenario or not. Figure 4.1(c) shows that there is more variation in predictability with the initial conditions than there is with the initial orientation. As may be seen from figure 4.3, the region where uncertainty

[4]The cumulative distribution graph of the doubling times corresponding to the maximal direction is always above the others.

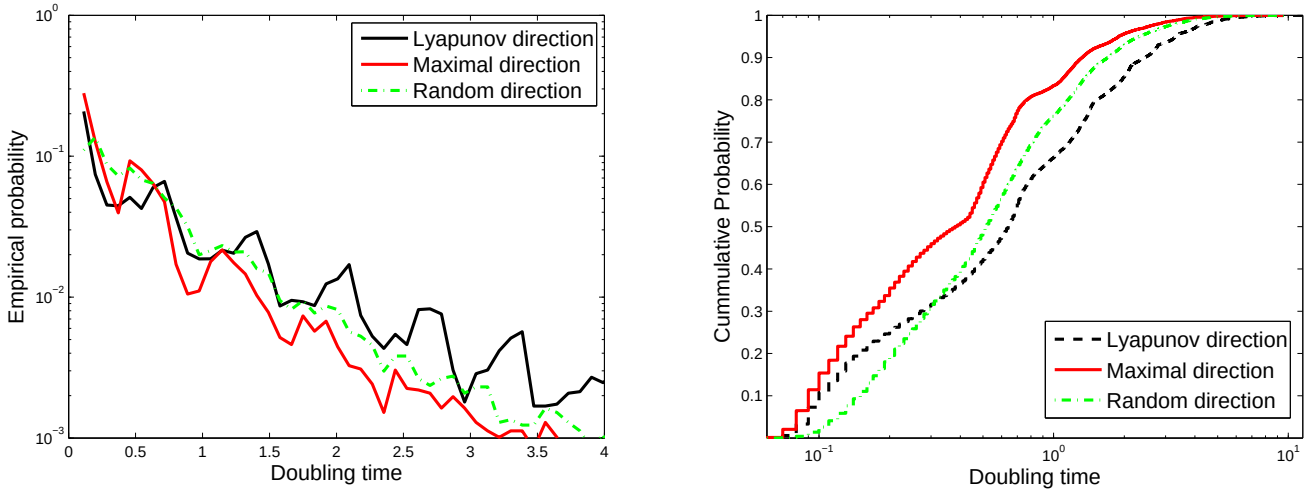


Figure 4.2: (left) Histograms of doubling times for the Lorenz attractor in the uncertainty directions of the diagrams shown in figure 4.1 after Smith [67] and (right) the corresponding cumulative distributions. On the vertical axis of left picture is the relative frequency of the bins used (or the proportion of sample points within a given bin).

doubling occurs does not reflect the natural measure of the underlying attractor. In fact, for all the uncertainty directions, doubling occurs in approximately the same region of the attractor as explained by Smith et al. [67].

We also considered the distributions of the doubling times of the MS system at parameter values $\gamma = 36$ and $\Gamma = 100$. The initial perturbations were made in the three uncertainty directions discussed above. These distributions in phase space are shown in figure 4.4. Again, we see in figure 4.4 (a) that the Lyapunov doubling times show more structure than the other directions, and their histogram (see § 4.6) is oscillatory. As with the Lorenz system, the regions where uncertainty doubling occurs (not shown) do not reflect the natural measure of the underlying attractor.

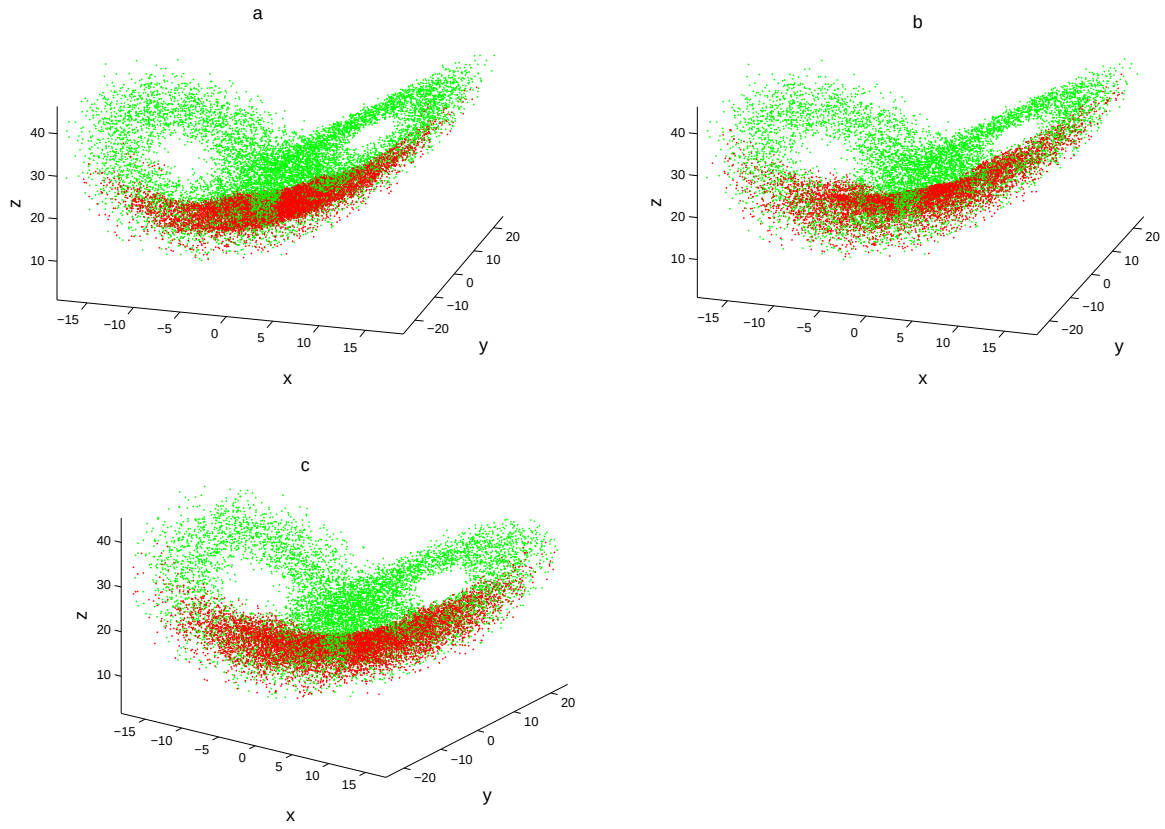


Figure 4.3: A view showing where the uncertainty doubling occurs in the three uncertainty directions described in figure 4.1. The starting points (green) reflect the natural measure on the attractor while the points where doubling occurs (red) do not.

4.2 The Imperfect Model Scenario

As remarked by Judd and Smith [26], the perfect model scenario is a fiction. In all practical situations we do not have a perfect model. Nevertheless, we could attempt computing measures of predictability from data without reference to any model at all as has been done by Eckmann and Ruelle [13]. The reliability of such measures needs to be investigated by an analysis of the fictitious perfect model scenario. In this section, we discuss the computation of q-pling times across a systems attractor with respect to an imperfect model. First, we will demonstrate that the distributions of q-pling times estimated from data based on local dynamics does not mimic what

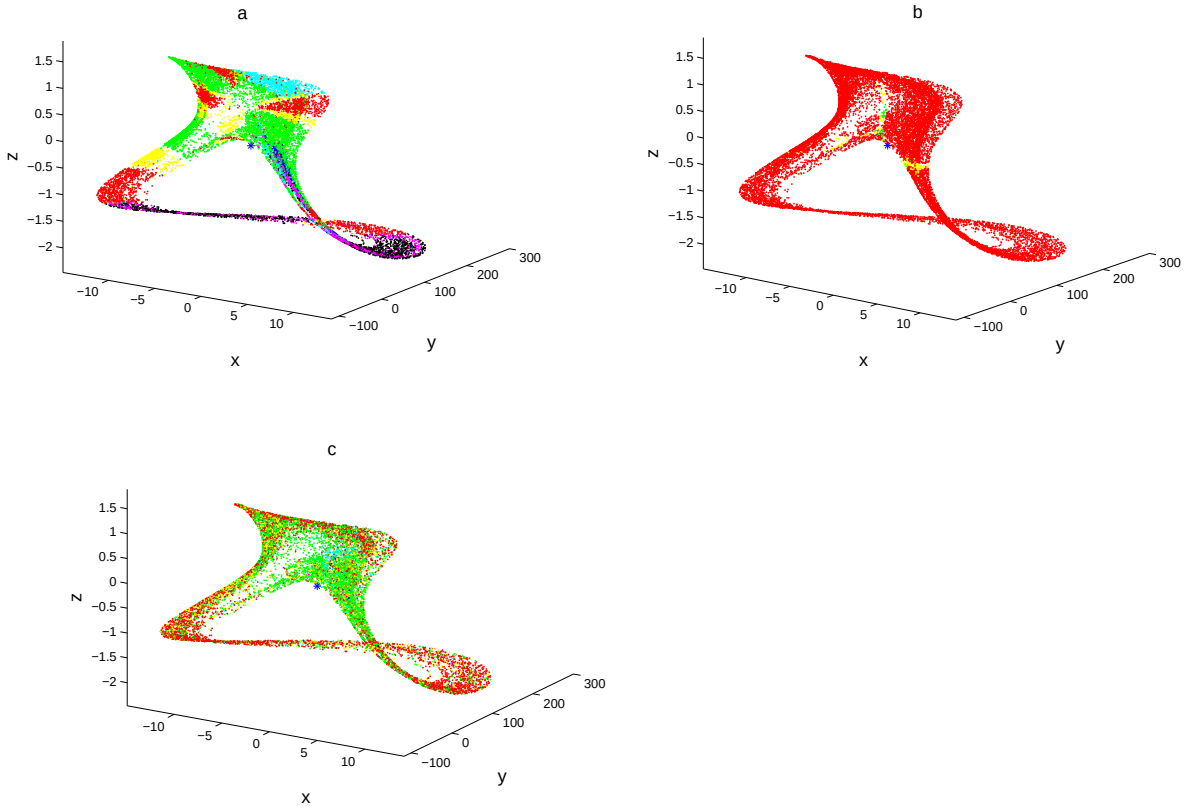


Figure 4.4: A view of the Moore-Spiegel attractor at $\gamma = 36$ showing the doubling times of initial perturbations in the (a) Lyapunov directions, (b) maximal direction and (c) random direction. Red indicates $\tau_2 < 0.05$, yellow $0.05 \leq \tau_2 < 0.1$, green $0.1 \leq \tau_2 < 0.5$, cyan $0.5 \leq \tau_2 < 0.8$, blue $0.8 \leq \tau_2 < 1$, magenta $1 \leq \tau_2 < 2$, and black $\tau_2 \geq 2$.

we would obtain using a perfect model. We then move on to discuss how to compute q-pling times from delay space models and measurement space models. The q-pling times are computed on the underlying system's attractor. The main part of the computations is finding the linear propagator.

4.2.1 Local Dynamics

Suppose we have some time series, $\{\boldsymbol{x}_i\}$ in m-dimensional space with sampling time, τ_s, and flow $\boldsymbol{\varphi}_\tau$. As in the perfect model scenario, we want to compute the matrices $T_i^\tau = D_{\boldsymbol{x}}\boldsymbol{\varphi}_\tau$ with $\tau = p\tau_s$. τ should be chosen so that the eigenvalues of T_i^τ are not

too small since we are going to multiply matrices proportional to τ^{-1}. We want to estimate the local dynamics from the time series to obtain the tangent matrices.

The derivatives, T_i^τ are obtained by a best linear fit of the map which, for $||\boldsymbol{x}_j - \boldsymbol{x}_i|| < \tilde{r}$, maps $\boldsymbol{x}_j - \boldsymbol{x}_i$ to

$$\boldsymbol{\varphi}_\tau(\boldsymbol{x}_i) - \boldsymbol{\varphi}_\tau(\boldsymbol{x}_j) = \boldsymbol{x}_{i+p} - \boldsymbol{x}_{j+p}$$

with $||\hat{\boldsymbol{x}}_{j+p} - \hat{\boldsymbol{x}}_{i+p}|| < \tilde{r}$. We may also require that $||\boldsymbol{x}_{j+k} - \boldsymbol{x}_{i+k}|| < \tilde{r}$ for all $1 \leq k < p$. Although m points are enough to determine the linear map, we shall seek many points $\hat{\boldsymbol{x}}_j$ in the neighbourhood of $\boldsymbol{x}_i$. We then use them to solve the least squares linear fit

$$T_i^\tau[\boldsymbol{x}_j - \boldsymbol{x}_i] \approx \boldsymbol{x}_{j+p} - \boldsymbol{x}_{i+p} \quad \Leftrightarrow \quad T_i^\tau \boldsymbol{\epsilon}_i^{(l)} \approx \boldsymbol{\epsilon}_{i+p}^{(l)}. \tag{4.5}$$

for $l = 1, \ldots, n_i$ and n_i is the number of points in the neighbourhood of $\boldsymbol{x}_i$. To solve (4.5), it is convenient to break it up into m matrix equations. Let $A_i = [\boldsymbol{\epsilon}_i^{(1)}, \boldsymbol{\epsilon}_i^{(2)}, \ldots, \boldsymbol{\epsilon}_i^{(n_i)}]^T$ and $\boldsymbol{b}_i^{(j)} = [\boldsymbol{\epsilon}_{i+p,j}^{(1)}, \boldsymbol{\epsilon}_{i+p,j}^{(2)}, \ldots, \boldsymbol{\epsilon}_{i+p,j}^{(n_i)}]^T$, where $\boldsymbol{\epsilon}_{i+p,j}^{(k)}$ is the jth entry of $\boldsymbol{\epsilon}_{i+p}^{(k)}$. If $T_i^{(j)}$ is the jth row of T_i^τ, then we can write (4.5) as

$$A_i T_i^{(j)} = \boldsymbol{b}_i^{(j)}, \quad j = 1, \ldots, m. \tag{4.6}$$

The m equations given by (4.6) can then be solved by least squares in a straight forward way. Better still, we can solve the single least squares equation

$$A_i X_i = B_i, \tag{4.7}$$

where $X_i = [T_i^{(\tau)}]^T$ and $B_i = [\boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}, \ldots, \boldsymbol{b}_i^{(m)}]$. It has been conjectured [13] that we should be able to obtain the unstable directions of T_i^τ with more confidence than the other directions.

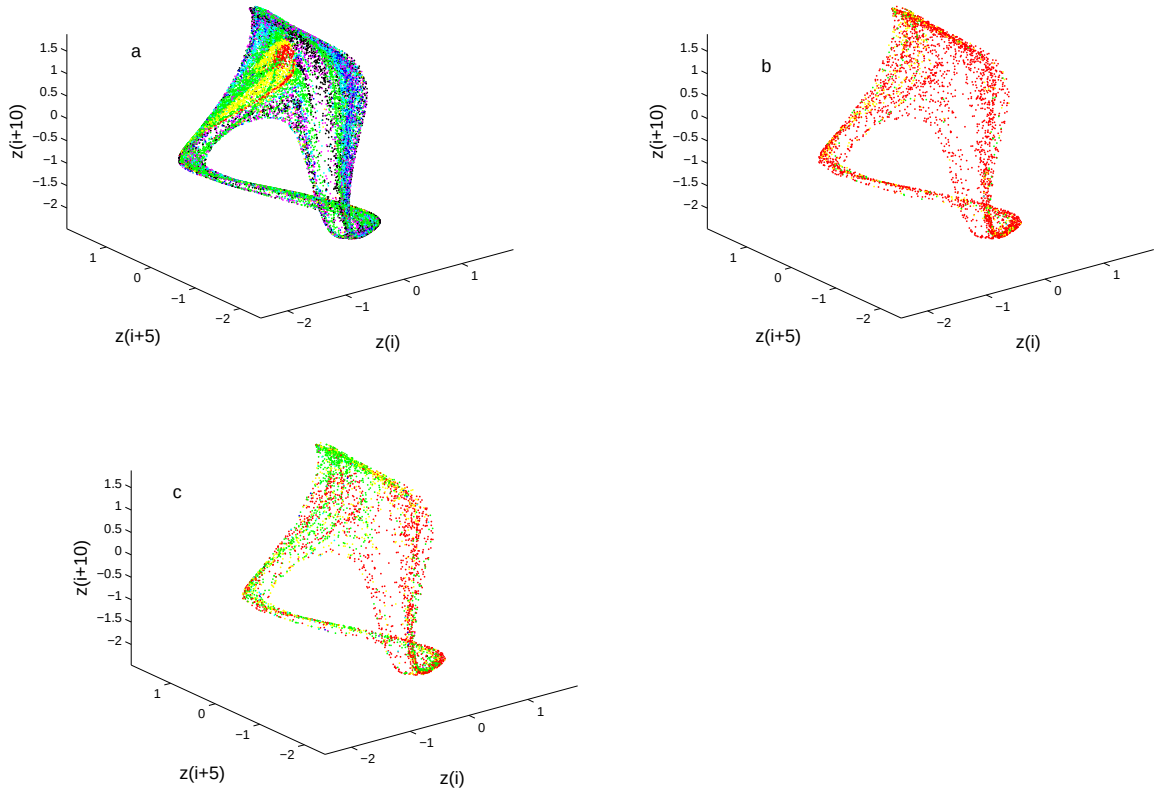


Figure 4.5: A view of the Moore-Spiegel attractor in embedding space at $\gamma = 36$ showing the doubling times for 10000 initial conditions for (a) the Lyapunov direction, (b) maximal direction and (c) random direction. Red indicates $\tau_2 < 0.05$, yellow $0.05 \leq \tau_2 < 0.1$, green $0.1 \leq \tau_2 < 0.5$, cyan $0.5 \leq \tau_2 < 0.8$, blue $0.8 \leq \tau_2 < 1$, magenta $1 \leq \tau_2 < 2$, and black $\tau_2 \geq 2$. The embedding dimension was $m = 5$ and the length of trajectories over which the linear propagator was computed with $p = 4$ iterations.

For the moment let us consider 10^6 data points (In the delay space of the z variable) of the Moore-Spiegel system at $\gamma = 36$ with integration step of 0.01, $p = 4$ and m=5. In figure 4.5, we have the distributions of doubling times in the Lyapunov vector direction, random vector direction and maximal singular vector direction, based on the foregoing method of approximating the local dynamics from time series. This differs significantly from the case of the perfect model scenario, where the variation of predictability in the Lyapunov direction is shown in figure 4.4. We have shown the two cases in figure 4.6, both in delay space for ease of comparison.

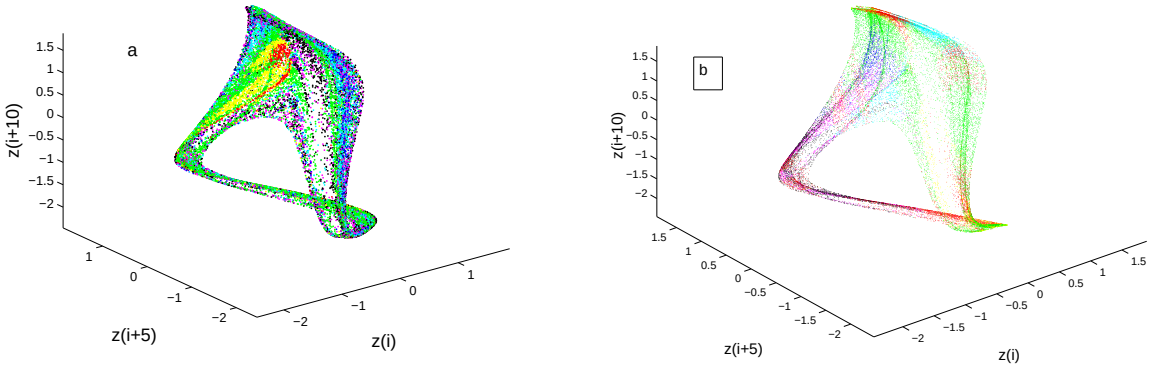


Figure 4.6: MS attractor doubling time distributions based on (a) local dynamics approximations and (b) the perfect model with time partitions as in figure 4.5.

4.2.2 Delay Space Models

In this subsection, we examine how to compute the variation of predictability based on an imperfect model built in delay space. Suppose our model is

$$x_{n+1} = \phi(\boldsymbol{x}_n), \tag{4.8}$$

where $\boldsymbol{x}_n$ is point on an m-dimensional delay space and x_{n+1} is a scalar. i.e.

$$\boldsymbol{x}_n = \begin{pmatrix} x_n \\ x_{n-k} \\ \vdots \\ x_{n-(m-1)k} \end{pmatrix},$$

where k is the time delay. Suppose $\boldsymbol{x}'_n$ is a perturbation of $\boldsymbol{x}_n$ such that $x'_{n+1} = \phi(\boldsymbol{x}'_n)$, and define $\varepsilon_{n+1} = x'_{n+1} - x_{n+1}$, $\boldsymbol{\varepsilon}_n = \boldsymbol{x}'_n - \boldsymbol{x}_n$. Then

$$\begin{aligned} \varepsilon_{n+1} &= \phi(\boldsymbol{x}'_n) - \phi(\boldsymbol{x}_n) \\ &\approx \nabla\phi(\boldsymbol{x}_n)\boldsymbol{\varepsilon}_n. \end{aligned} \tag{4.9}$$

If ϕ is a linear combination of radial basis functions[5] (RBF's), $\psi(\cdot)$, and some linear terms, i.e.

$$\phi(\boldsymbol{x}_n) = \sum_{i=1}^{n_c} \lambda_i \psi(||\boldsymbol{x}_n - \boldsymbol{c}_i||) + \boldsymbol{a} \cdot \boldsymbol{x}_n + a_0, \tag{4.10}$$

where $\boldsymbol{c}_i$ are centres and λ_i's are real coefficients, then

$$\nabla\phi(\boldsymbol{x}_n) = \sum_{i=1}^{n_c} \lambda_i \nabla\psi(||\boldsymbol{x}_n - \boldsymbol{c}_i||) + \boldsymbol{a}. \tag{4.11}$$

If we let $\mathbf{r}_i = \boldsymbol{x}_n - \boldsymbol{c}_i$ and $\psi(||\boldsymbol{x}_n - \boldsymbol{c}_i||) = \psi_i$, then equation (4.11) may be rewritten as

$$\nabla\phi(\boldsymbol{x}_n) = \sum_{i=1}^{n_c} \lambda_i \frac{\mathbf{r}_i}{r_i} \frac{\partial\psi_i}{\partial r_i} + \boldsymbol{a}. \tag{4.12}$$

If our model is built using cubic radial basis functions[6] and linear terms, then

$$\begin{aligned} \nabla\phi(\boldsymbol{x}_n) &= 3\sum_{i=1}^{n_c} \lambda_i r_i \mathbf{r}_i + \boldsymbol{a} \\ &= \left(a_m + 3\sum_{i=1}^{n_c} \lambda_i r_i (x_n - c_m^{(i)}), \ldots, a_1 + 3\sum_{i=1}^{n_c} \lambda_i r_i (x_{n-(m-1)} - c_1^{(i)}) \right) \end{aligned} \tag{4.13}$$

where a_i is the coefficient of x_{n-i+1} in the model.

Notice here that $\nabla\phi(\boldsymbol{x}_n)$ is a vector and not a matrix. Hence, we do not really have Jacobian matrices to use for the computation of the Lyapunov direction and singular vector direction. So, what can we use to determine the Lyapunov directions? To resolve this we consider the map

$$\boldsymbol{\Phi}(\boldsymbol{X}_n) = \begin{pmatrix} \phi(\boldsymbol{X}_n) \\ \phi_1(\boldsymbol{X}_n) \\ \vdots \\ \phi_{(m-1)k}(\boldsymbol{X}_n) \end{pmatrix}, \tag{4.14}$$

[5]The coefficients are fitted as discussed in § 3.2.3.
[6]That is, $\psi(r) = r^3$.

where $\phi_i(\boldsymbol{X}_n) = \boldsymbol{X}_n^{(i)}$ and

$$\boldsymbol{X}_n = \begin{pmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_{n-(m-1)k} \end{pmatrix}, \tag{4.15}$$

and $\boldsymbol{X}_n^{(i)}$ denotes the ith coordinate. This gives us the map

$$\boldsymbol{X}_{n+1} = \boldsymbol{\Phi}(\boldsymbol{X}_n), \tag{4.16}$$

When $k = 1$, the Jacobian is then given by

$$J(\boldsymbol{x}_n) = \begin{pmatrix} & \nabla f(\boldsymbol{x}_n) & \\ 1 & 0 & \cdots \\ 0 & \ddots & \ddots \end{pmatrix}, \tag{4.17}$$

with the characteristic equation [7]

$$\Lambda^m + \sum_{i=1}^{m} (-1)^{m-i+1} \left[a_i + \sum_{j=1}^{n_c} \lambda_j r_j \left[y_{n-(m-i)} - c_i^{(j)} \right] \right] \Lambda^{i-1} = 0, \tag{4.18}$$

where Λ is the eigenvalue of the Jacobian in (4.17). More generally, the Jacobian is then given by

$$D\boldsymbol{\Phi}(\boldsymbol{X}_n) = \begin{pmatrix} \frac{\partial \phi}{\partial x_n} & \frac{\partial \phi}{\partial x_{n-1}} & \cdots & \frac{\partial \phi}{\partial x_{n-(m-1)k}} \\ 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \end{pmatrix}. \tag{4.19}$$

It is useful to note that $\frac{\partial \phi}{\partial x_i} = 0$ for all i except $i = n - jk$, where $j = 0, \ldots, m-1$. The eigenvalues of this Jacobian are then given by solutions to the characteristic equation

$$(-\Lambda)^{(m-1)k+1} + \sum_{j=0}^{m-1} (-1)^{jk} (-\Lambda)^{[m-(j+1)]k} \frac{\partial \phi}{\partial x_{n-jk}} = 0. \tag{4.20}$$

As an example, let us consider doubling times of model M_1 on the circuit, which was built using cubic RBF's. Distributions of the doubling times with initial perturbations in the Lyapunov, maximal and random directions for this model on the circuit are shown in figure 4.7. Unlike those obtained by approximating Jacobians from the local dynamics according to § 4.2.1(shown in figure 4.8), these exhibit some nice organised

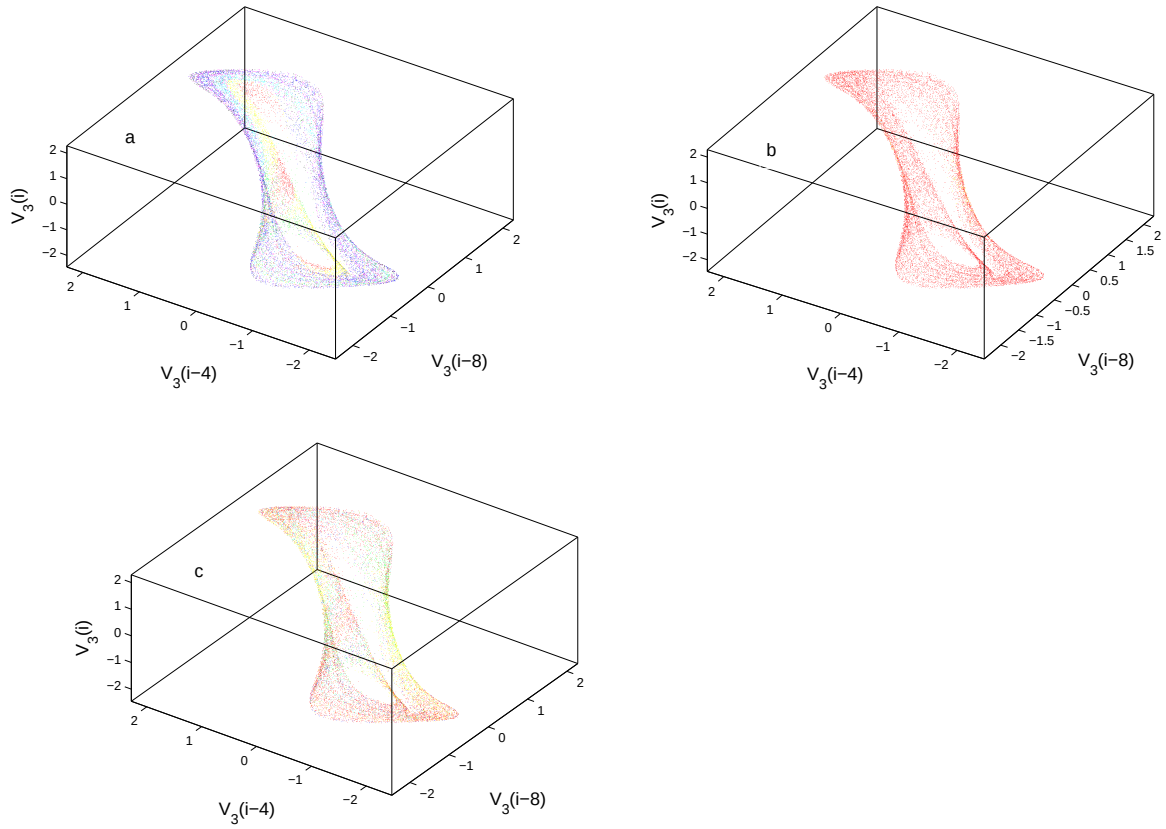


Figure 4.7: A view of the Moore-Spiegel circuit in delay space showing distributions of doubling times for a delay space model with initial errors in (a) Lyapunov, (b) maximal, and (c) random directions using data Set7. Red indicates $\tau_2 < 0.4$, yellow indicates $0.4 < \tau_2 < 0.9$, green indicates $0.9 < \tau_2 < 1.3$, cyan indicates $1.3 < \tau_2 < 3$, blue indicates $3 < \tau_2 < 8$, magenta indicates $8 < \tau_2 < 20$ and black indicates $\tau_2 > 20$.

structure typical in the perfect model scenario. That is, RBF in delay space fare better.

4.2.3 State space model

Alternatively, we may wish to model the system in state space, in which case our model becomes

$$\boldsymbol{x}_{n+1} = \boldsymbol{\phi}(\boldsymbol{x}_n), \tag{4.21}$$

[7]Provided we used cubic RBF's with linear terms to build our model.

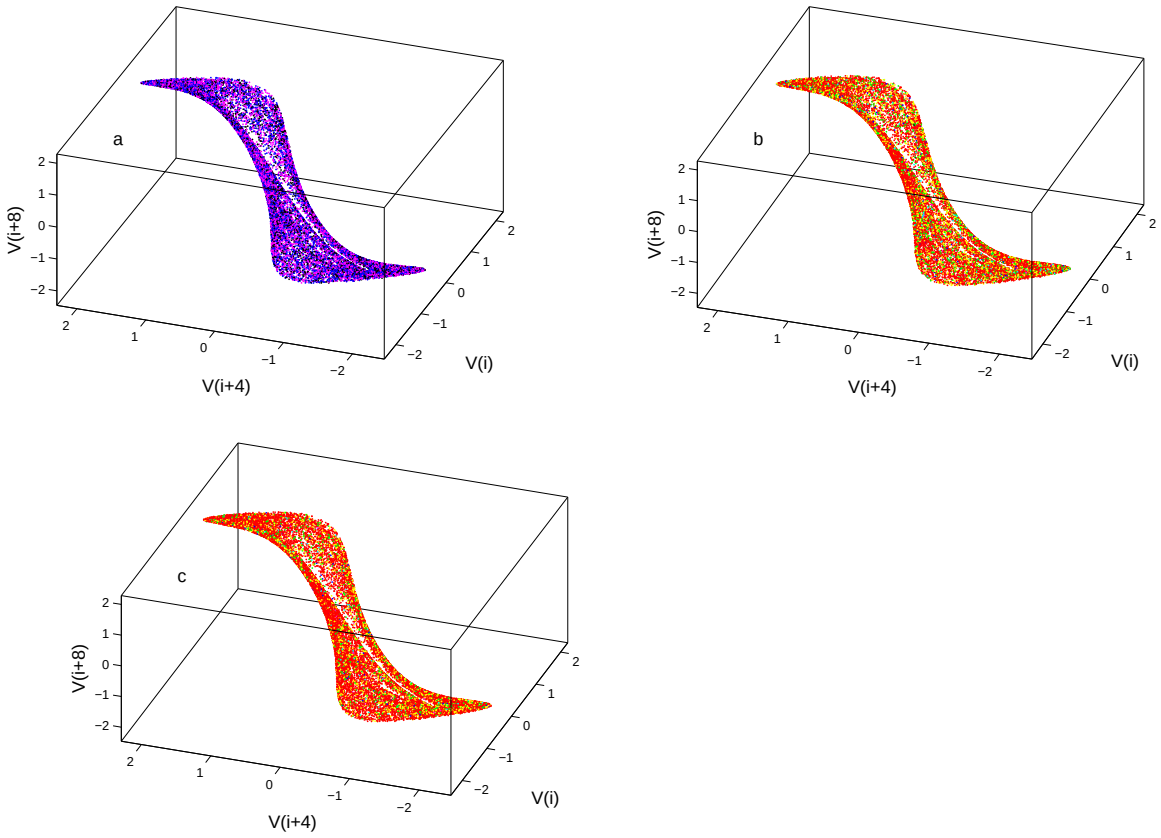


Figure 4.8: A view of the Moore-Spiegel circuit in delay space showing the doubling times of initial perturbations in the (a) Lyapunov direction (b), singular vector direction (c) and random direction with the Jacobian approximated from the local dynamics according to section 4.2.1 using data Set7. Red indicates $\tau_2 < 0.15$, yellow $0.15 \leq \tau_2 < 0.2$, green $0.2 \leq \tau_2 < 0.4$, cyan $0.4 \leq \tau_2 < 0.6$, blue $0.6 \leq \tau_2 < 0.9$, magenta $0.9 \leq \tau_2 < 2$, and black $\tau_2 \geq 2$. The length of trajectories over which the linear propagator was computed was $p = 4$.

where $\boldsymbol{\phi}$ is vector which is a linear combination of radial basis functions. That is, $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_m)^T$ and

$$\phi_i(\boldsymbol{x}_n) = \sum_{j=1}^{n_c} \lambda_{ij} \psi(|\boldsymbol{x}_n - \boldsymbol{c}_{ij}|) + \boldsymbol{a}_i \cdot \boldsymbol{x}_n + a_{i0}. \tag{4.22}$$

The centres $\boldsymbol{c}_{ij}$ may be written as a 3-dimensional array, $\boldsymbol{C}$, with entries, $\boldsymbol{C}_{ijk} = \boldsymbol{c}_{ik}^{(j)}$, where $\boldsymbol{c}_{ik}^{(j)}$ is the jth entry of $\boldsymbol{c}_{ik}$. We can also define the matrix $\boldsymbol{\lambda}$ with entries $\boldsymbol{\lambda} = (\lambda_{ij})$. The dynamics of initially small perturbations are governed by

$$\boldsymbol{\varepsilon}_{n+1} \approx D\boldsymbol{\phi}(\boldsymbol{x}_n)\boldsymbol{\varepsilon}_n, \tag{4.23}$$

where,

$$D\boldsymbol{\phi} = \begin{bmatrix} \nabla\phi_1 \\ \nabla\phi_2 \\ \vdots \\ \nabla\phi_m \end{bmatrix},$$

$$\nabla\phi_i(\boldsymbol{x}_n) = \sum_{j=1}^{n_c} \lambda_{ij} \frac{\mathbf{r}_{ij}}{r_{ij}} \frac{\partial\psi_{ij}}{\partial r_{ij}}, \tag{4.24}$$

$\mathbf{r}_{ij} = \boldsymbol{x}_n - \boldsymbol{c}_{ij}$, and $\psi_{ij} = \psi(|\mathbf{r}_{ij}|)$. If we use cubic RBF's, then equation (4.24) becomes

$$\nabla\phi_i(\boldsymbol{x}_n) = 3\sum_{j=1}^{n_c} \lambda_{ij} r_{ij} \mathbf{r}_{ij} + \boldsymbol{a}_i. \tag{4.25}$$

In § 4.4.1, we shall compare the q-pling time distributions for various models of the circuit with reference to Takens theorem. Before then, we will introduce a measure of similarity between q-pling time distributions and the briefly discuss Takens theorem in § 4.4.

4.3 Similarity Measure

In the previous sections, we considered q-pling times under various models of the MS system and circuit. The crucial question that remained unanswered was how to compare the distributions of q-pling times across an attractor. Here, we address this question.

Consider two models, M_1 and M_2 with q-pling times $\tau_q^{M_1}$ and $\tau_q^{M_2}$ respectively[8]. Let $\tau_q^{M_i} = \tau_q^{M_i}(t)$ be the q-pling time of model M_i at a point realised on the attractor at time t. Suppose the cumulative distribution function for the q-pling times of model M_i is $F_i(\tau_q^{M_i})$. If we partition each F_i with points $\{p_j\}_{j=0}^n$ with $p_0 = 0$ and $p_n = 1$, we

[8]The same notation is employed for OSEQ times.

can then define the jth q-pling time similarity set of the two sets of q-pling times by

$$\Gamma_s^j = \left\{ \left(\tau_q^{M_1}(t), \tau_q^{M_2}(t) \right) : F_1(\tau_q^{M_1}(t)), F_2(\tau_q^{M_2}(t)) \in [p_{j-1}, p_j] \right\}. \tag{4.26}$$

Whence the global similarity set, which depends on the partition $\{p_j\}_{j=0}^n$, is defined as

$$\Gamma_s^{(M_1,M_2)} = \bigcup_{j=1}^{n} \Gamma_s^j. \tag{4.27}$$

If l is a probability measure defined on the universal set containing $\Gamma_s^{(M_1,M_2)}$, the similarity between the q-pling time distributions for the two models will be $l(\Gamma_s^{(M_1,M_2)})$, where

$$l(\Gamma_s^{(M_1,M_2)}) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{\Gamma_s^{(M_1,M_2)}}(\boldsymbol{x}(t))\mathrm{d}t \tag{4.28}$$

and $\boldsymbol{x}(t)$ is the system trajectory in state space. The corresponding finite approximation over a discrete set of observations $\{\boldsymbol{s}_i\}_{i=1}^N$ is

$$l(\Gamma_s^{(M_1,M_2)}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\Gamma_s^{(M_1,M_2)}}(\boldsymbol{s}_i), \tag{4.29}$$

where $\boldsymbol{s}_i = \boldsymbol{h}(\boldsymbol{x}_i)$, $\boldsymbol{x}_i = \boldsymbol{x}(i\tau_s)$, $\boldsymbol{h}$ is some observation function and τ_s is the sampling time.

A few observations are worthy to be mentioning. To this end, let M_3 be a third model with q-pling times $\bigcup_{t\geq 0}\{\tau_q^{M_3}(t)\}$.

- If for some predefined $\epsilon > 0$,

$$1 - \epsilon < l(\Gamma_s^{(M_1,M_2)}) \leq 1, \tag{4.30}$$

  we say that model M_1 is similar to model M_2 up to ϵ and the partition in question. For a given partition, greater similarity between q-pling time distributions of two models across the attractor is reflected by smaller ϵ.

- Inequality (4.30) effectively defines an equivalence relation between models. Indeed, if we alternatively write (4.30) as $M_1 \sim M_2$, the reflexivity, symmetry and transitivity properties follow in a pretty straight forward way [9].

- If for a given partition

$$l(\Gamma_s^{(M_1,M_2)}) > l(\Gamma_s^{(M_1,M_3)}), \tag{4.31}$$

we say that model M_1 has q-pling times distribution more similar to that of model M_2 than to model M_3 on the attractor in question.

We shall employ this measure in section 4.4.

4.4 Takens and Predictability

In this section, we look at the application of Takens theorem [70] and its implications for the predictability of chaotic systems. Essentially, it states conditions under which a measurement function h yields, with probability one, a coordinate function $\boldsymbol{H}$ which is a differentiable imbedding [49]. This affords us the benefit of moving into another coordinate space and yet preserve ergodic measures [13, 49, 70]. To see this, let $\boldsymbol{\varphi}_t$ denote the dynamical flow on some manifold. Then the flow on the reconstructed manifold is given by

$$\boldsymbol{\phi}_t = \boldsymbol{H}\boldsymbol{\varphi}_t\boldsymbol{H}^{-1}. \tag{4.32}$$

Applying the chain rule to equation (4.32) yields

$$D\boldsymbol{\phi}_t = D\boldsymbol{H}D\boldsymbol{\varphi}_t D\boldsymbol{H}^{-1}, \tag{4.33}$$

[9] An equivalence relation satisfies reflexivity, $M_1 \sim M_1$, symmetry, which is $M_1 \sim M_2 \Rightarrow M_2 \sim M_1$ and transitivity which is, $M_1 \sim M_2$ and $M_2 \sim M_3$ implies $M_1 \sim M_3$.

which implies that the matrices $D\boldsymbol{\phi}_t$ and $D\boldsymbol{\varphi}_t$ are similar [19, 49] with the similarity transformation being $D\boldsymbol{H}$. It then follows that the eigenvalues of $D\boldsymbol{\varphi}_t$, denoted by $\lambda(D\boldsymbol{\varphi}_t)$, are contained in those of $D\boldsymbol{\phi}_t$, denoted by $\lambda(D\boldsymbol{\phi}_t)$. In capsule form,

$$\lambda(D\boldsymbol{\varphi}_t) \subseteq \lambda(D\boldsymbol{\phi}_t). \tag{4.34}$$

Although (4.34) guarantees the preservation of global quantities like Lyapunov exponents, it places no restriction on local measures of predictability. This, in turn, means finite time measures in the imbedding space may be rather different from those in the system state space. This will inevitably be true when our models are imperfect. The third contributing factor is the presence of "spurious" exponents (eigenvalues/singular values) that creep in when we move into embedding space. Recall that to compute the q-pling times in delay space, we introduced the vector in equation (4.15). This may be related to the delay vector $\boldsymbol{x}_n$ by the mapping:

$$\boldsymbol{G} : \boldsymbol{X}_n \rightarrow \boldsymbol{x}_n. \tag{4.35}$$

This mapping is clearly one to one. If the dynamics of $\boldsymbol{X}_n$ are gorvened by $\boldsymbol{\Phi}_t$, the flow on the delay space is gorvened by $\boldsymbol{G}\boldsymbol{\Phi}_t$ and this leads to the relation

$$\boldsymbol{G}\boldsymbol{\Phi}_t = \boldsymbol{H}\boldsymbol{\varphi}_t\boldsymbol{H}^{-1}. \tag{4.36}$$

whence the chain rule yields

$$D\boldsymbol{G}D\boldsymbol{\Phi}_t = D\boldsymbol{H}D\boldsymbol{\varphi}_tD\boldsymbol{H}^{-1}. \tag{4.37}$$

From equation (4.36), we cannot conclude that $\boldsymbol{\Phi}_t$ and $\boldsymbol{\varphi}_t$ are isomorphic and 4.37 does not guarantee similarity between $D\boldsymbol{\Phi}_t$ and $D\boldsymbol{\varphi}_t$. This means that if one has two models, one in delay space and another in system state space, the variations in predictability will inevitably be different. In the next subsection, we shall quantify the differences in q-pling time distributions for various models of the circuit.

4.4.1 Q-pling times

In this subsection, we compare variations in the performance of different models of the circuit built in different coordinate spaces. Q-pling times make sense when we want to see variations in model performance on a micro scale, at which level the linearised dynamics are a good approximation.

In figure 4.8, we saw that distributions of doubling times of the circuit obtained by approximating the Jacobians from the local dynamics in delay space vary almost uniformly[10] across the attractor for the three directions of initial perturbations considered. These exhibit striking similarities to similar computations in system state space (not shown). Next, we considered q-pling times of four RBF models of the circuit, M_1, M_2, M_4 and M_5. Models M_1 and M_5 were built in 3D delay space, model M_2 in 4D delay space and model M_4 in 3D measurement space. The variations of their doubling times are shown in figure 4.9, all projected into 3D delay space for ease of comparison.

It is interesting to note that the two models built in 3D delay space, have strikingly similar pictures (see figures 4.9(a) and (d)), and are clearly different from the other two. The initial uncertainty was in the Lyapunov direction and the doubling time was defined to be the time for the uncertainty in the forecast of the prediction coordinate, V_3, to double. Also, the other two figures are different from each other (see figure 4.9(b) and (c)).

We then considered the octapling of errors initially on the axis of the prediction

[10]Due to poor approximation.

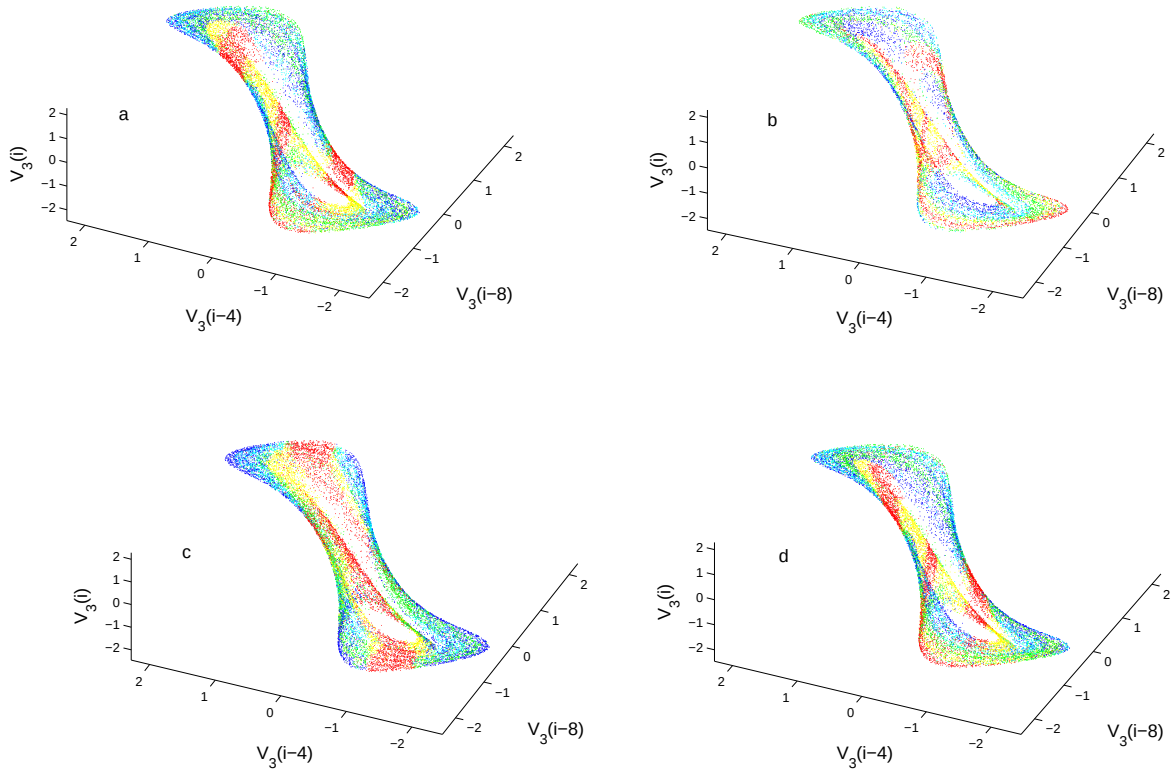


Figure 4.9: Distributions of doubling times for models (a) M_1, (b) M_2, (c) M_4 and (d) M_5 of the circuit. Models M_1 and M_5 were built in 3D delay space, model M_2 in 4D delay space and model M_4 in measurement space. From measurement space, the distribution of doubling times for model M_4 were then mapped into delay space. Red indicates $F(\tau_2) < 0.2$, yellow indicates $0.2 < F(\tau_2) < 0.4$, green indicates $0.4 < F(\tau_2) < 0.6$, cyan indicates $0.6 < F(\tau_2) < 0.8$ and blue indicates $F(\tau_2) > 0.8$. $F(\cdot)$ is the cumulative distribution function of the doubling times. In each case we considered the doubling of the prediction variable with the initial vector in the Lyapunov direction.

coordinate to obtain pictures shown in figure 4.10. Again, the models in 3D delay space show striking similarities in contrast to models in measurement space and 4D delay space. A table of similarity measures for these is shown in table 4.1. Notice that $l(\Gamma_s^{(M_1,M_5)}) = 0.683$, and is the highest value. Next in magnitude is $l(\Gamma_s^{(M_1,M_2)}) = 0.351$. These results support the conclusion that the greatest similarity is exhibited by model M_1 and M_5, which were built in 3D delay space, followed by similarity between these two with model M_2, another delay space model. Model M_4

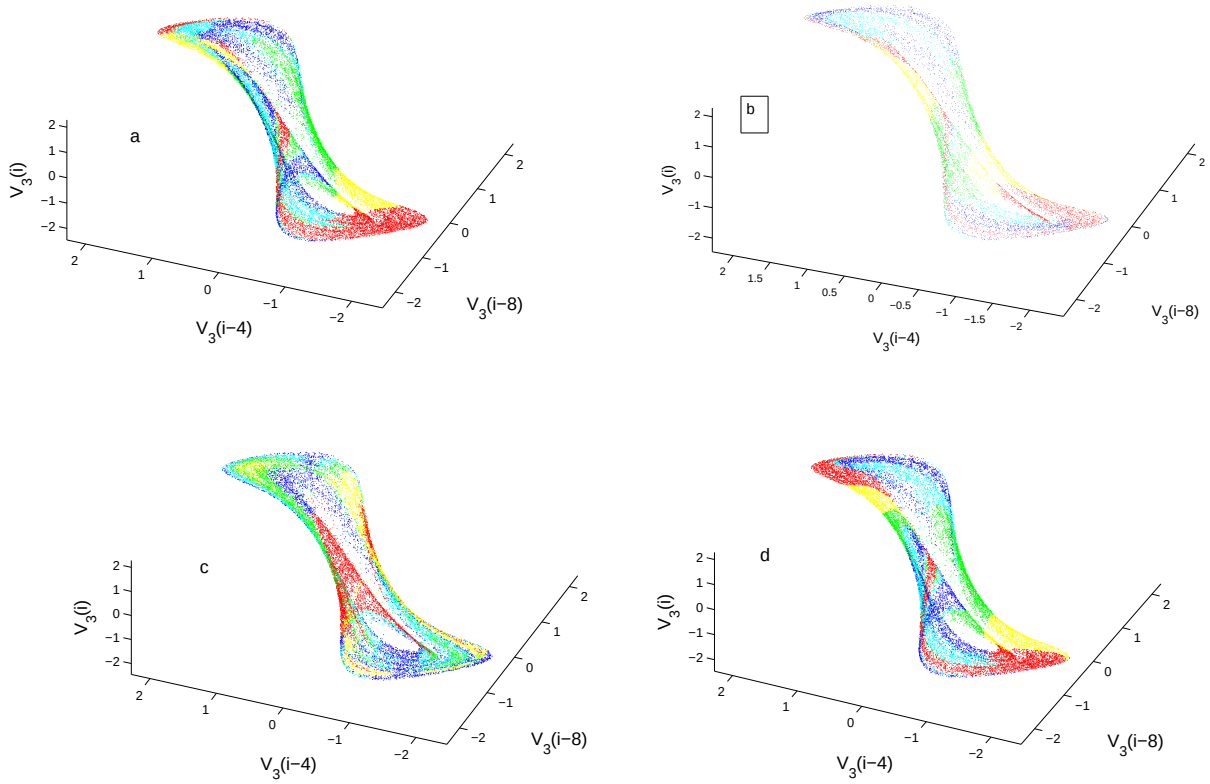


Figure 4.10: View of the MS circuit showing distributions of octapling times for models (a) M_1, (b) M_2, (c) M_4 and (d) M_5 of the circuit. Models M_1 and M_5 were built in 3D delay space, model M_2 in 4D delay space and model M_4 in measurement space. From measurement space, the distribution of octa-pling times for model M_4 were then mapped into delay space. Red indicates $F(\tau_2) < 0.2$, yellow indicates $0.2 < F(\tau_2) < 0.4$, green indicates $0.4 < F(\tau_2) < 0.6$, cyan indicates $0.6 < F(\tau_2) < 0.8$ and blue indicates $F(\tau_2) > 0.8$. $F(\cdot)$ is the cumulative distribution function of the doubling times. In each case we considered the octa-pling of the prediction variable with the initial error on the axis of the prediction variable.

Model	M_1	M_2	M_4	M_5
M_1	1	0.351	0.150	0.683
M_2	0.351	1	0.196	0.325
M_4	0.150	0.196	1	0.162
M_5	0.683	0.325	0.162	1

Table 4.1: Table of values of $l(\Gamma_s^{(i,j)})$, the similarity measure, for the models indicated on the first column and first row of the table whose octapling time variations are shown in figure 4.10. According to this table, all the models in delay space exhibit the least similarity to model M_4 (model in measurement space).

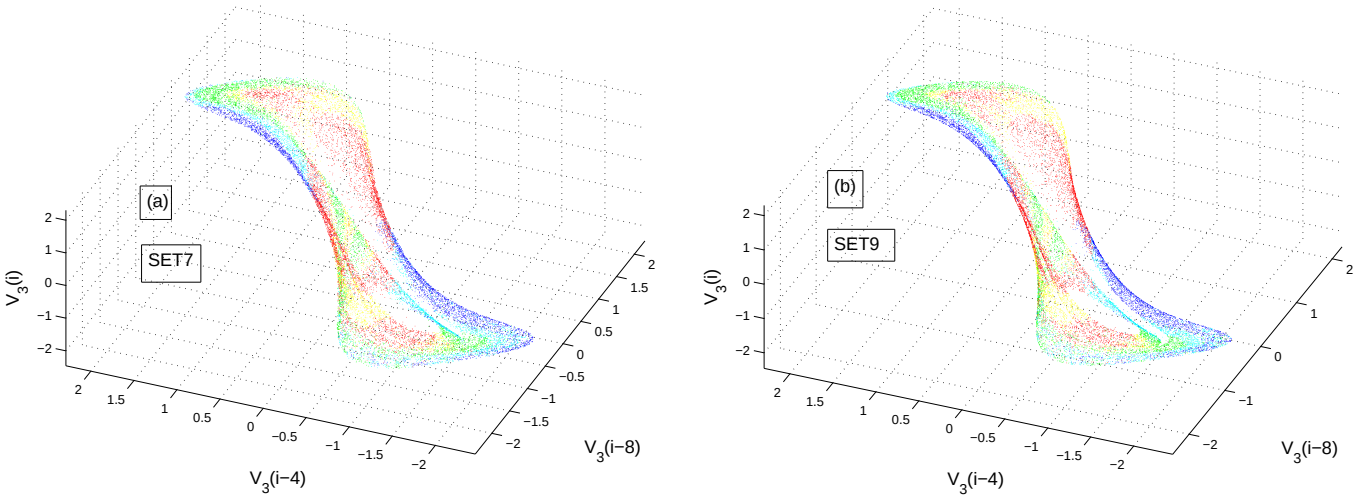


Figure 4.11: Views of the MS circuit showing distributions of doubling times for model 4 of the circuit on data sets 7 and 9. Red indicates $F(\tau_2) < 0.2$, yellow indicates $0.2 < F(\tau_2) < 0.4$, green indicates $0.4 < F(\tau_2) < 0.6$, cyan indicates $0.6 < F(\tau_2) < 0.8$ and blue indicates $F(\tau_2) > 0.8$. $F(\cdot)$ is the cumulative distribution function of the doubling times. We used $||\boldsymbol{\epsilon_0}||=10^{-6}$. The similarity between these two pictures suggests that the circuit dynamics are not altered by the ambient temperature fluctuations.

manifests the least similarity with the rest of the models. It seems plausible to conclude that these similarities and differences are largely due to differences in modelling spaces rather than model error.

We also investigated distributions of doubling times of the circuit by looking at data sets obtained on different days. Recall that each complete data set from the circuit was taken over 14 hours duration. These yielded views of the attractor in delay space shown in figure 4.11. The striking similarity between these two figures is evidence that the predictability of the circuit did not change, especially with respect to ambient temperature fluctuations. This is very much welcome because, if true, it gives us the go ahead to use the limitless amounts of data in model building, which models we can then use to explore limits to predictability.

4.4.2 Q-pling regions

In section 4.1, we saw that the regions where doubling occurs on the Lorenz attractor do not reflect the natural measure of the underlying attractor. In the perfect model scenario, the q-pling regions are where the underlying flow is very unstable. When the model is imperfect, they represent where the model is most unstable. In this section, we have looked at the q-pling times of four (imperfect) models and saw that those of model M_1 and M_5 were strikingly similar. The question that still stands is: What about the regions where q-pling occurs (or simply *q-pling regions*)? Will model 1 and 5 still exhibit the most similarity when we look at their q-pling regions? Indeed they do as shown in figure 4.12.

Again, we see that figures 4.12(b) and (c) are clearly different from each other and (a) and (c). Not withstanding that, the q-pling regions of the four models share the common aspect of not reflecting the underlying circuit attractor. This is not always the case. For instance, octa-pling regions for initial perturbations in the Lyapunov directions for the octa-pling time distributions shown in figure 4.8 reflect the underlying measure as shown in figure 4.13.

4.4.3 Conclusions

These results have profound implications for the modelling of chaotic systems. They suggest that instead of seeking a single, best, high resolution model, it is persuasive to seek multiple models that out-perform each other on the various regions. It is also advisable to build the models in different coordinate spaces. The question of how to combine the models will be addressed in chapter 6.

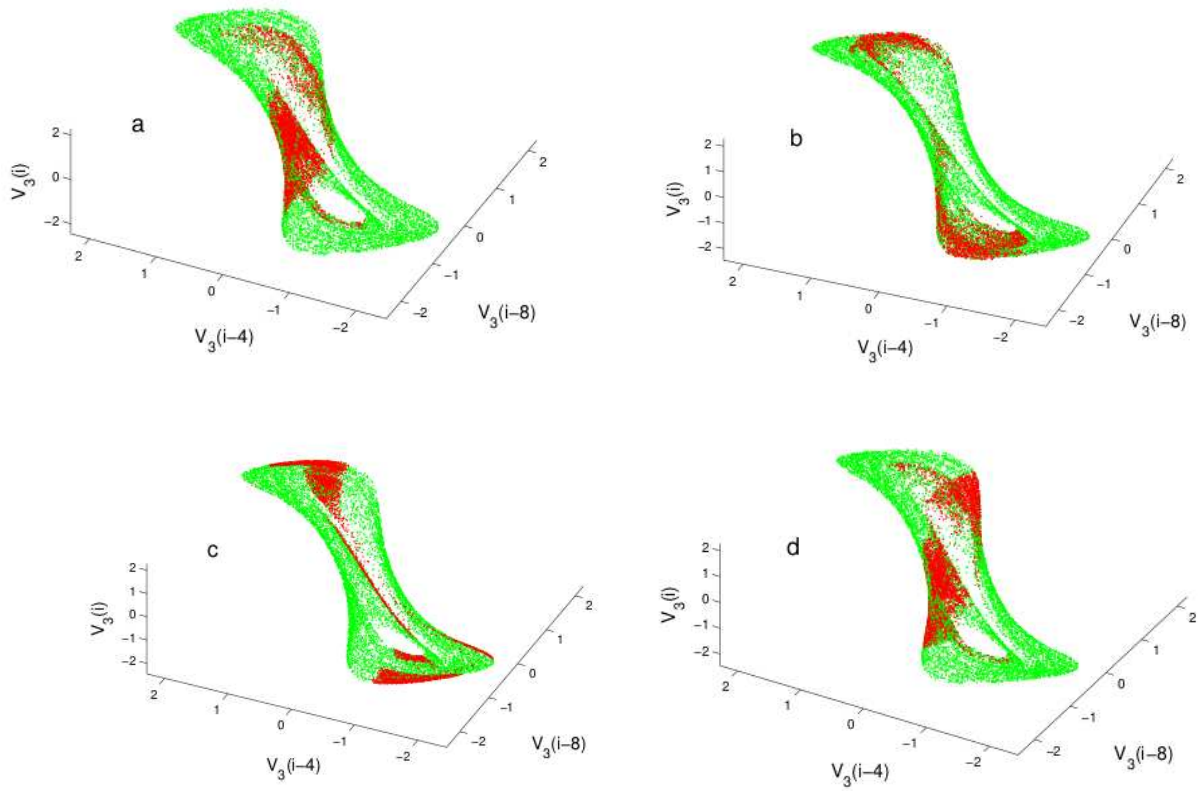


Figure 4.12: Regions where uncertainty octapling occurs for diagrams shown in figure 4.10. Green reflects the underlying circuit attractor and red indicates regions where octapling occurs.

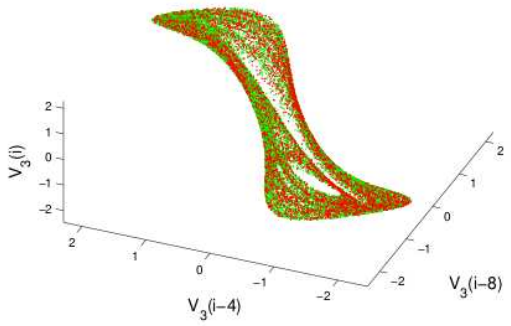


Figure 4.13: The region where uncertainty octapling occurs for the diagram shown in figure 4.8(a). Green dots indicate the initial conditions and red dots indicate where doubling occurs. Notice that the regions where octapling time occurs reflect the underlying natural measure of the circuit.

4.5 One-step-error q-pling times

In the previous sections, we considered what happens when some initial small vector is iterated by a sequence of tangent matrices. This inherently places a restriction on the error size so that the assumptions of linearity remain valid. Also, the q-pling times obtained using an imperfect model do not really tell us what happens to the dynamics of the underlying flow. Rather, they tell us how the models behaves at various regions of the attractor. Large q-pling times indicate that the model is relatively stable under iteration in a given region of the system's attractor. However, a model may be stable while diverging from the underlying system's trajectory. Therefore, we introduce one-step-error q-pling times (OSEQ-times) as simple way to quantify regional losses in predictability. To define these, let us consider observations, $\{\boldsymbol{x}_n\}_{n=0}^{N}$, of a dynamical system which is modelled by the map, $\boldsymbol{\phi}(\boldsymbol{x})$. Let $\boldsymbol{\varepsilon}_0$ be some finite pre-assigned threshold. The one-step-error q-pling time is defined as

$$\tau_q^{(1)}(\boldsymbol{\varepsilon}_0, \boldsymbol{x}_0) = \inf_{n\geq 0} \left\{ n\tau_s \middle| ||\boldsymbol{x}_n - \boldsymbol{\phi}^n(\boldsymbol{x}_0)|| \geq q||\boldsymbol{\varepsilon}_0|| \right\} \tag{4.38}$$

This allows us to explore the variation of models' predictability loss at macro scale.

To compute the OSEQ-times of a given model, there is need to use the same value of $||\boldsymbol{\varepsilon}_0||$ across the whole attractor. In fact, when we have to compare multiple models, we need to choose a uniform threshold for all the models so that none are more heavily penalised than others. To choose this threshold, we consider the distributions of one-step-errors (OSEs). For example, if we consider two models, say model M_1 and M_2, with global mean and median of OSEs $\mu^{(1)}, m_1$ and $\mu^{(2)}, m_2$ respectively, we can choose $\epsilon_0 \geq \{\mu^{(1)}, \mu^{(2)}, m_1, m_2\}$. The OSEQ-time is then the minimum time for

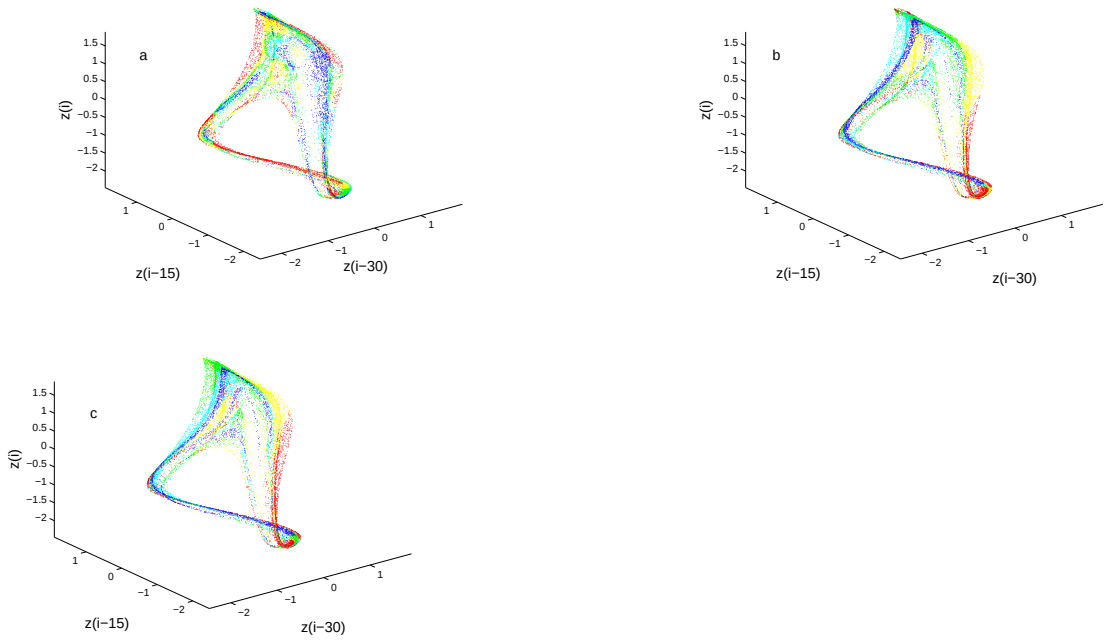


Figure 4.14: A view of the MS attractor showing OSEQ times for model M_{1e} with (a) $q = 2$, (b) $q = 64$ and (c) $q = 128$. The distributions of the OSEQ times are as follows: red reflects $F(\tau_q) < 0.2$, yellow $0.2 < F(\tau_q) < 0.4$, green $0.4 < F(\tau_q) < 0.6$, cyan $0.6 < F(\tau_q) < 0.8$ and blue $F(\tau_q) > 0.8$, where $F(\cdot)$ is the cumulative distribution function of τ_q.

the forecast error to exceed $q\epsilon_0$. In our considerations, we define the forecast error with respect to some variable of interest. If the variable we seek to forecast is x_n and the forecast is $\hat{x}_n$, then the forecast error is $\epsilon = |x_n - \hat{x}_n|$.

Let us consider models of the MS system. We use model M_{1e} with $q = 2, 64, 128$. Views of the variations of OSEQ times on the attractor are shown in figures 4.14. The initial error was chosen to be $\epsilon_0 = 0.001$. From this figure, it is clear that the variability of the OSEQT is organised across the attractor. Secondly, as q is increased, there appears to be convergence in the structure of the variation of predictability. This is understandable since at higher values of q, what we get is the variation at a macro scale, at which scale a given model would begin to fail in the same sorts of regions.

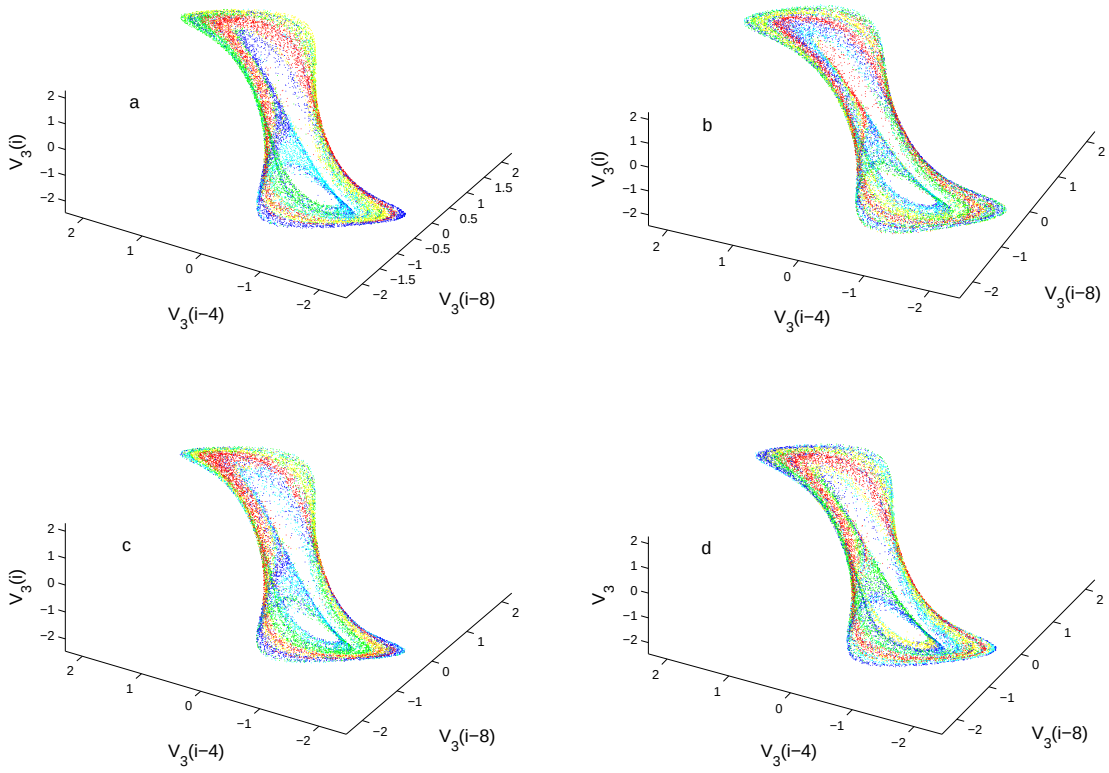


Figure 4.15: Views of the circuit attractor showing variations of OSE-quadrupling times of models (a) M_1, (b) M_2, (c) M_4 and (d) M_5.

We can also note that at low q ($q = 2$, for example), there appears to be randomness in the variation of predictability, and this goes away at higher q-pling times.

Turning to the circuit: For each model, we set $\epsilon_0 = 0.2$ and obtained the quadrupling ($q = 4$) times with colour plots shown in figure 4.15. The figures are very similar suggesting that model failure at macro scale could be largely due to the dynamics of the circuit rather than model error. It means that, medium term forecasts fail at the same sorts of regions. It is, however, still the case that performances of the models across the circuit are different. A table showing values of the similarity measure is shown in table 4.2. What we can learn from this table is that models M_1 and M_5

Model	M_1	M_2	M_4	M_5
M_1	1	0.332	0.427	0.554
M_2	0.332	1	0.388	0.273
M_4	0.427	0.388	1	0.346
M_5	0.554	0.273	0.346	1

Table 4.2: Table of values of $l(\Gamma_s^{(i,j)})$, the similarity measure, for the models indicated on the first column and first row of the table whose OSE quadrupling time variations are shown in figure 4.15.

continue to exhibit the greatest similarity.

The cumulative distributions of the OSE-quadrupling times are shown in figure 4.16. They suggest that model M_2 is actually the best and model M_4 is the worst. Models M_1 and M_5 have barely distinguishable OSE-quadrupling times. One might suspect that model M_2 will win when we use skill scores in chapter 6.

Unlike the case with q-pling regions, the regions where one-step-error q-pling occurs are similar for all the models as seen in figure4.17. They indicate that all the models fail in the same sort of regions, so that chaos could be the major limit to long term prediction.

4.6 Lyapunov q-pling times

The question we wish to address in this section is how the oscillatory nature of the histogram of the Lyapunov doubling times comes about. The histogram for the Lorenz attractor is shown again in figure 4.18 and how the doubling times are distributed on the attractor in figure 4.19(a). The colour partitioning was obtained by using the extrema of the histogram as the partition points. Unlike the one obtained in figure 4.1(a) by using percentiles, different regions are more clearly marked and reveal a

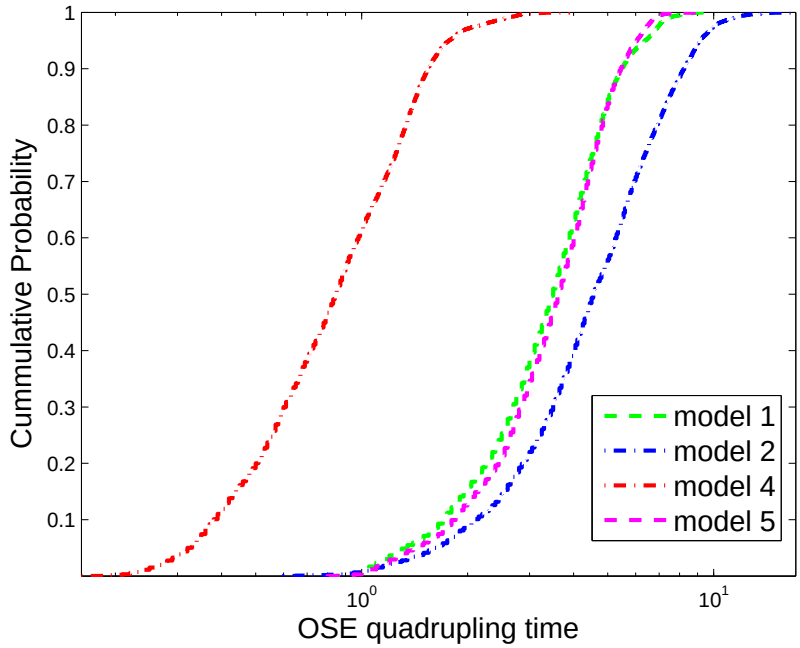


Figure 4.16: Cumulative distributions of OSE quadrupling times, of the circuit models, corresponding to figure 4.15.

banded structure. It indicates that predictability increases outwards from inside the internal bands. For instance, the yellow strip is encapsulated by two strips of cyan and blue, which regions are more predictable than the yellow one. The centre of the attractor is the least predictable as it is coloured in red and this is because it is in the neighbourhood of the saddle point at the origin.

The oscillatory nature of the doubling time distributions is in stark contradistinction to distributions of waiting times in stochastic systems which are exponentially distributed [50, 67]. This non-exponential nature of the distribution of the τ_2 has been attributed to determinism in chaotic systems [67], albeit without explanation. Here, we shall go into the structure of the Lorenz attractor to see how this comes about.

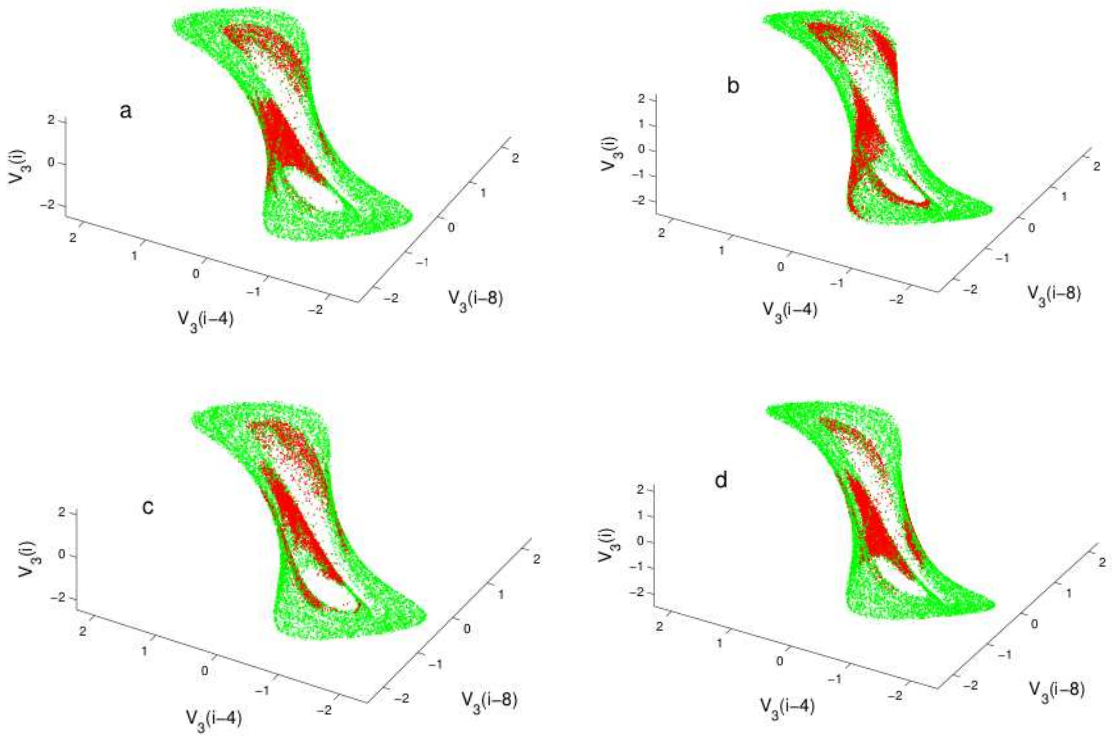


Figure 4.17: Regions where OSE-quadrupling occurs (red) and on the circuit attractor (green) for the models with quadrupling times shown in figure 4.15.

Recently, Tucker [72], using interval arithmetic to solve Smale's 14th problem [63], proved that the geometrical model proposed by Guckenheimer and Holmes [21, 22] is indeed correct. In particular, he proved that the Lorenz strange attractor exists. In this thesis, we propose that the oscillatory nature exhibited by the distribution of doubling times is a signature of the self-similarity of the Lorenz attractor. The Lorenz system has three equilibrium points, one at the origin and the other two at

$$C_{\pm} = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1). \tag{4.39}$$

These two are stable foci, the origin is a saddle splitting the attractor into two by its two dimensional stable manifold, $W^s(0)$. It is this manifold that also introduces a discontinuity into the corresponding Poincaré return map [17, 22, 72], rendering the

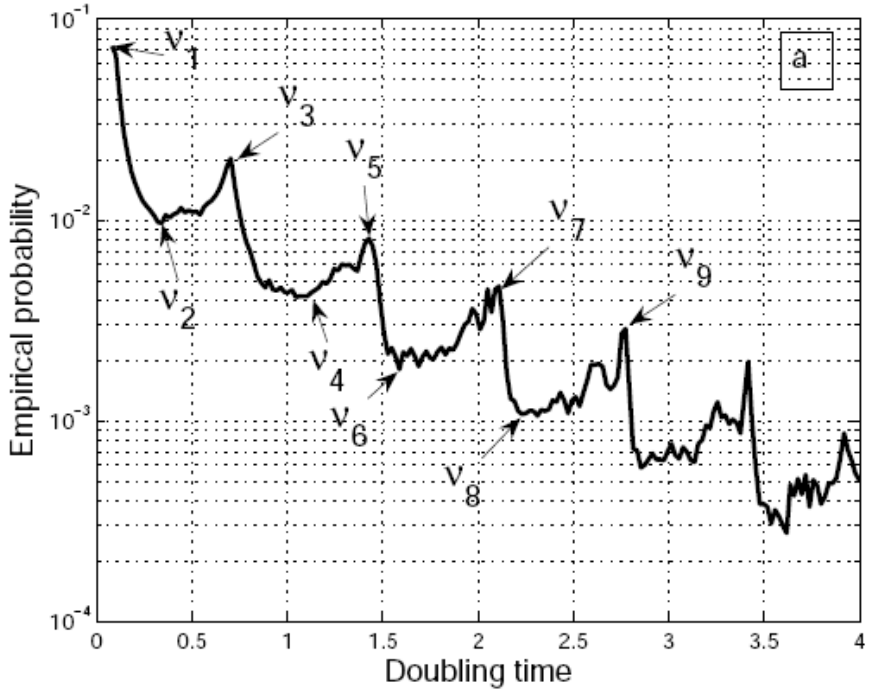


Figure 4.18: Histogram of Lyapunov doubling times of the Lorenz attractor. Notice the oscillatory nature, a signature of self-similarity. The successive extrema have been denoted ν_i, $i = 1, \ldots, 9$.

attractor non-hyperbolic. The attractor is bounded away from the nonzero steady states, $C_\pm$. The dynamics of the flow have been explained by looking at a Poincaré map defined on a rectangle Σ, whose opposite sides are formed by the one-dimensional stable manifolds of these equilibria, $W^s(C_\pm)$. This rectangle is contained in the plane $z = r - 1$, and $\dot{z} < 0$ on its interior. It intersects $W^s(0)$ on some line, say D. All trajectories pass down through Σ, and then spiral around either C_- or C_+.

A plot of the points whose doubling times correspond to the extrema of the histogram in figure 4.18 indicates that they are separated by Σ into two parts as shown in figure 4.19 (b), with points where the minima occur being above Σ and those where the maxima occur being below Σ. The successive extrema are labelled ν_i, with the

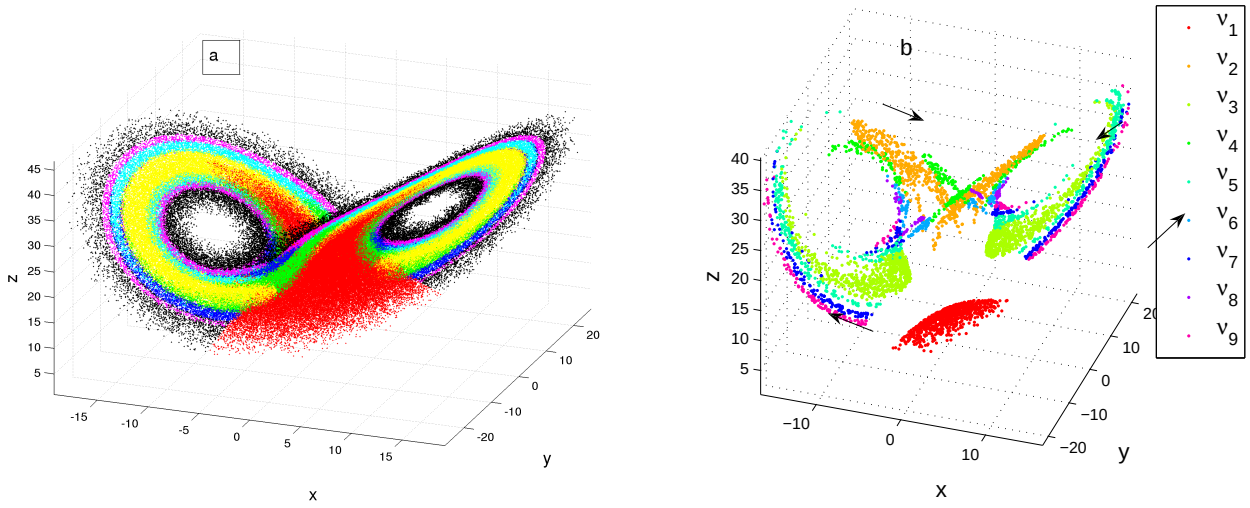


Figure 4.19: (a) View of the Lorenz attractor showing distributions of doubling times in the Lyapunov with the colour scheme dictated by the extrema of figure 4.18: Red reflects $\tau_2 < 0.31 = \nu_2$, yellow indicates $\nu_2 = 0.31 < \tau_2 < 0.71 = \nu_3$, green reflects $\nu_3 = 0.71 < \tau_2 < 1.04 = \nu_4$, cyan reflects $\nu_4 = 1.04 < \tau_2 < 1.39 = \nu_5$, blue reflects $\nu_5 = 1.39 < \tau_2 < 1.58 = \nu_6$, and black reflects $\tau_2 > 1.58 = \nu_6$ and (b) points on the attractor whose doubling times correspond to extrema on the PDF of figure 4.18.

odd subscripts denoting maxima and the even subscripts denoting minima. Points corresponding to the first maximum are the red ones at the very bottom. This means that points with the shortest doubling times lie in this region of the attractor, and their measure is leading. The points coloured in gold are where the first minimum occurs, and they lie above Σ. The flow comes to these points after going round either of the equilibria, $C_{\pm}$ and while this happens, the measure of the regions where the intervening doubling times occur would be decreasing monotonically. From the ν_2 to the ν_3 region, we then have an increase in the measure of the regions with the intervening doubling times; that is, until we get to the region coloured light green (or lime). From then on (i.e. ν_i, $i \geq 4$), regions of the doubling times where the next minimum (maximum) occurs sandwich regions of the previous minimum (maximum). In this sense, distributions of the Lyapunov doubling times reveal the self similar na-

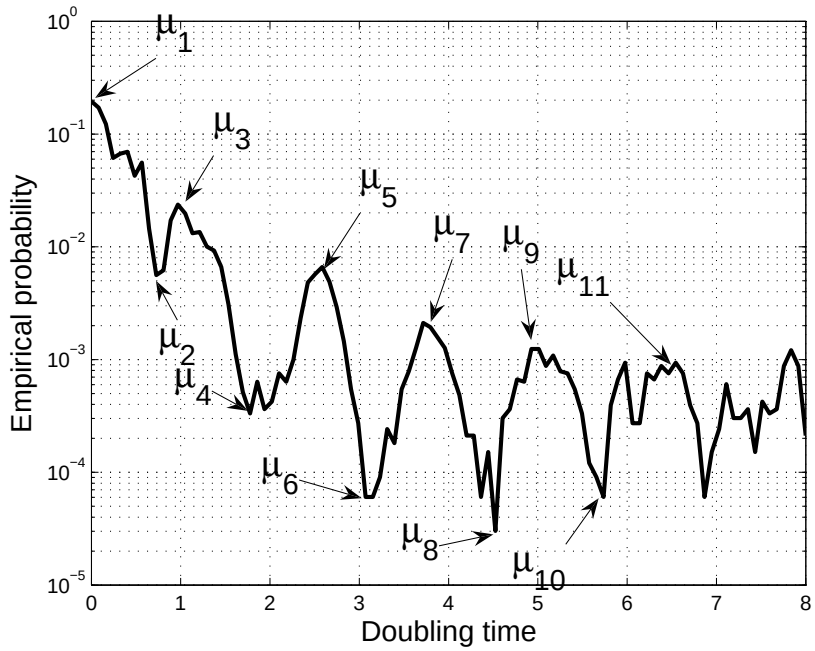


Figure 4.20: Histogram of Lyapunov doubling times of the MS attractor. Notice the oscillatory nature, a signature of self-similarity. The successive extrema have been denoted μ_i, $i = 1, \ldots, 11$.

ture of the underlying attractor.

Let us now turn to the MS attractor. The histogram of doubling times is shown in figure 4.20. The distribution exhibits an oscillatory behaviour not much weaker than that of the Lorenz attractor. Nevertheless, the MS attractor does not yield the kind of beauty exhibited by the Lorenz attractor in the sense of doubling times, although manifesting the underlying self similarity in the histogram of doubling times. Points on the attractor where successive extrema occur on the histogram are shown on figure 4.21. Observing from figure 4.20, most of the measure is taken up by points that double very quickly. Regions corresponding to μ_2 and μ_3 are in a bi-metallic strip sort of formation (The dark and light green points in figure 4.21). These regions of slightly longer doubling times occur where the flow comes toward the only stationary

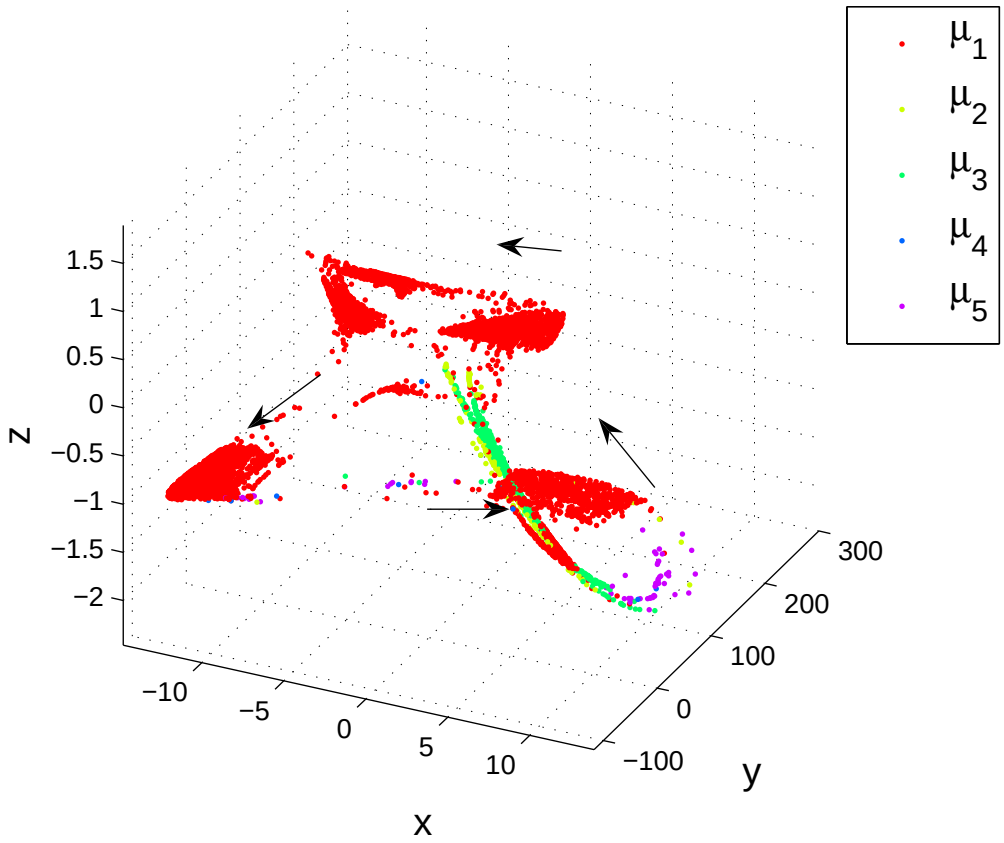


Figure 4.21: Points on the MS attractor whose doubling times correspond to successive extrema on the PDF of figure 4.20. The arrows indicate the direction of the flow.

point at the origin. This is reminiscent of the Lorenz case where regions of minima occurred above the plane Σ where the flow was approaching the saddle at the origin. Overall, there is no clear splitting between the maxima and minima regions in the MS attractor.

Finally, we consider the MS circuit. The distribution of doubling times of the circuit with respect to model M_4 are given in figure 4.22. We show these for two data sets that were collected on different days; SET7 and SET9. These are oscillatory like we found in the perfect model scenario. Another important point to note is that these

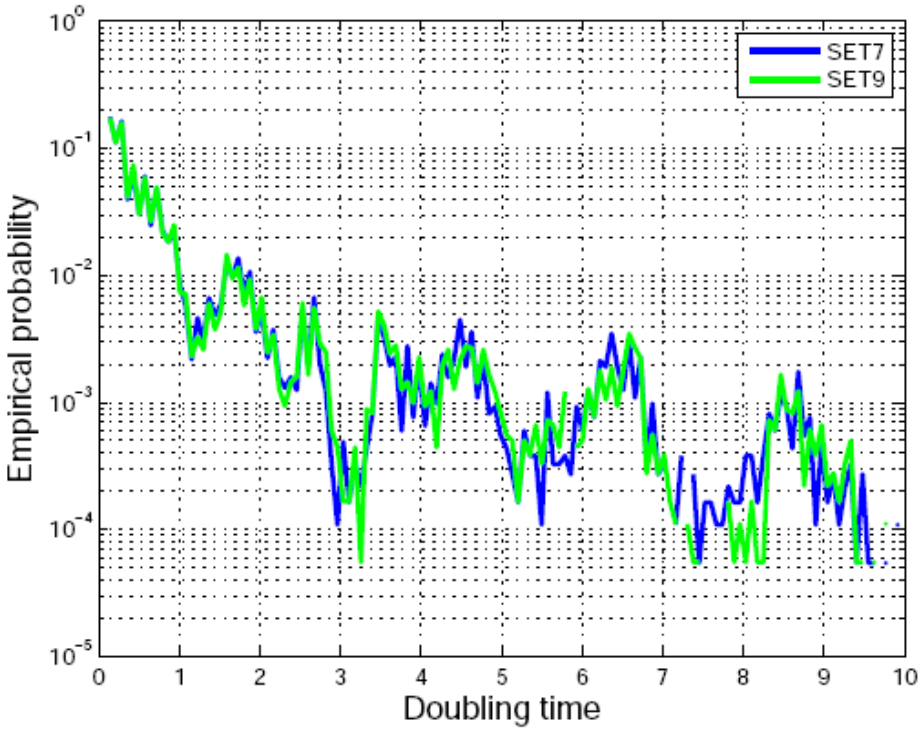


Figure 4.22: Histogram of Lyapunov doubling times of the MS circuit using model M_4 on data sets 7 and 9. It is oscillatory just like those obtained in the perfect model scenario. We used $||\epsilon_0|| = 10^{-6}$.

distributions are very much alike, suggesting that the behaviour of the model is not different for the data sets, notwithstanding ambient temperature variations. In fact, views of the attractor showing the Lyapunov doublings corresponding to model 4 are very similar (see figure 4.11). We shall not investigate the extrema of the histogram in figure 4.22 since it is more about the performance of the model than the properties of the circuit.

At macro-scale, distributions of one-step-error q-pling times tend to show much weaker oscillations (see figure 4.23) and may even resemble exponential decay, depending on the quality of the model. Since model M_2 is of very good quality, it is

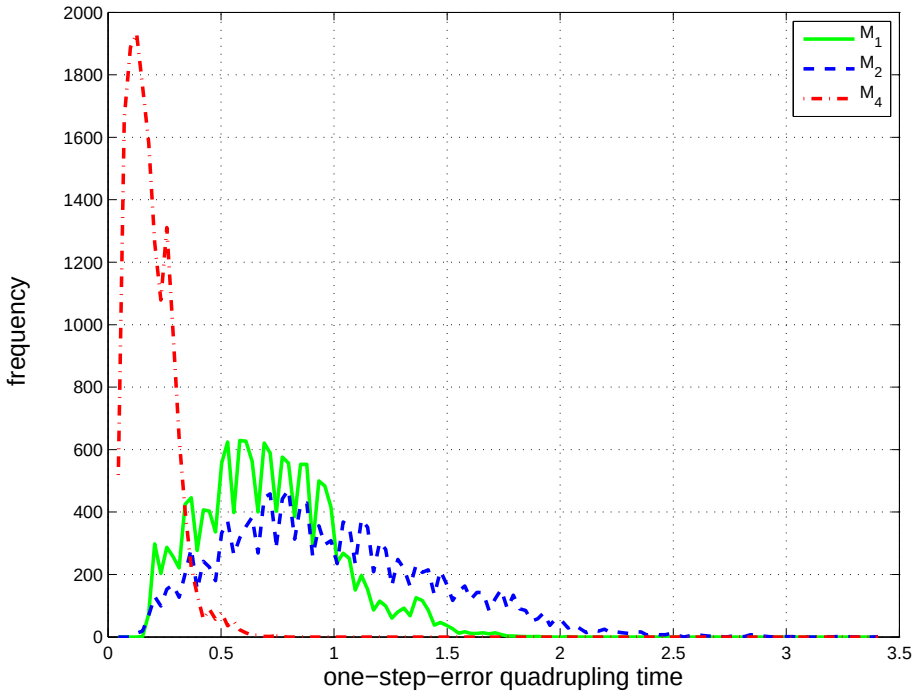


Figure 4.23: OSE quadrupling times for the models M_1 (green), M_2 (blue) and M_4 (red) of the circuit.

centred away from zero.

4.7 Conclusions

In this chapter, we investigated variations in predictability both in the perfect model scenario and the imperfect model scenario. In the perfect model scenario, we looked at the Lorenz attractor and the MS attractor. In the IMS, we looked at the MS system and the circuit. These variations in predictability were quantified by the use of q-pling and OSEQ-times. Considering the imperfect model scenario was a step further than a lot of traditional manoeuvres which often investigate properties of the models as though they were properties of the underlying system.

As one might have expected, we found that distributions of q-pling and OSEQ times

depend on which model one is using, albeit these were well organised. It was also seen that these do not only depend upon the model, but also on the kind of coordinate space in which the model lies. For that reason, we suggested that building models in various coordinate spaces to optimise forecasts might be better than using a single high resolution model.

We also found that at macro scale, OSEQ-times yield variation structures that are strikingly similar across different models, which suggests that limitations to long term prediction of the circuit could be due to chaos. Unlike qpling times, OSEQ-times yield very week oscillatory histograms, and may even appear exponetial.

Finally, we unravelled the reason for the oscillatory nature of the histograms of doubling times of the Lorenz and MS system, and found it is a signature for self similarity in the Lorenz attractor, but not in the MS attractor. It was, however, found pointless to do the same thing in the imperfect model scenario where a lot depends on the underlying model.

Original work in this chapter includes:

- The similarity measure introduced in section 4.3 that we used to compare variations in predictability.
- The concept of one-step-error q-pling (OSEQ) times and the use of cumulative distributions of OSEQ times to rank models.
- That the coordinate space in which the model lies largely affects variations in

predictability and this necessitates using multiple models in forecasting. It also means our failure to forecast as good at certain times as others may not be that we have entered a region where the circuit dynamics are very unstable, but that our model is not good enough to capture that region. Predictability variation is model dependent.

- Showing that the circuit exhibited strikingly similar q-pling time distributions over two different temperature regimes.

- Unravelling the oscillatory nature of the histograms of doubling times of both the Lorenz and MS attractor. Indeed the oscillatory nature is a signature of determinism.

In the next chapter, we shall discuss ensemble prediction and the concept of *skill scores*. Skill scores can be used to rank models. The ideas of that chapter will be carried over to chapter 6, where we discuss combining dynamical models.

Chapter 5

Ensemble Prediction

Given a perfect model, the only limit to predictability is uncertainty in the initial conditions. On the other hand, given an imperfect model without uncertainty in the initial conditions, model error will impede prediction. This is because the system state space and the model state space are different. However, there is a corresponding initial condition on the solution manifold of the model from which we could hope to gain some improved forecasts. In section 5.1, we demonstrate that modelling the slightly perturbed MS system with the MS system can give bad results if one puts in the exact initial conditions. But for some initial conditions, some trajectories obtained from the perturbed initial conditions stay close to the true trajectory longer than those from the unperturbed initial conditions. Such perturbations of the initial conditions are called initial *ensembles*. The main point of § 5.1 is that ensemble prediction is the way to go when confronted with even the simplest case of parametric uncertainty.

Despite the fact that there is consensus on using ensembles to forecast chaotic systems, the issue of how to go about making ensembles has not been resolved. Although knowing the statistical form of the observational uncertainty in the perfect model

scenario (PMS) would help us collect an ensemble of initial conditions [67], this will definitely not suffice in the imperfect model scenario (IMS) because model error has to be taken into account. The question of what we are to do is an ongoing debate in current research. For instance, Lyons [40] argues for a stochastic approach to this problem, since the different regions across an attractor will necessarily be different. In an ECMWF group meeting chaired by Lyons [1], it was concluded that it would be better to find a way of searching selectively for best ensemble members, in a way reminiscent of how a spell checker works. The point is that, "on encountering a dubious word, a spell checker presents an ensemble of alternative words rather than giving the user an entire dictionary and set of associated probabilities for each word contained therein".

In this chapter, our goal is to collect initial ensembles that optimise prediction[1]. For this reason, we opt for an ensemble that optimises prediction in the sense of some *skill score*, which we shall define later. Given an observation, $\boldsymbol{x}_0$, we make random perturbations from a Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\mathrm{diag}(\epsilon_1^2, \epsilon_2^2, \ldots, \epsilon_n^2)$. The main point of this chapter is that, given the statistical form of the random perturbations, the problem of collecting an initial ensemble is tantamount to determining the size of the perturbations, $\mathrm{diag}(\epsilon_1^2, \epsilon_2^2, \ldots, \epsilon_n^2)$. To this end, we shall introduce skill scores. For the moment, let it suffice to say that a skill score measures the *reliability* and *sharpness* of forecasts. Suppose the forecast probability density function of a random variable X from an ensemble of initial conditions using an imperfect model is $f(x)$. Reliability is a measure of how "close" $f(x)$ is to

[1]We shall use prediction and forecasting interchangeably.

the corresponding perfect forecast $p(x)$. On the other hand, sharpness measures how tightly packed the forecast ensembles are. If our forecasts are reliable, and yet have a standard deviation comparable to that of the *climatology*, the long term probability density function of the underlying system, then the forecasts are not useful, and we may as well just use the climatology. Collecting initial ensembles that optimise prediction in the sense of skill scores amounts to finding a perturbation size that gives us the best possible skill (decomposition of sharpness and reliability). We don't want to gain either attribute at the expense of the other. In § 5.2, we discuss skill scores and then introduce the information based skill score, *Ignorance* in § 5.3. Since skill scores are functions of forecast probability density functions (PDFs) and verifications, we discuss a way of obtaining PDFs from a discrete set of ensembles in sections 5.4 and 5.5. Section 5.6 is concerned with the use of the Ignorance skill score to find an optimum initial ensemble perturbation.

In § 5.7, we summarise the main points of the chapter and give a list of what constitute original work. We assume that the reader is familiar with some basic statistics. An integral without limits of integration should be understood to mean that we are integrating over the whole real line.

5.1 Perturbed MS system

In this chapter, we shall argue and demonstrate that ensemble prediction is invaluable even in the clean data case when there is model error. To this end we shall use the MS system as an example.

Following Balmforth and Craster [5], the MS system can be written in potential form as

$$\ddot{x} = -\frac{\partial V}{\partial x} - \nu\dot{x} \tag{5.1}$$

$$\dot{z} = A(x,z) \tag{5.2}$$

with $V(x,z) = \lambda\left(\frac{x^4}{12} - \frac{x^2}{2} + xz\right)$ and $A = -\varepsilon\left[z + (x^3/3 - x/\lambda)\right]$, except for a difference in notation and the damping term $\nu\dot{x}$. In terms of the old variables in equation (2.10), $\lambda = 1 - \gamma/\Gamma$ and $\varepsilon = \Gamma^{-1/2}$. The unperturbed MS system corresponds to $\nu = 0$. The question we then pose is: How does perturbing the MS system affect its predictability? This question is closely related to *structural stability*. Given a system

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \tag{5.3}$$

whose perturbed version is

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \nu\boldsymbol{g}(\boldsymbol{x}), \tag{5.4}$$

we say that (5.3) is structurally stable if for all sufficiently small ν, the respective flows for equations (5.3) and (5.4), $\boldsymbol{\varphi_f}(\boldsymbol{x},t)$ and $\boldsymbol{\varphi_{f+\nu g}}(\boldsymbol{x},t)$ are equivalent [21, 22]. This means that there exists a *homeomorphism* $\boldsymbol{h}$ such that

$$\boldsymbol{\varphi_{f+\nu g}}(\boldsymbol{h}(\boldsymbol{x})) = \boldsymbol{h}(\boldsymbol{\varphi_f}(\boldsymbol{x})), \tag{5.5}$$

where the time dependence has been suppressed because it is irrelevant. A *homeomorphism* is a continuous, one-to-one map with a continuous inverse. Equation (5.5) effectively says that the flows of equation (5.3) and (5.4) are isomorphic. In other words, there exists $\boldsymbol{h}$ such that the diagramatical representation of equation (5.5) shown in figure 5.1 commutes, R and D are the respective manifolds. Thus, structural stability guarantees that the underlying attractors will have similar topologies [2].

[2]For a discussion of topological objects, refer to Lee [36].

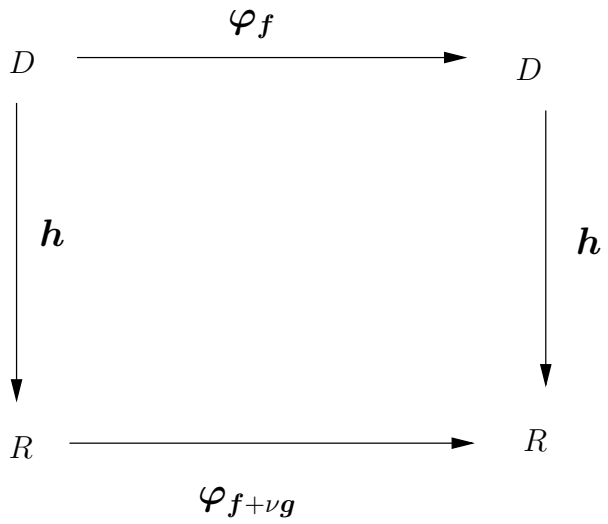


Figure 5.1: Diagramatical representation of equation (5.5).

Although this means that the trajectory patterns are qualitatively similar, they could be quantitatively very different. A much stronger requirement would be that $\boldsymbol{h}$ be a diffeomorphism[3]. If the flows are quantitatively similar, it would still be ill-advised to perform a point forecast by substituting the "exact" initial condition of the underlying system without first transforming it with some diffeomorphism $\boldsymbol{h}$. The reason for this is that the system state space and model state space will be different up to the function $\boldsymbol{h}$. However, the point is that if point forecasts go wrong, we may be better off seeking the function $\boldsymbol{h}$ for better forecasts. In the simplest case, ensemble prediction may be thought of as performing many searches for the function $\boldsymbol{h}$ of the form

$$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\xi}, \tag{5.6}$$

where $\boldsymbol{\xi}$ is a random variable drawn from a distribution of mean $\mathbf{0}$ and covariance matrix $C = \mathrm{diag}(\epsilon_1^2, \epsilon_2^2, \ldots, \epsilon_n^2)$. i.e.

$$\boldsymbol{\xi} \sim \Xi(\mathbf{0}, C).$$

[3]A functions that is differentiable together with its inverse.

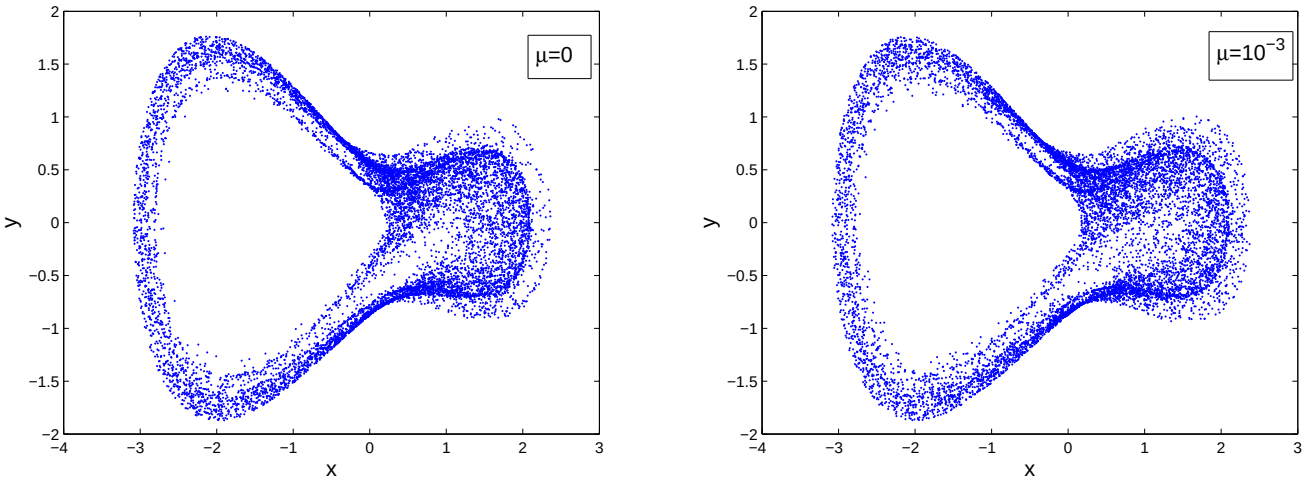


Figure 5.2: Projections of the perturbed Moore Spiegel attractor with $\mu = 0, 10^{-3}$, $\lambda = 0.64$ and $\varepsilon = 0.1$ onto the (x, y)-plane. Notice that the two attractors cannot be distinguished by eye, yet modelling the $\mu = 10^{-3}$ system with the $\mu = 0$ system fails in some regions of state space (see figure 5.3).

Such a search is in vain if the underlying system is not structurally stable. The construction of these distributions is not the subject of our study, but in our computations we use $\boldsymbol{\Xi}(\mathbf{0}, C) = \boldsymbol{\mathcal{N}}(\mathbf{0}, C)$.

To illustrate the point we are driving at, we integrated the MS system with $\lambda = 0.64$, $\varepsilon = 0.1$ and $\mu = 0, 10^{-3}$ and obtained attractors whose projections onto the (x, y) plane are shown in figure 5.2, where $y = \dot{x}$. We then considered modelling the perturbed system (with $\mu \neq 0$) with the original ($\mu = 0$) Moore-Spiegel system. We made ensemble perturbations at various initial conditions, each ensemble containing 32 members, including the unperturbed initial condition. The perturbations were drawn from a normal distribution of standard deviation $\epsilon_x = 10^{-3}$ and zero mean and were made on all the variables, with ϵ_x being the standard deviation of the perturbations of the x variable. For the other variables, we chose independent perturbations

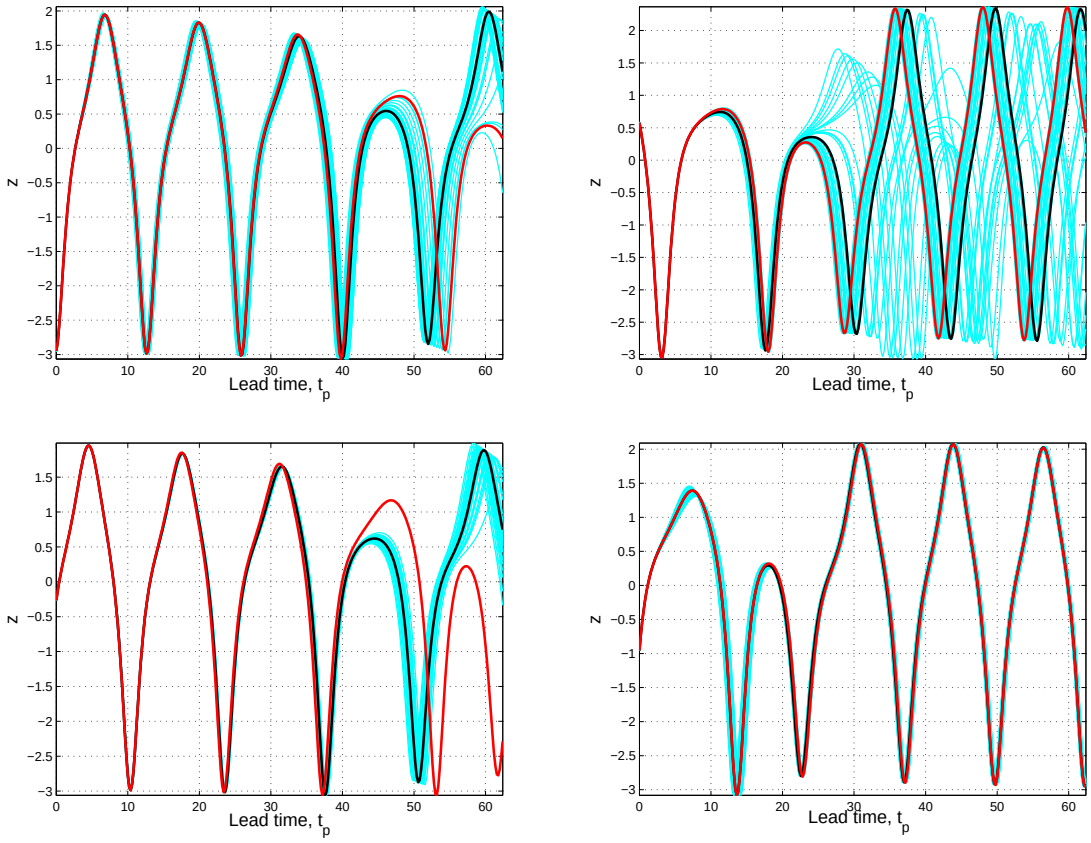


Figure 5.3: Ensemble predictions of the perturbed Moore-Spiegel system ($\mu = 10^{-3}$) using original Moore-Spiegel system as the model. Each ensemble had 32 members obtained by perturbing the initial conditions with Gaussian perturbations of standard deviation $\epsilon_x = 10^{-3}$ for the x-variable and ϵ_y and ϵ_z for the other variables according to equation (5.7). In each figure, the cyan trajectories are the ensemble members, the black trajectory (control) corresponds to unperturbed initial conditions and the red one is the truth. Notice that, in (a)-(c), the control trajectory stays close to the truth for a while before going astray. In (a) and (b), the ensembles do pick the verification when the control goes astray and in (c) both the ensembles and the control go astray after a while. However, most regions like the one in (d) are well tracked by the control.

with standard deviations

$$\begin{aligned} \epsilon_y &= \epsilon_x \frac{\sigma_y}{\sigma_x}, \\ \epsilon_z &= \epsilon_x \frac{\sigma_z}{\sigma_x}, \end{aligned} \tag{5.7}$$

where σ_x,σ_y, σ_z are the standard deviations for the x, y and z variable respectively[4].

Graphs shown in figure 5.3 are a small sample of those obtained by integrating the

[4]Suppose the probability density function of $\boldsymbol{x} = (x, y, z)$ is $p(\boldsymbol{x})$, so that the marginal PDF of x is $p(x)$. Then the mean of x is defined by $\mu_x = \int xp(x)\mathrm{d}x$ and the standard deviation is $\sigma_x = \int (x - \mu_x)^2 p(x)\mathrm{d}x$. Likewise definitions are made for the other variables.

initial ensemble perturbations. In each graph, the cyan trajectories are the ensemble obtained from the perturbed initial conditions, the black trajectory (control) is obtained from the unperturbed initial condition, and the red trajectory is the truth (verification). Notice that in figures 5.3 (a)-(c), the control stays close to the verification for a while before going astray. In figures 5.3 (a) and (b), when the control goes astray, there are still some trajectories from the ensemble that stay close to the verification. However, in figure 5.3 (c), they all begin to go astray together. It should be pointed out that most of the time we get both the control and the ensemble providing good forecasts over a number of cycles. The point is that even when the control trajectory goes astray completely, ensemble predictions with initial perturbations drawn from the Gaussian distribution yield superior results in the sense of *shadowing times*[5]. This example demonstrates how invaluable ensemble prediction is in the imperfect model scenario, with no observational noise. Also, in operational weather forecasting, it is common to use the trajectory corresponding to the "true" initial condition as a control, but we notice here that, at their very worst, ensembles perform as good as the control.

In the next section, we formally introduce *skill scores*, an invaluable tool in probabilistic prediction. The discussion paves way for *Ignorance*, the only skill score that we shall employ.

[5]Loosely put, a model trajectory that stays close, according to some norm and prescribed tolerance, to the observations for a longer time than another is said to cast longer shadows [66].

5.2 Skill Scores

Given a model, we often want to score its performance. In the preceding discussions, it has been argued that model error and/or observational uncertainty make probabilistic prediction on the basis of an ensemble of initial conditions a possible option. Forecasting an ensemble generated at an initial condition $\boldsymbol{x}_0$ provides us with a forecast probability density function (PDF), $f_t(\boldsymbol{x};\boldsymbol{x}_0)$, of the random variable $\boldsymbol{X}$ at time t. How good this forecast is can be determined by the use of some *score*, S, with respect to the perfect forecast PDF, $p_t(\boldsymbol{x};\boldsymbol{x}_0)$. We denote the score of $f_t(\boldsymbol{x};\boldsymbol{x}_0)$ at $\boldsymbol{X}$ by $S(f_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{X})$ [7, 9, 55, 56]. While $\boldsymbol{X}$ denotes an observed value of the variable, $\boldsymbol{x}$ is any possible value in the range of $\boldsymbol{X}$. We need to find the score over multiple values and we do so by computing the expectation

$$\mathbb{E}[S(f_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{X})] = \int S(f_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{z})p_t(\boldsymbol{z};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{z}, \tag{5.8}$$

where the $\boldsymbol{x}$ is used for notation only and does not imply that the expectation is a function of $\boldsymbol{x}$. It has been argued that the scores to pick in (5.14) are *proper skill scores* [7, 9, 33]. *Skill* is a property of the forecasts that S should be capable of measuring and *propriety* (being *proper*) is a property of the score. The score S is proper if for any two probability densities $f_t(\boldsymbol{x};\boldsymbol{x}_0)$ and $p_t(\boldsymbol{x};\boldsymbol{x}_0)$,

$$\int S(f(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{z})p_t(\boldsymbol{z};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{z} \geq \int S(p_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{z})p_t(\boldsymbol{z};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{z}. \tag{5.9}$$

Hence the minimum of the left is obtained if $f_t(\boldsymbol{x};\boldsymbol{x}_0) = p_t(\boldsymbol{x};\boldsymbol{x}_0)$. The score is strictly proper if equality occurs in (5.9) only if $f_t(\boldsymbol{x};\boldsymbol{x}_0) = p_t(\boldsymbol{x};\boldsymbol{x}_0)$. The point of (5.9) is that a proper score gives a lower score to a more accurate forecast. Thus, for any two forecasts, $f_t(\boldsymbol{x};\boldsymbol{x}_0)$ is more accurate than $g_t(\boldsymbol{x};\boldsymbol{x}_0)$ if,

$$\int S(f_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{z})p_t(\boldsymbol{z};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{z} \leq \int S(g_t(\boldsymbol{x},\boldsymbol{x}_0),\boldsymbol{z})p_t(\boldsymbol{z};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{z}, \tag{5.10}$$

The term on the left-hand-side of equation (5.10) is *expected forecast skill* of $f_t(\boldsymbol{x};\boldsymbol{x}_0)$ and the one on the right-hand-side is the expected forecast skill of $g_t(\boldsymbol{x};\boldsymbol{x}_0)$. It says that $f_t(\boldsymbol{x};\boldsymbol{x}_0)$ is more skillful than $g_t(\boldsymbol{x};\boldsymbol{x}_0)$. Let us define *skill*. We can write

$$\begin{aligned}\int S(f_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{z})p_t(\boldsymbol{z};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{z} &= \int S(f_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{z})f_t(\boldsymbol{z};\boldsymbol{x}_0)\mathrm{d}\boldsymbol{z} \\ &+ \int S(f_t(\boldsymbol{x};\boldsymbol{x}_0),\boldsymbol{z})\left[p_t(\boldsymbol{z};\boldsymbol{x}_0)-f_t(\boldsymbol{z};\boldsymbol{x}_0)\right]\mathrm{d}\boldsymbol{z}.\end{aligned} \tag{5.11}$$

The first term after equality is the *sharpness* term and the second term is the *reliability* term. It is desirable that the sharpness term be as small as possible. We want forecasts to make the reliability term as close to zero as possible and the sharpness term as negative as possible. Although we agree with the persuasions of Gneiting et al. [18] that effort has to be spent on measuring sharpness, ultimate forecast calibration can be obtained by the use of a skill score because it measures both reliability and sharpness.

The performance of a model cannot be evaluated based on a single forecast [66]. This is because different models will out-perform each other at different times. We will need a sequence of forecast probability density functions, $\{f_t(\boldsymbol{x};\boldsymbol{x}_\tau)\}_{\tau\geq 0}$, with corresponding perfect forecasts $p_t(\boldsymbol{x};\boldsymbol{x}_\tau)$, where $\boldsymbol{x}_\tau$ is the initial condition from which each ensemble forecast is made and t is the lead time. Then, at lead time t, the overall forecast skill on the attractor is

$$\mathbb{E}[S(t)] = \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}[S(f_t(\boldsymbol{x};\boldsymbol{x}_\tau),\boldsymbol{X}^{(\tau)})]\mathrm{d}\tau, \tag{5.12}$$

provided the limit exists. $\boldsymbol{X}^{(\tau)}$ is the random variable being forecast from the initial ensemble corresponding to $\boldsymbol{x}_\tau$. If the underlying system is ergodic, we can write

$$\mathbb{E}[S(t)] = \lim_{T\to\infty}\frac{1}{T}\int_0^T S(f_t(\boldsymbol{x};\boldsymbol{x}_\tau),\boldsymbol{X}^{(\tau)})\mathrm{d}\tau. \tag{5.13}$$

Operationally, we never have the perfect forecast $p_t(\boldsymbol{x}; \boldsymbol{x}_0)$ which we could use to score the forecast $f_t(\boldsymbol{x}; \boldsymbol{x}_0)$. In fact, if we had the perfect forecast there would be no point in using $f_t(\boldsymbol{x}; \boldsymbol{x}_0)$. For each forecast, the underlying system can only furnish us with one verification $\boldsymbol{X}$. Therefore, it will be best to use (5.13) to score forecasts rather than (5.12). We can discretise time according to $\tau_i = (i-1)\tau_s$, for $i = 1, 2, .., N$ and τ_s is the sampling time. This gives us a sequence of forecast PDFs, $\{f_t(\boldsymbol{x}; \boldsymbol{x}_i)\}_{i=1}^N$, corresponding verifications $\{\boldsymbol{X}^{(i)}\}_{i=1}^N$ and some score S. The following empirical score is then used to value the t-ahead forecast system [9]:

$$\langle S \rangle(t) = \frac{1}{N} \sum_{i=1}^{N} S(f_t(\boldsymbol{x}; \boldsymbol{x}_i), \boldsymbol{X}^{(i)}). \tag{5.14}$$

which is an approximation of (5.13).

In the next section, we shall look at *Ignorance*, the only skill score that we shall employ.

5.3 Ignorance

We will here introduce the skill score proposed by Roulston and Smith [55]. It was motivated by ideas from information theory, for which we are greatly indebted to Shannon [60], and they called it the information-based Ignorance. Before introducing Ignorance, we shall first discuss the concept of *entropy*. Consider a random variable $\boldsymbol{X}$ that can take discrete values $\{\boldsymbol{x}_i\}_{i=1}^n$ with probabilities, $p_i = P(\boldsymbol{X} = \boldsymbol{x}_i)$, such that

$$\sum_{i=1}^{n} p_i = 1. \tag{5.15}$$

Entropy, $H(\boldsymbol{X})$, may be thought of as the amount of *information* given us (or *uncertainty* removed) by the realisation of $\boldsymbol{X}$ [31, 24]. The entropy depends only on the

probabilities, $\{p_i\}_{i=1}^n$. Hence, we can write

$$H(\boldsymbol{X}) = H(p_1, p_2, \cdots, p_n). \tag{5.16}$$

The entropy is required to have two basic properties. First, it has to assume its maximum when $p_i = \frac{1}{n}$. Also, if $\boldsymbol{Y}$ is another random variable that assumes discrete values, then H needs to satisfy the relation

$$H(\boldsymbol{X}, \boldsymbol{Y}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}|\boldsymbol{X}), \tag{5.17}$$

where $H(\boldsymbol{X}, \boldsymbol{Y})$ is the amount of information gained by the observation of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and $H(\boldsymbol{Y}|\boldsymbol{X})$ is the mathematical expectation of the amount of additional information gained by observing $\boldsymbol{Y}$ after the realisation of $\boldsymbol{X}$ [31]. In other words,

$$H(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{i=1} p_i H_i(\boldsymbol{Y}), \tag{5.18}$$

where $H_i(\boldsymbol{Y})$ is the entropy of $\boldsymbol{Y}$ given that $\boldsymbol{X} = \boldsymbol{x}_i$. $H(\boldsymbol{X}, \boldsymbol{Y})$ is also called *mutual information* [14, 30]. To add the third property, suppose $\boldsymbol{Y}$ can take the values $\{\boldsymbol{x}_i\}_{i=1}^{n+1}$ with $p_i = P(\boldsymbol{Y} = \boldsymbol{x}_i)$ and $p_{n+1} = 0$. It is then desirable that

$$H(p_1, p_2, \cdots, p_n, 0) = H(p_1, p_2, \cdots, p_n). \tag{5.19}$$

If for any n, H is a continuous function with respect to all its arguments and has the properties (5.17), (5.19) and maximisation by $p_i = 1/n$, it can be shown [31] that then

$$H(\boldsymbol{X}) = H(p_1, p_2, \cdots, p_n) = -\lambda \sum_{i=1}^{n} p_i \log p_i, \tag{5.20}$$

where $\lambda > 0$ is a constant. If the logarithm is base 2 then the corresponding units are called *bits*, and if base-e the units are called *nats*. Henceforth we shall fix $\lambda = 1$. To prove that the entropy as defined in (5.20) assumes its maximum when the points $\boldsymbol{x}_i$

have equal probabilities, $1/n$, we apply Jensen's inequality[6]. Consider a continuous convex function $\upsilon(t)$. Then

$$\upsilon\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) \leq \frac{1}{n}\left(\sum_{i=1}^{n} \upsilon(a_i)\right)$$

where a_i are any positive numbers. Setting $a_i = p_i$ and $\upsilon(t) = t\log t$, and using (5.15) yields

$$\upsilon\left(\frac{1}{n}\right) \leq \frac{1}{n}\sum_{i=1}^{n} p_i \log p_i = -\frac{1}{n}H\left(p_1, p_2, \cdots, p_n\right)$$

which implies that

$$H(\boldsymbol{X}) = H(p_1, p_2, \cdots, p_n) \leq \log n = H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right), \tag{5.21}$$

with equality if and only if $p_i = 1/n$. The reader is referred to [31] for proofs of the other properties, which are more straight forward. Entropy expresses average information where each event corresponding to i has information content [57, 60, 74],

$$I_i = -\log p_i. \tag{5.22}$$

Since the p_i's are less than 1, each I_i is positive. We define $p_i \log p_i = 0$ if $p_i = 0$. Based on this idea of information, Roulston and Smith [55] defined

$$\text{ign}(f_t; \boldsymbol{x}) = -\log f_t(\boldsymbol{x}; \boldsymbol{x}_0), \tag{5.23}$$

which is the information deficit or *Ignorance* that a forecaster in possession of the forecast PDF $f_t(\boldsymbol{x}; \boldsymbol{x}_0)$ has before the observation $\boldsymbol{x}$ is communicated to him [55]. Unlike the expression in (5.22), Ignorance may assume negative values when the value of $f_t(\boldsymbol{x}; \boldsymbol{x_0})$ is greater than 1. This is the score that shall be used to evaluate forecasts, so that

$$S(f_t, \boldsymbol{x}) = \text{ign}(f_t, \boldsymbol{x}). \tag{5.24}$$

[6]Consider a function $\phi(x)$ that is continuous and convex. Then Jensen's inequality is given by $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$ [24, 74].

If the corresponding perfect forecast is $p_t(\boldsymbol{x}; \boldsymbol{x}_0)$, then the expected Ignorance is

$$\mathbb{E}[\text{ign}(f_t(\boldsymbol{x}; \boldsymbol{x}_0), \boldsymbol{X})] = -\int p_t(\boldsymbol{z}; \boldsymbol{x}_0) \log f_t(\boldsymbol{z}; \boldsymbol{x}_0) \mathrm{d}\boldsymbol{z}. \tag{5.25}$$

It follows from the positivity of the Kullback-Leiber inequality that Ignorance is proper. The expected Ignorance is bounded below by the *density entropy* [24] of the perfect forecast, which is given by

$$H(X) = -\int p_t(\boldsymbol{z}; \boldsymbol{x}_0) \log p_t(\boldsymbol{z}; \boldsymbol{x}_0) \mathrm{d}\boldsymbol{z}. \tag{5.26}$$

In section 5.2 and equation (5.14), we introduced *empirical skill* as the quantity of computational interest. Likewise, in our computations, we will need *empirical mean Ignorance*, which is defined as,

$$\langle \text{ign} \rangle(t) = \frac{1}{N} \sum_{i=1}^{N} \text{ign}(f_t(\boldsymbol{x}; \boldsymbol{x}_i), \boldsymbol{X}^{(i)}), \tag{5.27}$$

where $\boldsymbol{x}_i$ is the point on the time series from which an initial ensemble to be forecast is generated and $\boldsymbol{X}^{(i)}$ is the verification. In the next section, we discuss the way to construct continuous PDFs from a finite number of points and explain the employment of Ignorance in this.

5.4 Dressing

In the preceding discussions, we found that, to score the performance of a model, we need the forecast probability density function. In practice, the forecasts we get comprise ensembles of discrete points. Given an ensemble of points, $\{X_i\}_{i=1}^{n}$, we assume that they are an independent and identically distributed random draw from some underlying forecast distribution, which we assume to be continuous. How can we use the ensemble to estimate the underlying forecast PDF? Wilks [76] compared

different ways of estimating the forecast PDF and found that continuous PDF estimation techniques out-perform discrete estimates. In particular, he found that the PDF fitting technique proposed by Roulston and Smith [56], which takes into account the historical forecast errors, is among those that provide the best results. Fitting PDFs to ensembles by taking into account historical errors is called *dressing* the ensembles [56]. The process of dressing involves putting *kernels* on each of the ensemble members. In this section, we shall address why, given the choice between estimating forecast PDFs using either histograms or kernels, we settle for the latter. At the same time, we shall explain how to go about doing the dressing. Our attention shall be restricted to uni-variate data.

Suppose an ensemble of points $\{X_i\}_{i=1}^n$ was drawn from some underlying forecast distribution with PDF $f(x)$. Let $\hat{f}_h(x)$ be a histogram estimate of the PDF. To follow Scott [59], we consider the histogram to be defined on an equally spaced mesh $\{x_{ni:-\infty<i<\infty}\}$, so that $h_n = x_{n(i+1)} - x_{ni}$. The mean integrated squared error is then given by

$$MISE_h = \int E[\hat{f}_h(x) - f(x)]^2 dx. \tag{5.28}$$

This error depends on the bin width h_n and the relative position of the mesh. By considering the bias and variance at some interval containing a fixed x and employing (5.28), it can be shown that [59]

$$MISE_h = \frac{1}{nh_n} + \frac{1}{12} h_n^2 \int f'(x)^2 dx + O(1/n + h_n^3). \tag{5.29}$$

The optimum bin width is the minimiser of (5.29) and it is

$$h_n^* = \left\{ \frac{6}{\Psi(f)} \right\}^{1/3} n^{-1/3}, \tag{5.30}$$

where $\Psi(f) = \int f'(x)^2 dx$, so that the minimum possible error over h_n is

$$MISE_h^* = \frac{3}{2}\left\{\frac{\Psi(f)}{6}\right\}^{1/3} n^{-2/3} + O(1/n + h_n^3). \tag{5.31}$$

It is clear that $MISE_h \to 0$ as $n \to \infty$ provided

$$(nh_n)_{n\to\infty} \to \infty. \tag{5.32}$$

Alternatively, we could use *Kernels* to estimate $f(x)$. This means using linear combinations [62, 23]

$$\hat{f}_k(x) = \frac{1}{n}\sum_{i=1}^{n} w(x, X_i), \tag{5.33}$$

where $w(x, X_i)$ are the weighting functions placed on each of the ensemble members and satisfying certain properties. In particular, we can choose

$$w(x, X_i) = \frac{1}{\sigma_n} K(\frac{x - X_i}{\sigma_n}). \tag{5.34}$$

PDF estimates based on these sorts of weighting functions are called Parzen density estimators [52]. σ_n is called the smoothing parameter [7]. It can be likened to the bin-width of the histogram [62]. If the weighting functions $K(t)$ satisfy the following properties [62]

$$K(t) \geq 0, \tag{5.35}$$

$$\int K(t)dt = 1 \tag{5.36}$$

$$\int tK(t)dt = 0, \tag{5.37}$$

$$\int t^2 K(t)dt = k_2 \neq 0, \tag{5.38}$$

and are symmetric [8], they are called Kernels. Then the PDF estimate is asymptotically unbiased [52], meaning that if

$$\lim_{n\to\infty} \sigma_n = 0 \tag{5.39}$$

[7] The smoothing parameter is also called window width or bandwidth [62].

[8] This condition can be relaxed since it is merely intended for ease of exposition.

then

$$\lim_{n\to\infty} E[\hat{f}_k(x)] = f(x). \tag{5.40}$$

The PDF estimate is also consistent [9] [52] in the sense of mean squared error if in addition to (5.39),

$$\lim_{n\to\infty} n\sigma_n = \infty. \tag{5.41}$$

Consistency follows from (5.40) and the fact that (5.41) implies that [52]

$$\lim_{n\to\infty} \mathrm{Var}[\hat{f}_k(x)] = 0. \tag{5.42}$$

Note that the mean integrated squared error is given by [52, 62]

$$\mathrm{MISE}k = \int \left\{ \mathrm{Var}[\hat{f}_k(x)] + (E[\hat{f}_k(x)] - f(x))^2 \right\} dx. \tag{5.43}$$

Application of Taylor series expansions to f, with the assumption that σ_n is small, yields the approximation [62]

$$\mathrm{MISE}k = \frac{1}{4}\sigma_n^4 k_2^2 \int f''(x)^2 dx + \frac{1}{n\sigma_n} \int K(t)^2 dt, \tag{5.44}$$

which is minimised by [62]

$$\sigma_n^* = k_2^{-2/5} \left\{ \int K(t)^2 dt \right\}^{1/5} \left\{ \int f''(x)^2 dx \right\}^{-1/5} n^{-1/5}. \tag{5.45}$$

The minimum mean integrated squared error is obtained by substituting (5.45) into (5.44) to obtain [62]

$$\mathrm{MISE}k^* = \frac{5}{4} C(K)\Phi(f) n^{-4/5}, \tag{5.46}$$

where

$$C(K) = k_2^{2/5} \left\{ \int K(t)^2 dt \right\}^{4/5}, \quad \text{and} \quad \Phi(f) = \left\{ \int f''(x)^2 dx \right\}^{1/5}. \tag{5.47}$$

[9] The estimate is consistent if $E[\hat{f}_k(x) - f(x)]^2 \to 0$ as $n \to \infty$ [62].

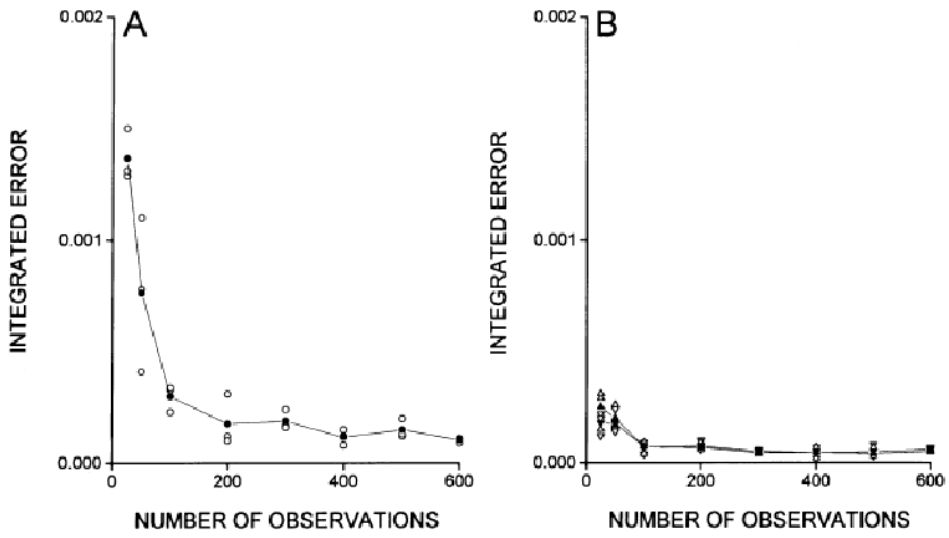


Figure 5.4: Integrated mean square error as a function of n (number of observations) using A) histogram estimates and B) kernel estimates. The circles (or upward and downward triangles) correspond to three random number selections for each n. For downward triangles, Gaussian kernels were used and rectangular kernels were used for upward triangles. The kernels were used with optimum widths. From Glanovic [15].

The dependence of the error on the choice of kernel is not as significant as one might expect. This has been substantiated in [62] by computing the efficiencies[10] of various kernels with respect to the *Epanechnikov kernel*, defined by

$$K_e = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right), & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0, & \text{otherwise.} \end{cases} \tag{5.48}$$

Let us now get back to the question of whether to use the comparatively cheap histogram estimates or to use kernel estimates. Scott [59] discredited kernel estimates for the higher order of accuracy. However, Glavinovic [15] demonstrated that histogram estimates perform relatively very poorly especially at low values of n (see figure 5.4, for example) by considering a beta distribution with parameters 10 and 3 respectively. Scott used the fact that if the optimum window width of the histogram and kernel estimates changes by a factor c, the integrated mean squared errors change by factors

[10]Efficiency was defined as $\text{eff}(K) = \{C(K_e)/C(K)\}^{5/4}$.

$(c^3+2)/3c$ and $(c^5+4)/5c$ respectively and

$$(c^3+2)/3c \leq (c^5+4)/5c,$$

provided $c > 0$. However, comparing the actual values of the integrated mean squared errors may give a different picture since then the n dependence is brought into play. In particular, if we consider standard normal kernels and underlying PDF, then the errors are

$$\text{MISE}^*(n) = 0.42970n^{-2/3}, \quad \text{and} \quad \text{MISE}k^*(n) = 0.33287n^{-4/5}$$

In weather forecasting centres, computational power places limitations on the ensemble size, which immediately makes kernel estimates the most desirable.

Given a matrix of ensembles, we could naively make PDF estimates for each ensemble by using the foregoing procedure. However, such estimates would be of limited value [55]. In particular, the forecasts may lack *skill*, which we defined in section 5.2 as a decomposition into sharpness and reliability. Roulston and Smith [55] argued that this pitfall can be remedied by dressing each of the ensemble members with the distribution of the historical errors of the best member forecasts. The error distribution so obtained could then be used as the uncertainty associated with each ensemble member. In this thesis, we do something slightly different. Instead of computing the distribution of historical errors, we estimate them by kernels. A kernel with optimum parameters would then correspond to the distribution of historical errors of best member forecasts [8, 56]. The problem is then reduced to estimating the parameters of the chosen kernel. Optimum parameters would then be those that minimise some skill score. We explain the procedure more carefully in sequel.

Suppose the set of our dth ensemble is $\{z_i^{(d)}\}_{i=1}^{N_e}$ and z_d is the corresponding verification. Then, the form of the Kernels we choose is $\frac{1}{\sigma}K\left(\frac{x-\mu}{\sigma}\right)$ [8]. We then choose σ and μ so as to minimise Ignorance at z_d. Define

$$\rho_{\sigma,\mu}^{(d)}(z_d) = \frac{1}{N_e}\sum_{i=1}^{N_e}\frac{1}{\sigma}K\left(\frac{z_d - z_i^{(d)} - \mu}{\sigma}\right), \tag{5.49}$$

where z_d is the verification of the dth ensemble. Now let us suppose that we have D ensembles and verifications. For each of these we denote the corresponding probability density function by $\rho_\sigma, \mu^{(d)}$. The average Ignorance is then given by

$$\langle \text{ign} \rangle(\sigma, \mu) = \frac{1}{D}\sum_{d=1}^{D} -\log \rho_{\sigma,\mu}^{(d)}(z_d). \tag{5.50}$$

We then wish to find σ and μ that minimise the average Ignorance. There are different minimisation algorithms that can be used for this such as the Newton-Raphson method[11], secant method, to mention a few [53]. Matlab also has a variety of functions such as *fminsearch.* The function fminsearch uses the Nelder-Mead simplex search method [53]. To perform the search, we need the additional constraint that $\sigma > 0$ since the spread cannot be negative. For scalar data, we choose

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right). \tag{5.51}$$

It is easy to spot that

$$\int_{-\infty}^{\infty}\rho_{\sigma,\mu}^{(d)}(z)dz = 1 \tag{5.52}$$

since $\int_{-\infty}^{\infty}\frac{1}{\sigma}K(\frac{z-z_i^{(d)}-\mu}{\sigma})dz = 1$ for $i = 1, \ldots, N_e$. Also $\rho_{\sigma,\mu}(z) \geq 0$. Thus $\rho_{\sigma,\mu}(z)$ is a probability density function. In the next section, we shall explain how climatology can be used to safeguard us against possible pitfalls in the calculation of optimum parameters.

[11] To apply this method, one would have to first find the derivative and then find its zero.

5.5 Climatology

In the preceding formulation for calculating optimum dressing parameters, we have ignored the fact that a few poor ensembles can result in a big bandwidth being chosen, so that the dressed ensemble PDFs would not be sharp. By poor ensembles we mean those that are clustered away from their respective verifications. This can happen because we choose one spread, σ, for all the dressing kernels and if one of the ensembles is far away from the verification, σ is adjusted to be big enough to avoid assigning zero to the PDF [12] and this in turn causes us to lose skill. Large values of σ result in PDFs that are not sharp.

To surmount the foregoing problem, we use the *climatology* in the dressing of the ensembles. By *climatology* we mean the PDF of the underlying attractor, and this can be estimated from past observations. To obtain a climatology, we use historical data set, $\{y_i\}_{i=1}^{M}$, which may be the learning set used to build the model. An approximation of the climatology is then given by [8],

$$\rho_{cl}(y) = \frac{1}{\sigma_c M} \sum_{j=1}^{M} K\left(\frac{y - y_j - \mu_c}{\sigma_c}\right), \tag{5.53}$$

where K is a kernel, and σ_c is the climatological kernel spread and μ_c is the associated offset. We then need σ_c, μ_c so as to optimise the average Ignorance [8]

$$-\frac{1}{M} \sum_{i=1}^{M} \log\left[\frac{1}{\sigma_c (M-1)} \sum_{j=1, j\neq i}^{M} K\left(\frac{y_i - y_j - \mu_c}{\sigma_c}\right)\right]. \tag{5.54}$$

With σ_c and μ_c thus chosen, we then need to optimise the average Ignorance of [8]

$$\rho^{(d)}(z) = \alpha \rho^{(d)}_{\sigma,\mu}(z) + (1-\alpha)\rho_{cl}(z)$$

[12] Since $\lim_{\sigma \to 0^+} \text{ign} = \infty$

to choose σ and α, with $\alpha \in [0, 1]$. This is a constrained optimisation problem. The average Ignorance over D observations is now given by

$$\langle \text{ign} \rangle(\alpha, \sigma, \mu) = \frac{1}{D} \sum_{d=1}^{D} -\log \rho^{(d)}(z_d). \tag{5.55}$$

After fitting a PDF to an ensemble, it may be necessary to measure the sharpness of the resulting PDF. One way to do this is to compute the variance of the PDF [54, 76].

To compute the variance, first note that

$$\int_{-\infty}^{+\infty} z \rho_{\sigma,\mu}^{(d)}(z) dz = \frac{1}{N} \sum_{i=1}^{N} z_i^{(d)} + \mu = \bar{z}^{(d)} + \mu, \tag{5.56}$$

where $\bar{z}^{(d)} = \frac{1}{N_e} \sum_{i=1}^{N} z_i^{(d)}$ is the raw-ensemble mean. Hence the dressed-ensemble mean is equal to the raw-ensemble mean plus the offset parameter, μ. Now let $k(z; z_i^{(d)}) = \frac{1}{\sigma} K((z - z_i^{(d)} - \mu)/\sigma)$ so that[13] $\rho_{\sigma,\mu}^{(d)}(z) = \frac{1}{N_e} \sum_{j=1}^{N_e} k(z; z_i^{(d)})$. In this case,

$$\begin{aligned} &\int_{-\infty}^{\infty} \left(z - z_i^{(d)} - \mu \right)^2 k(z; z_i^{(d)}) dz = \sigma^2 \\ \Rightarrow &\int_{-\infty}^{\infty} z^2 k(z; z_i^{(d)}) dz = \sigma^2 + (z_i^{(d)} + \mu)^2. \end{aligned} \tag{5.57}$$

The variance of the dressed ensemble is then given by

$$\begin{aligned} \text{Var}[Z] &= \int_{-\infty}^{\infty} (z - \bar{z}^{(d)} - \mu)^2 \rho_{\sigma,\mu}^{(d)}(z) dz \\ &= \frac{1}{N_e} \sum_{i=1}^{N_e} \int_{-\infty}^{\infty} (z - \bar{z}^{(d)} - \mu)^2 k(z; z_i^{(d)}) dz, \\ &= \frac{1}{N_e} \sum_{i=1}^{N_e} \left[\int_{-\infty}^{\infty} z^2 k(z; z_i^{(d)}) dz - 2(\bar{z}^{(d)} + \mu)(z_i^{(d)} + \mu) + (\bar{z}^{(d)} + \mu)^2 \right] \end{aligned} \tag{5.58}$$

Substituting (5.57) into (5.58) yields

$$\int_{-\infty}^{\infty} (z - \bar{z}^{(d)} - \mu)^2 \rho_{\sigma,\mu}^{(d)}(z) dz = \sigma^2 + \frac{1}{N_e} \sum_{i=1}^{N_e} \left(z_i^{(d)} - \bar{z}^{(d)} \right)^2, \tag{5.59}$$

[13] In particular, we may use $k(z; z_i^{(d)}) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(z - z_i^{(d)} - \mu)^2}{2\sigma^2}}$

after a few algebraic manipulations.

Let us now turn to the case when $\alpha \neq 1$. If we denote the dressed ensemble mean by μ_d, then

$$\begin{aligned} \mu_d &= \int_{-\infty}^{\infty} z\rho^{(d)}(z)\mathrm{d}z \\ &= \alpha(\bar{z}^{(d)} + \mu) + (1-\alpha)(\bar{y} + \mu_c), \end{aligned} \tag{5.60}$$

where $\bar{y} = \frac{1}{M}\sum_{i=1}^{M} y_i$. The variance of the dressed ensemble, σ_d^2 is then given by,

$$\begin{aligned} \sigma_d^2 &= \alpha \int_{-\infty}^{\infty} z^2 \rho_{\sigma,\mu}^{(d)}(z)\mathrm{d}z + (1-\alpha)\int_{-\infty}^{\infty} z^2\rho_{cl}(z)\mathrm{d}z - \mu_d^2. \\ &= \alpha\left[\sigma^2 + \left(\bar{z}^{(d)} + \mu\right)^2 + \frac{1}{N_e}\sum_{i=1}^{N_e}\left(z_i^{(d)} - \bar{z}^{(d)}\right)^2\right] \\ &\quad +(1-\alpha)\left[\sigma_c^2 + (\bar{y} + \mu_c)^2 + \frac{1}{M}\sum_{i=1}^{M}(y_i - \bar{y})^2\right] - \mu_d^2. \end{aligned} \tag{5.61}$$

To illustrate the benefits of blending with climatology, we made computations on training data consisting of 1024 points sampled from SET7 every 256 points. Initial ensembles were then made by adding Gaussian perturbations of standard deviation $\epsilon = 5 \times 10^{-3}$. Ensemble predictions were then obtained by iterating these initial ensembles forward in time using model M_2 (see chapter 3). Dressed ensembles at lead 6.4ms lead time are shown in figure 5.5. The contrasts in sharpness may be attributed to the sizes of the respective bandwidths, which are $\sigma = 0.2195$ (with climatology shown in figure 5.6) and $\sigma = 0.4002$ (without climatology) at a lead time of 6.4 ms. Notice that in each case, the variance would diagnose the PDFs not containing the climatological blend as being sharper, contrary to the diagnosis of entropy. The entropy diagnosis supports visual evidence. Of course, things are not always what they appear to be. There are cases where the variance and entropy are in agreement, such

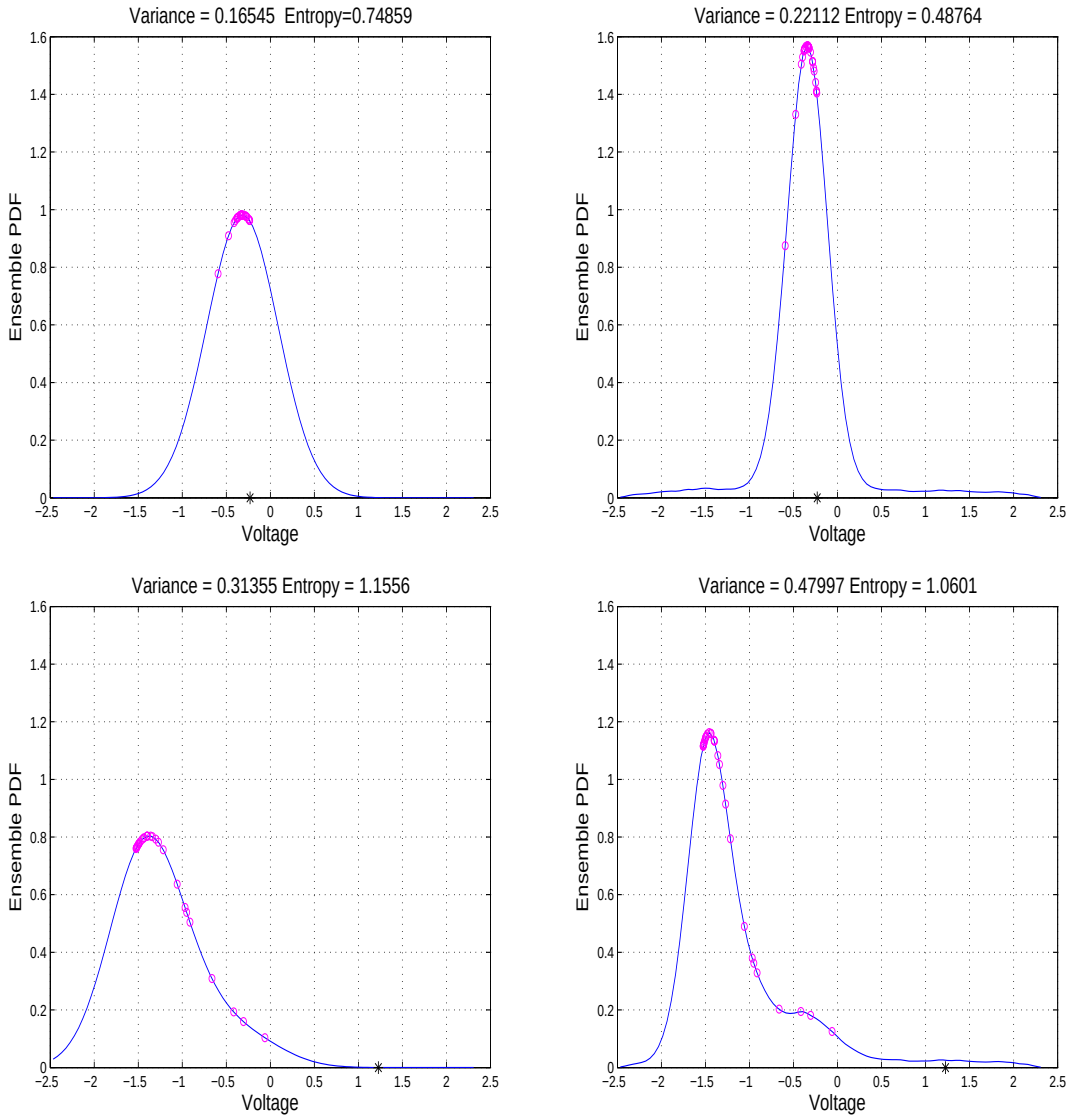


Figure 5.5: PDFs obtained by dressing the ensemble forecasts (At lead lead time of 6.4ms) of the circuit without (left) and with (right) the climatology. The magenta circles are the forecasts and the black asterisk is the verification. Clearly the right distributions are sharper than the left ones and this fact is reflected by the respective entropies, yet the variances of the left ones are lower.

as the one indicated in figure 5.7.

To measure the effects of using the climatology on the skill of the PDF dressings, we scored the performance of the PDFs on the training and testing data. The graphs of average Ignorance versus lead lead time are shown in figure 5.8 for the two

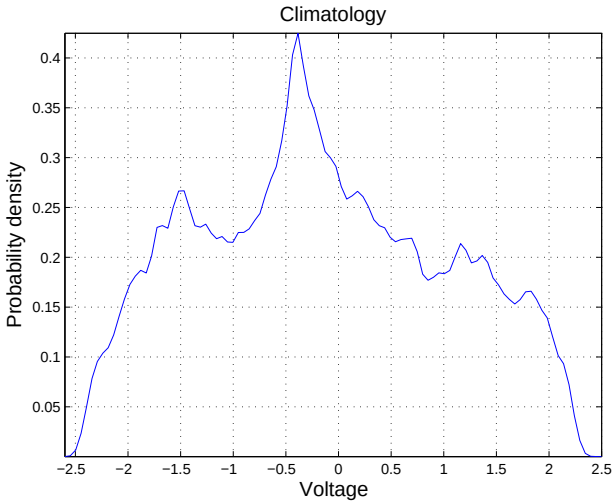


Figure 5.6: Shows the climatology of the circuit made from 8192 data points and then used to compute the PDFs of figures 5.5-5.7.

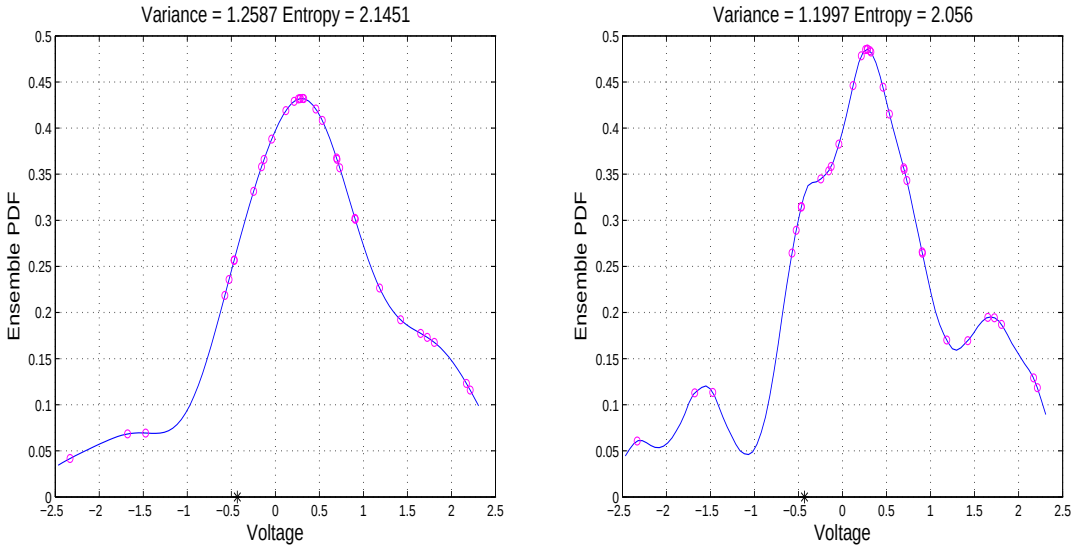


Figure 5.7: PDFs obtained by dressing the ensemble forecasts (At lead time of 6.4ms) of the circuit without (left) and with (right) climatology. The magenta circles are the forecasts and the black asterisk is the verification. In this case the diagnosis of both variance and entropy are in agreement.

cases when the ensembles are dressed with and without climatology. Both dressing methods provide PDFs of relatively equal skill up to lead times of about 2.5 ms on the training data. After that lead time, PDFs containing the climatological blend provide superior skill. On the testing data, the PDFs with climatological blend provide better

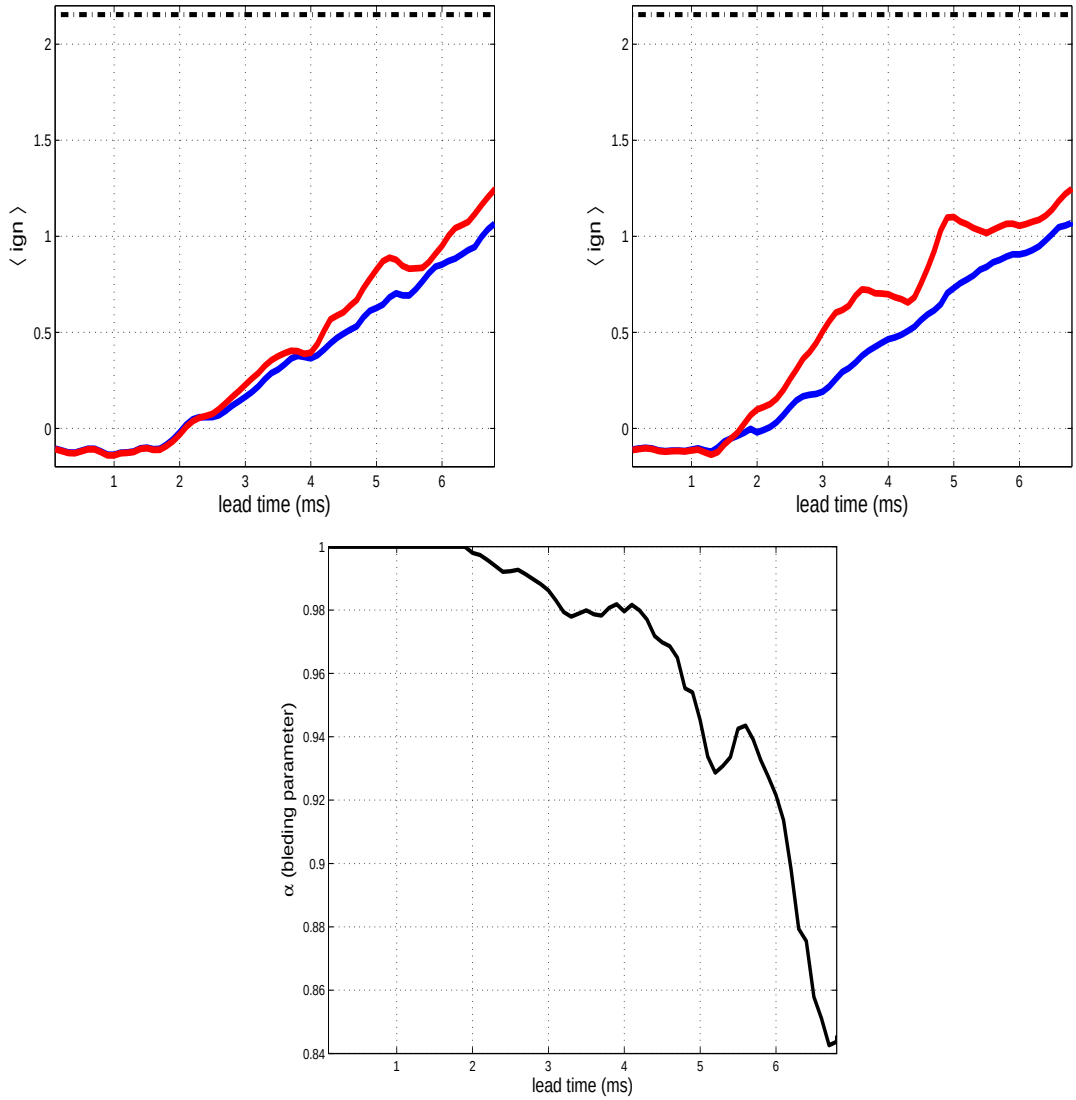


Figure 5.8: Graphs of average Ignorance in the training (top left) and testing (top right) period. The lines graphs correspond to dressing with climatology (blue), without climatology (red) and black dashed line is Ignorance of the climatology. On the bottom is the graph of the blending parameter versus lead time.

skill at an earlier lead time about of about 2 ms. The graph of α versus lead time indicates that the difference in skill takes effect when $\alpha < 1$ (see the bottom figure in figure 5.8). The results indicate that blending with climatology provides more skillful PDFs at higher lead times.

In the next section, we employ the Ignorance skill score and kernel dressing to determine the size of perturbations necessary for optimum predictions. This will be done in both the PMS and IMS.

5.6 Initial-Perturbation Spread

Given a set of observations from which we want to make future predictions with some model, what should be the standard deviation of the initial perturbations? This question has to be answered before predictions are made and we can not settle for any perturbation level because if the standard deviation is too big, we may as well just use the climatology. On the other hand, if it is too small, our forecasts may be unreasonably over confident. Therefore, the desirable standard deviation of the perturbations is one that gives us the best skill. In the perfect model scenario, when the only limit to predictability is due to uncertainty in the initial conditions, we might guess that the optimum perturbation level should be equal to the noise level. But if our model is imperfect, even if there is no observational noise, we still need initial perturbations to mitigate model error. We shall consider the cases (i) Perfect model scenario and (ii) imperfect model scenario, in both the noise free and noisy data cases. Our noise and perturbations are drawn from Gaussian distributions with standard deviations δ and ϵ respectively. This means $\delta = 0$ will correspond to clean data. We expect an optimum perturbation level to be one that minimises the Ignorance skill score.

This section is organised as follows: In § 5.6.1, we explore how Ignorance changes with the standard deviation of the perturbations in the PMS. This is done by looking at MS data. In § 5.6.2, we shift our attention to the IMS by first considering MS

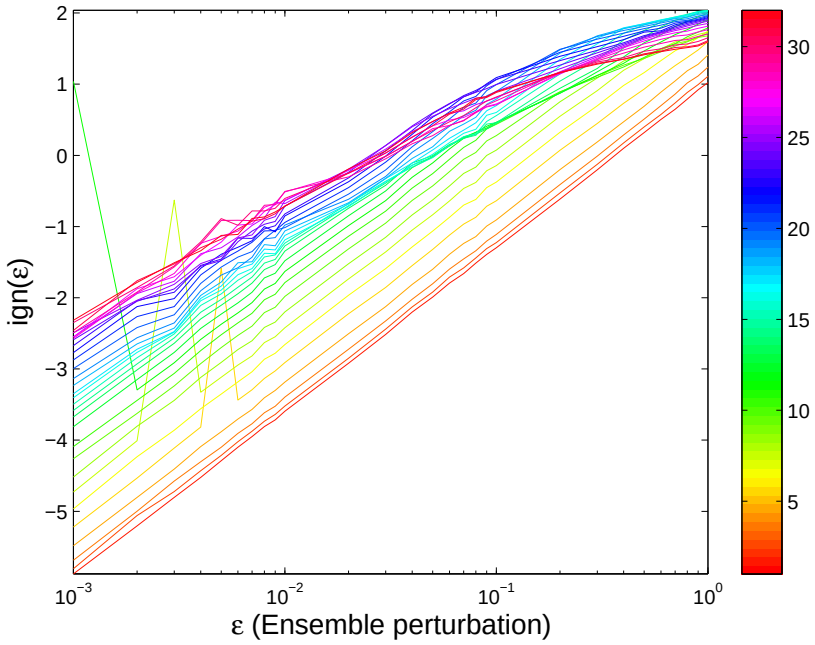


Figure 5.9: The graphs of average Ignorance versus ensemble perturbation with 512 ensembles for various lead times (according to the right colour bar), each ensemble containing 32 members. The MS data was noise free and we used a perfect M-S model. The perturbations were Gaussian with standard deviation ε.

data and then circuit data. All imperfect models employed are radial basis function models.

5.6.1 The Perfect Model Scenario (PMS)

Here we shall consider MS data [14] at parameter values $\gamma = 36$ and $\Gamma = 100$. The model in question shall be the corresponding MS system as defined in equation (2.10). For a given observational noise level, we shall vary ϵ logarithmically between 10^{-3} and 1. For $\delta = 0$, we have shown the graphs of average Ignorance versus ϵ in figure 5.9. The different colour lines correspond to different lead times (time ahead at which the forecast is made) up to 32 time steps. Notice that the graph generally yields straight lines except at higher lead times and perturbation levels. Indeed as the perturbation

[14] MS data is the data obtained by numerical integrations of the MS system after ignoring transients.

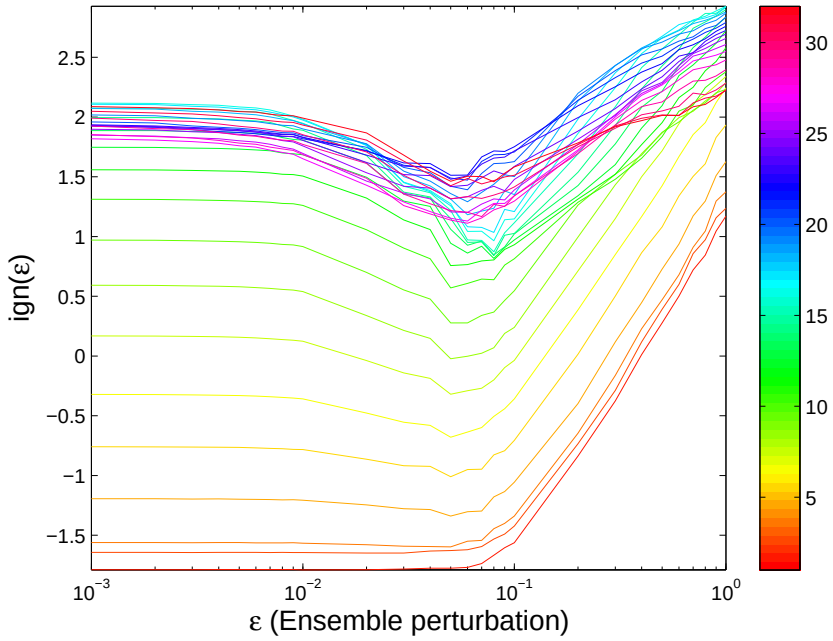


Figure 5.10: The graph of ignorance versus ensemble perturbation with 512 ensembles for various lead times (according to the right colour bar), each ensemble containing 32 members. The MS data was corrupted with observational noise of standard deviation 5×10^{-2} and we used perfect M-S model.

level increases, we would expect that the ensembles at low lead times to be approximately flattened Gaussians. Lower values of average Ignorance for the high lead times at larger values of ϵ reflect return of skill. What we can draw from this is that linear graphs like those in figure 5.9 would suggest that the underlying model is perfect and the data is noise free.

Next, we considered the case when the data was corrupted with additive noise of standard deviation, $\delta = 5 \times 10^{-2}$. The corresponding graphs of average Ignorance versus ϵ is shown in figure 5.10. At low perturbation levels, all the graphs are almost flat since the perturbations are drowned by the noise level. However, as the perturbation level increases, the higher lead time ensembles begin to fall. This is because

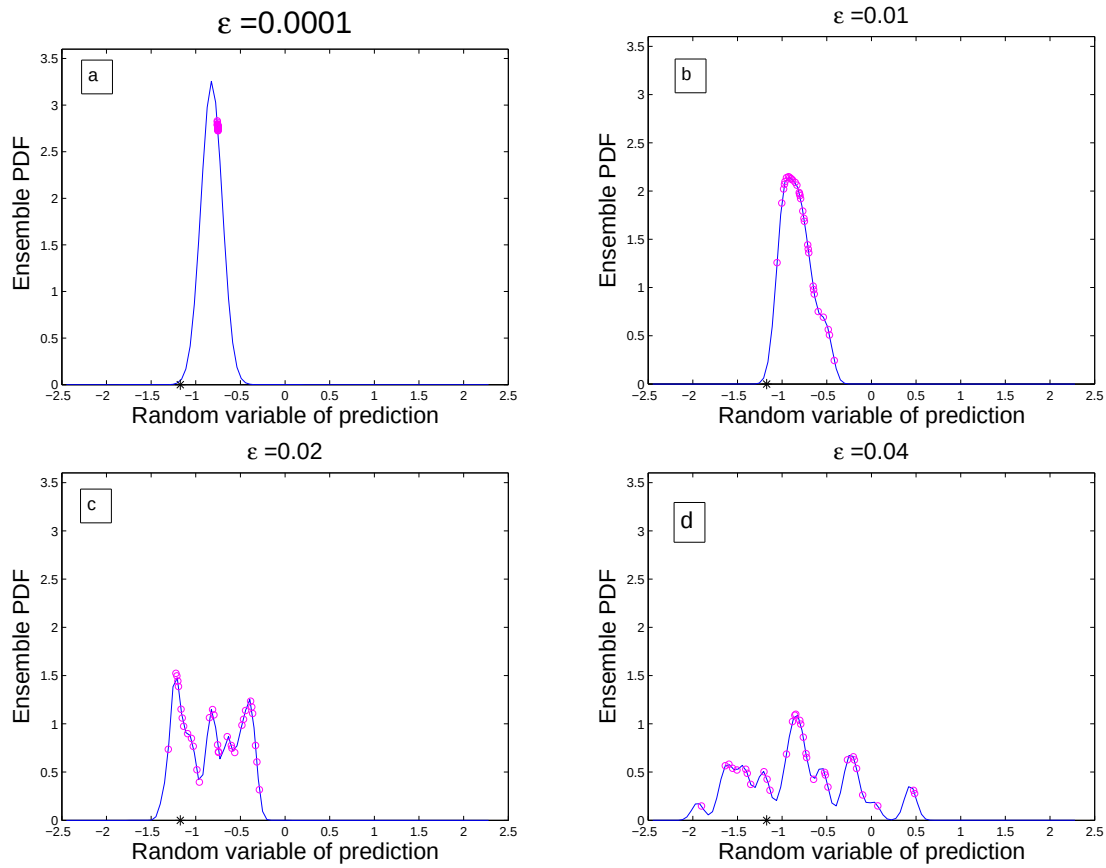


Figure 5.11: A series of ensembles for at lead time 16 of the circuit obtained with an imperfect model. Notice that as ϵ increases, the value of $\rho(x)$ increases and then decreases, where x is indicated by the asterisk (verification). The corresponding graph of ignorance versus ϵ is shown in figure 5.15.

the ensembles at higher lead times spread out, and in the process, the verifications which were initially far at the tails of the distribution tend to be encapsulated by the ensembles as we gain skill (see figure 5.11). At low lead times, the verifications are generally at the centre of the ensembles. However, as the distribution spreads out and flattens, Ignorance increases. That is why at low lead times, the graphs are initially flat and then begin to increase linearly. The value of ϵ where this happens is $\epsilon \approx \delta$. It is the same value at which graphs of the higher lead times attain their minima.

To explain the foregoing observations further, let us consider two PDFs of the per-

fect forecast and the imperfect forecasts, $p_t(x;\sigma_p(\epsilon),\mu_p)$ and $f_t(x;\sigma_f(\epsilon),\mu_f)$, where σ_p (rep. σ_f) and μ_p (resp. μ_f) are the standard deviation and mean respectively. Suppose our forecast, f_t, is Gaussian, meaning that

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

Then the expected skill of f_t is

$$\begin{aligned}\mathbb{E}[\text{ign}(f_t,X)] &= -\int_{-\infty}^{\infty} p_t(x;\sigma_p^2,\mu_p)\log f_t(x;\sigma_f^2,\mu_f)\mathrm{d}x \\ &= \frac{1}{2}\log(2\pi\sigma_f^2) + \frac{\sigma_p^2}{2\sigma_f^2} + \frac{1}{2\sigma_f^2}(\mu_p-\mu_f)^2. \end{aligned} \tag{5.62}$$

We assume that the standard deviations, σ_p and σ_f, are monotonic increasing functions of ϵ. If $\sigma_p = \sigma_f$ then (5.62) reduces to

$$\mathbb{E}[\text{ign}(f_t,X)] = \frac{1}{2}\log(2\pi e\sigma_f^2) + \frac{1}{2\sigma_f^2}(\mu_p-\mu_f)^2 \tag{5.63}$$

and the expected skill is minimised by

$$\sigma_f = |\mu_p - \mu_f|. \tag{5.64}$$

If $\mu_p = \mu_f$, then

$$\mathbb{E}[\text{ign}(f_t,X)] = \frac{1}{2}\log(2\pi e\sigma_f^2), \tag{5.65}$$

which is a monotonic increasing function of σ_f. This may explain why we obtained straight line graphs in the noise free PMS. They are obtained when the perfect forecast ensemble and the imperfect forecast ensemble have equal means and variances.

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

$$\min_{\sigma_f>0} \mathbb{E}[\text{ign}(f_t,X)] = \frac{1}{2}\log\left[2\pi e^2(\mu_p-\mu_f)^2\right]. \tag{5.66}$$

In particular,

$$\min_{\epsilon>0} \mathbb{E}[\text{ign}(f_0, X)] = \frac{1}{2} \log \left[2\pi e^2 \xi_0^2\right], \tag{5.67}$$

where $\xi_0 = \mu_p - \mu_f \sim N(0, \delta^2)$. Here, μ_p is the mean of the initial ensemble containing the unperturbed initial condition, x_0, and lies on the attractor and μ_f is the mean of the dressed ensemble obtained by making Gaussian perturbations of variance ϵ^2 and mean 0. For a more general case, at lead time t, we define $\xi_t = \mu_p(t) - \mu_f(t)$. If $\epsilon > \xi_0$, then minimum in (5.67) will not be attained by increasing ϵ because it can only be attained when $\sigma_f = \epsilon = |\xi_0|$. However, over a window of time series, the average may be constant for a while as witnessed in figure 5.10. We assume that at 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 $\epsilon = \delta$.

5.6.2 The Imperfect Model Scenario (IMS)

We now carry over the ideas of the preceding subsection to the imperfect model scenario. We consider models of the form

$$x_n = \phi(\boldsymbol{x}_{n-1}) + \epsilon_n, \tag{5.68}$$

where ϵ_n is the dynamical noise, $\boldsymbol{x}_{n-1}$ is the delay vector, and ϕ is a function expressing the deterministic part of the model. The ϵ_n's are *iid*. The deterministic part of the models was built from noise free data using cubic radial basis functions. The dynamical noise is a term that may be used to take model error into account at each time step. It should not be thought to imply that the underlying system is stochastic. We draw it from a Gaussian distribution with zero mean and standard deviation equal to the standard deviation of the one-step-error of the corresponding deterministic model.

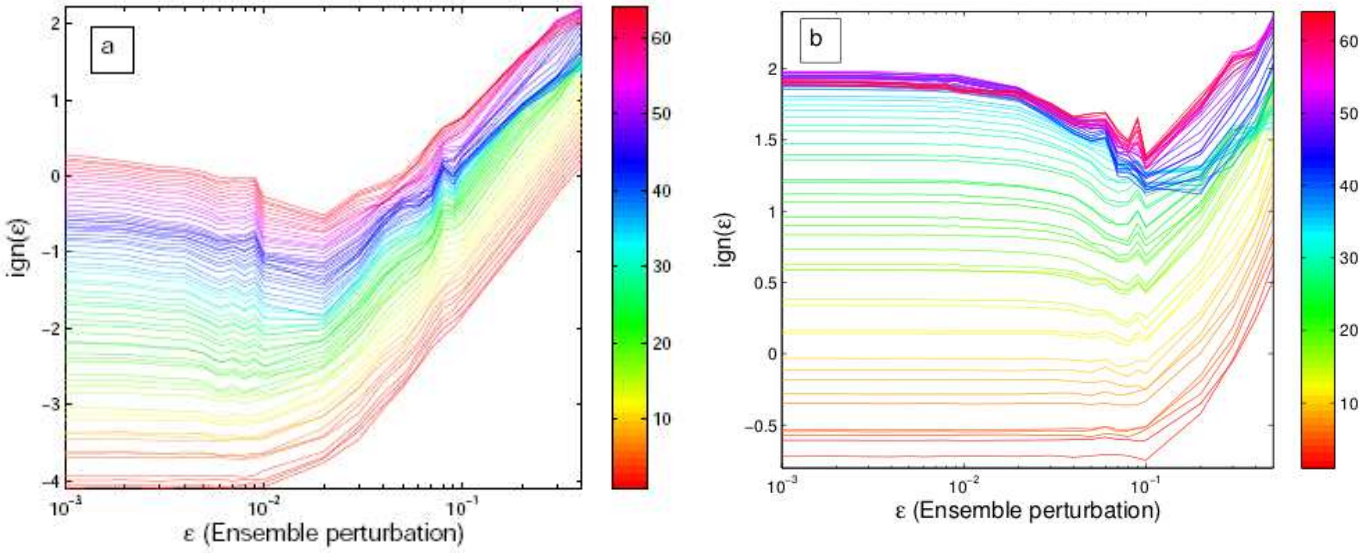


Figure 5.12: Graphs of average Ignorance versus logarithmically varying standard deviation, ϵ, of ensemble perturbations with initial observational error of standard deviation $\delta = 10^{-2}$ in (a) and 10^{-1} in (b) on MS data with an imperfect model. 128 initial conditions with a time step 64 between 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. 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.

Let us first consider MS-data with observational noise, $\delta = 10^{-2}, 10^{-1}$, and $\epsilon_n = 0$. The graphs of Ignorance versus ensemble perturbation for various lead times are shown in figure 5.12. We notice that the low lead time graphs begin to rise at $\epsilon \approx \delta$. At a slightly bigger value of ϵ, 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.

Perhaps a worry would be: What if the noise of the underlying system is not Gaussian? For example, can we detect the noise level if it is uniformly distributed with standard deviation δ? Let us consider this case with the distribution of the noise being $U[a, b]$, where $a = -b$, in which case $\delta^2 = b^2/3$. We have plotted graphs of

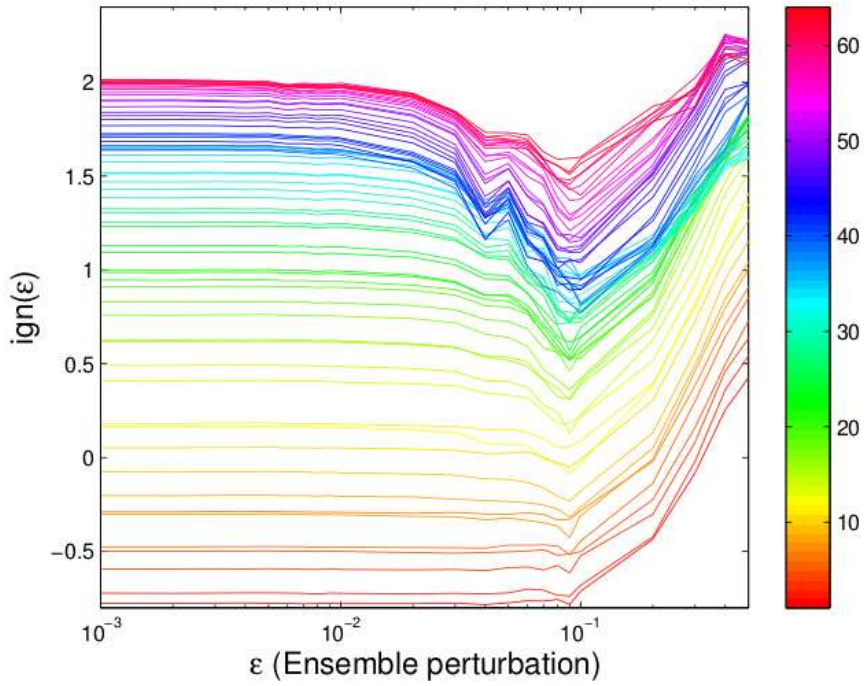


Figure 5.13: Graphs of ignorance versus logarithmically varying standard deviation, ϵ, of ensemble perturbations with uniformly distributed observational error of standard deviation $\delta = 10^{-1}$ on MS data with an imperfect model. 128 initial conditions with a time step 64 between them were used. 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 colorbar on the right.

average Ignorance versus ensemble perturbation in figure 5.13 with $\delta = 10^{-1}$. Again we see some critical behaviour at $\epsilon = \delta$.

What happens when the data is noise free? There are two cases we consider: (i) dynamical noise is present and (ii) there is no dynamical noise. Graphs for these two cases are shown in figure 5.14. The main difference between the graphs is that the ones for a model with dynamical noise generally exhibit lower values of average Ignorance, and this is more pronounced at high lead times. Apart from that, the graphs are qualitatively similar to the perfect and imperfect model scenario with observational noise on the data.

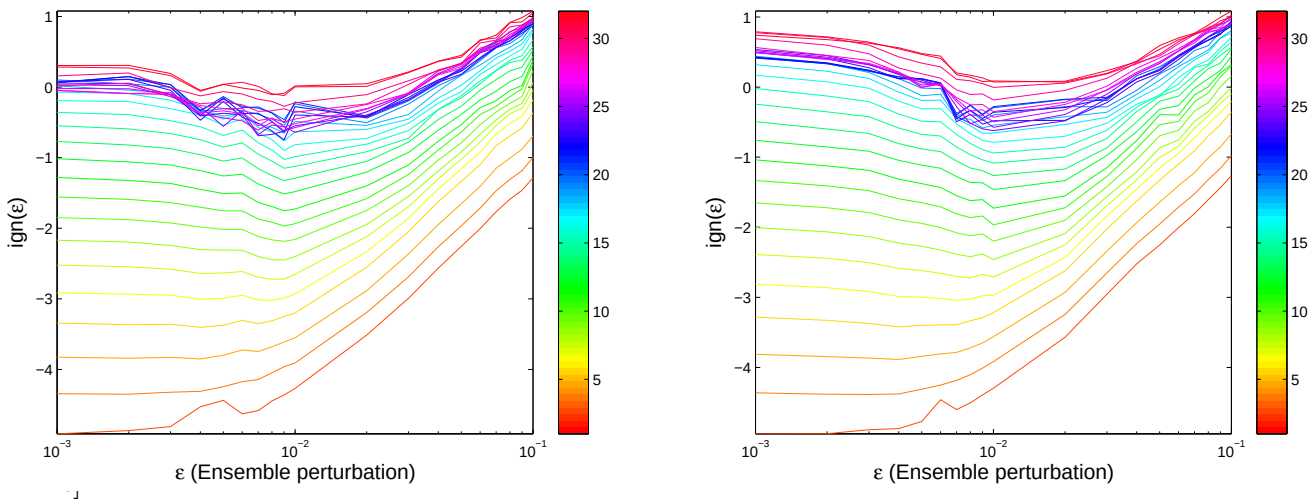


Figure 5.14: The graphs of ignorance versus ensemble perturbation with 256 ensembles, each ensemble containing 32 members and using a cubic global model with dynamical noise of standard deviation 3×10^{-3} (left) and a without dynamical noise (right) on noise free M-S data. The colour bar on the right shows the lead times for the different graphs of Ignorance.

The foregoing discussions can be summed up as follows: Whereas there is similarity in the graphs of average Ignorance for the PMS with observational noise and the IMS, there is a clear difference with the PMS on clean data shown in figure 5.9. 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. It should prove better than using tools like *Talagrand* diagrams, which can only eliminate bad models[15]; and their success hinges on whether the initial conditions are perfect or not. Also, if the noise level dominates model error, we may be in a position to detect the noise level. Otherwise, in general, we cannot be sure if the problem is model error or observational noise.

[15]It has been demonstrated that flat Talagrand diagrams are not a sufficient condition for forecasts to be perfect [18].

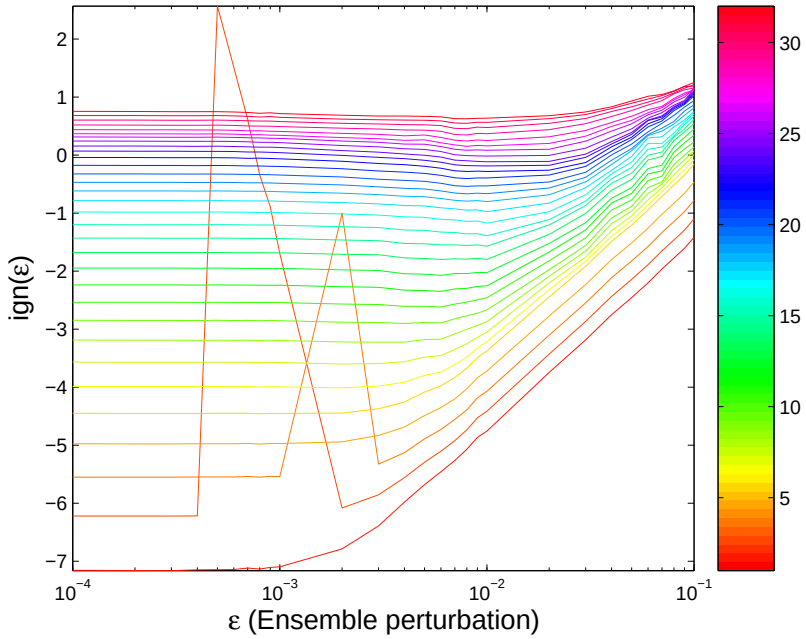


Figure 5.15: Graph of ignorance versus logarithmically standard deviation, ϵ, of Gaussian perturbations on circuit data. 128 initial conditions with a time step 64 between them were used. 32 initial ensembles were generated in each and iterated forward up to 64 time steps. The multiple lines correspond to different lead times according to the colorbar on the right.

Before ending this section, we consider the circuit. Some of the questions we wish to answer for the circuit are, what perturbation level should we use for a given model? Is there measurement noise? These questions are addressed by the use of Ignorance as we have explained in the preceding paragraphs. A graph of average Ignorance versus perturbation level is shown in figure 5.15. Notice that the graph of the first lead time begins to rise at $\epsilon \approx 10^{-3}$, which is quite small. This suggests that the noise is very low. This is comparable to both the noise in the circuit and the standard deviation of the one-step errors of the model. The graphs look very much like those obtained with MS-data without observational noise, but with an imperfect model (see figure 5.14).

5.7 Conclusions

In this chapter, by looking at the perturbed MS system, we reinforced the invaluability of ensemble prediction. We also re-interpreted ensemble prediction as a way of coping with model error when the underlying system is structurally stable. In particular, we showed portions of the time series where ensemble prediction out-performs point forecasts of the MS system in the sense of shadowing times.

The concept of skill scores was then systematically developed. The decomposition of a skill score into sharpness and reliability proposed by Broecker et al. [7] was modified. What they had as the sharpness term did not depend on the forecast and could, therefore, not be the sharpness of the forecast. We also showed how the empirical skill score relates to the overall theoretical skill of the time series. In our discussion of dressing by the use of the Ignorance score, we derived the variance of the forecasts.

Thirdly, we demonstrated that Ignorance can be used to determine an optimum initial-perturbation spread for a given system and model. The standard deviation of the optimum perturbation turns out to be the maximum of that of observational uncertainty and one-step-errors of the model in question. Although it is critical that we use Gaussian perturbations, the distribution of the underlying uncertainty or model error seems not to play a critical role. It turns out that we can also diagnose the fictitious case of a perfect model with perfect initial conditions.

Things that are new in this chapter are:

- Re-interpreting ensemble prediction a way of making searches for diffeomor-

phisms that map the initial conditions from the system state space to the model state space.

- Demonstrating with the example in § 5.1 that predicting the perturbed MS-system with the unperturbed MS system from exact initial condition can be out-performed by ensemble prediction.

- An alternative decomposition of a skill score into reliability and sharpness and showing that the reliability term is positive definite. The expected empirical score was also related to the theoretical score.

- The variance of the dressed ensembles and its failure to faithfully diagnose the sharpness of the PDF.

- Demonstrating that blending with climatology increases the skill of the PDFs.

- The use of the Ignorance skill score to find the optimum standard deviation of initial ensemble perturbations and an explanation of the graphs of Ignorance versus perturbation with the special case where the forecast PDFs are Gaussian.

- The explanation of the graphs of Ignorance versus initial-perturbation spread.

In the next section, we shall consider how to combine ensemble outputs of different models to make a single forecast PDF that is as skillful as possible.

Chapter 6

Multiple Models

In the preceding chapters we have noted that different global models of the circuit performed differently across the different regions of the attractor. Put another way, distributions of q-pling times of various models on the attractor differ. This was noticed to be the case in chapter 4. These suggested that variations in predictability loss by the various models would differ.

As discussed in the previous chapter, model performance may be scored by the use of Ignorance, which is a skill score. The variation in the performance of each model against the others may follow an alternating fashion and, on average, one of them will be the best. Should we use the single best model and ignore the rest? How does using a single best model fare against using the combined not so good models? These are the questions addressed in this chapter.

A new way of combining forecasts of multiple models with particular reference to the circuit is presented. To turn the model output ensembles into skillful PDFs, we appeal to the kernel dressing technique discussed in section 5.4 and 5.5. Our approach is somewhat reminiscent to that of Raftery et al. [54].

This chapter is organised as follows: In § 6.1, we give an example of how two models of the circuit out-perform each other at different regions of the attractor at a macro-scale. In § 6.2, we go through the theory employed to combined probabilistic forecasts. The theory is applied to models of the circuit in § 6.3. Finally, the conclusions and a list of new things are given in § 6.4

6.1 Multiple Model Ensembles

Suppose we have J multiple, dynamical models, $\{M_1, \ldots, M_J\}$ of some physical system of interest, and we can make observations $\boldsymbol{x}$ in multiple dimensions. At a given initial condition, $\boldsymbol{x}_0$, instead of just using only $\boldsymbol{x}_0$ to make future predictions, we can generate many other initial conditions that are close to the observation and then use each of these to predict the future. At forecast time t, instead of just one prediction, we will have an ensemble of predictions. Suppose the quantity we wish to predict is a scalar random variable, say Y. Model M_j will yield an ensemble, $\{y_i^{(j)}\}_{i=1}^N$, of predictions. Should we pick a single model for forecasting and ignore the rest of the models?

To address this question let us consider two models of the circuit, M_1 and M_2. In Figure 6.1, we show the ensemble predictions for two different initial conditions x_0^1 and x_0^2 using models M_1 (green) and M_2 (blue). It is clear that the different models display different performances for different initial conditions. In fact, on the left, M_1 stays close to the verification longer than M_2 and vice versa on the right. This is a motivation for using both models to make predictions. If we use a single model, we would not capture some behaviour at some regions of the attractor. But how shall

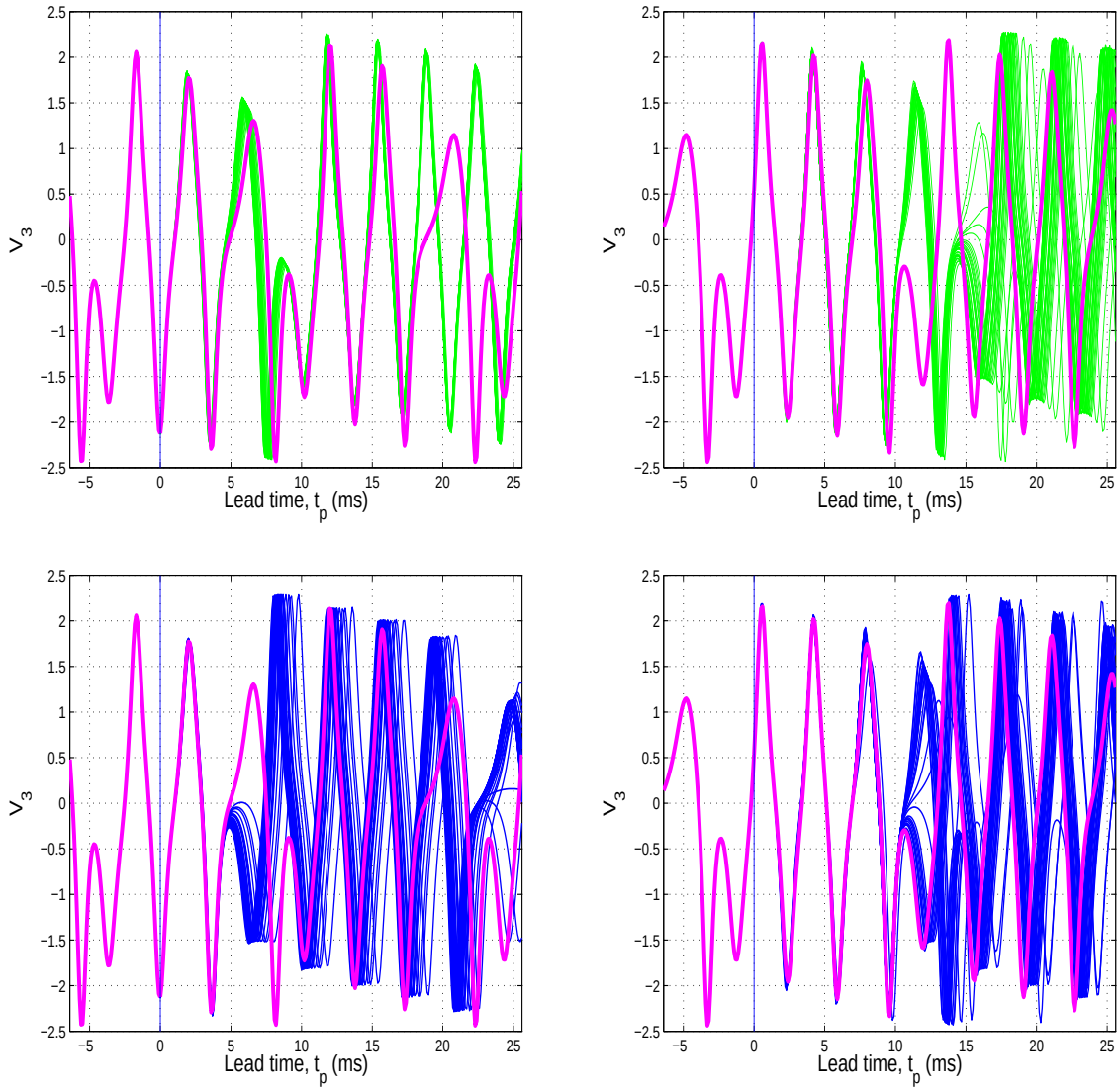


Figure 6.1: The graphs show ensemble predictions of a voltage signal, V_3 (magenta), by two models, M_1(green) and M_2 (blue), from two initial conditions x_0^1 (left) and x_0^2 (right).The vertical line at $t = 0$ indicates the time from which predictions are made. Notice that in the left pictures, M_1 stays close to the verification for a longer time than M_2 and vice versa in the right pictures.

we combine these predictions in an optimal manner?

6.2 Combining Probabilistic Forecasts

Given J dynamical models, $\{M_1, \ldots, M_J\}$ and training data $\mathcal{D}_T = \{z_d\}_{d=1}^{D}$. Then the Bayesian law of total probability says that the forecast PDF, $p(y|\mathcal{D}_T)$, is [54]

$$p(y|\mathcal{D}_T) = \sum_{j=1}^{J} p(y|M_j)p(M_j|\mathcal{D}_T), \tag{6.1}$$

where $p(y|M_j)$ is the forecast PDF of y based on model M_j alone, and $p(M_j|\mathcal{D}_T)$ is the posterior probability of model M_j being correct given the training data. Since we acknowledge that our models are wrong, we should rather think of $p(M_j|\mathcal{D}_T)$ as a weight reflecting how well model M_j fits the training data. The posterior model probabilities satisfy the condition

$$\sum_{j=1}^{J} p(M_j|\mathcal{D}_T) = 1. \tag{6.2}$$

Whereas obtaining more data does help improve the PDFs, as long as the models are imperfect, we will reach a saturation point when the PDFs cease to improve. The following formalism shall be employed to blend dressed ensembles of individual models.

Given a dth ensemble of points obtained by forecasting with model M_j, $\{y_i^{(d,j)}\}_{i=1}^{N}$, we can fit a probability density function

$$\rho_j^{(d)}(y) = \frac{1}{\sigma_j N} \sum_{i=1}^{N} K\left(\frac{y - y_i^{(d,j)} - \mu_j}{\sigma_j}\right),$$

where K is the kernel, σ_j is the kernel spread and μ_j is the offset parameter. To account for model error, we incorporate the climatology $\rho_{cl}(y)$, obtained from historical data, so that the PDF of model M_j is

$$f_j^{(d)}(y) = \alpha_j \rho_j^{(d)}(y) + (1 - \alpha_j)\rho_{cl}(y), \tag{6.3}$$

where $0 \leq \alpha_j \leq 1$. With sufficient training data, if the forecasts are very good relative to the climatology, we expect $\alpha_j \sim 1$. Optimum parameters μ_j and σ_j are determined over a historical archive of ensemble forecast-verification pairs. After training the individual models, we can then blend the PDFs to obtain:

$$f^{(d)}(y) = \sum_{j=1}^{J} w_j \rho_j^{(d)}(y) + \alpha \rho_{cl}(y), \tag{6.4}$$

where the coefficients $\alpha, w_j \geq 0$ satisfy the condition

$$\alpha + \sum_{j=1}^{J} w_j = 1. \tag{6.5}$$

The weights of $f^{(d)}(y)$ can be found by minimising the average Ignorance,

$$\langle \text{ign} \rangle(\boldsymbol{\omega}) = -\frac{1}{D} \sum_{d=1}^{D} \log_2 f^{(d)}(z_d) \tag{6.6}$$

over $\mathcal{D}_T$, subject to (6.5), where $\boldsymbol{\omega} = (w_1, \ldots, w_J, \alpha)$ is the vector of weights. Minimising the Ignorance is equivalent to maximising the associated likelihood function,

$$L(\boldsymbol{\omega}) = \sum_{d=1}^{D} \log_2 f^{(d)}(z_d). \tag{6.7}$$

Suppose τ_s is the sampling time such that $\tau = \tau_s d$. In the limit $\tau_s \to 0$, recall that the empirical skill is an approximation of

$$\mathbb{E}[\text{ign(f)}](\boldsymbol{\omega}) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbb{E}[\text{ign}(f^{(\tau)}, Y)] \mathrm{d}\tau, \tag{6.8}$$

where

$$\mathbb{E}[\text{ign}(f^{(\tau)}, Y)] = -\int_{-\infty}^{\infty} p^{(\tau)}(y) \log f^{(\tau)}(y) \mathrm{d}y$$

and $p^{(\tau)}(y)$ is the corresponding perfect forecast at time τ. If $\boldsymbol{\omega}_*$ is the minimiser of the overall score of the forecast, $\mathbb{E}[\text{ign(f)}](\boldsymbol{\omega})$, we can express this as

$$\mathbb{E}[\text{ign(f)}](\boldsymbol{\omega}_*) \leq \mathbb{E}[\text{ign(f)}](\boldsymbol{\omega}). \tag{6.9}$$

If we plug in $\boldsymbol{\omega} = (0, \ldots, w_j, 0, \ldots, 1 - \alpha_j)$ with $w_j = \alpha_j$, then (6.9) reduces to

$$\mathbb{E}[\text{ign(f)}](\boldsymbol{\omega}_*) \leq \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[\text{ign}(f_j^{(\tau)}, Y)]\mathrm{d}\tau. \tag{6.10}$$

Therefore, the blended forecast will out-perform each of the individual model ensembles on average. In other words, the mixture model will always beat a component.

We can also extract deterministic forecasts by taking the expectations of Y with respect to the PDFs. To this end, we can first note, from equation (5.56), that

$$\int y\rho_j^{(d)}(y)\mathrm{d}y = \bar{y}^{(d,j)} + \mu_j,$$

where $\bar{y}^{(d,j)}$ is the mean of the dth ensemble of model M_j. The expectation of y given forecasts by model M_j is

$$\begin{aligned} \mathbb{E}[Y|\{y_i^{(d,j)}\}_{i=1}^N] &= \int y f_j^{(d)}(y)\mathrm{d}y \\ &= \alpha_j(\bar{y}^{(d,j)} + \mu_j) + (1 - \alpha_j)(\bar{z} + \mu_c), \end{aligned} \tag{6.11}$$

where $\bar{z}$ is the mean of the historical data used to construct the climatology and μ_c is the climatological offset. The expectation of Y given the forecasts by all the models is given by

$$\mathbb{E}[Y|\{y_i^{(d,j)}\}_{i=1}^N, j = 1, \ldots, J] = \sum_{j=1}^{J} w_j(\bar{y}^{(d,j)} + \mu_j) + \alpha(\bar{z} + \mu_c). \tag{6.12}$$

Equations (6.11) and (6.12) may be used as a forecasts of y.We may compare the distributions of absolute prediction errors of the deterministic forecasts to evaluate model performance.

6.3 Application to the Circuit

Let us consider four global models of the circuit, M_j, $j = 1, \ldots, 4$. First the models shall be ranked at various lead times using the ignorance score. The training and

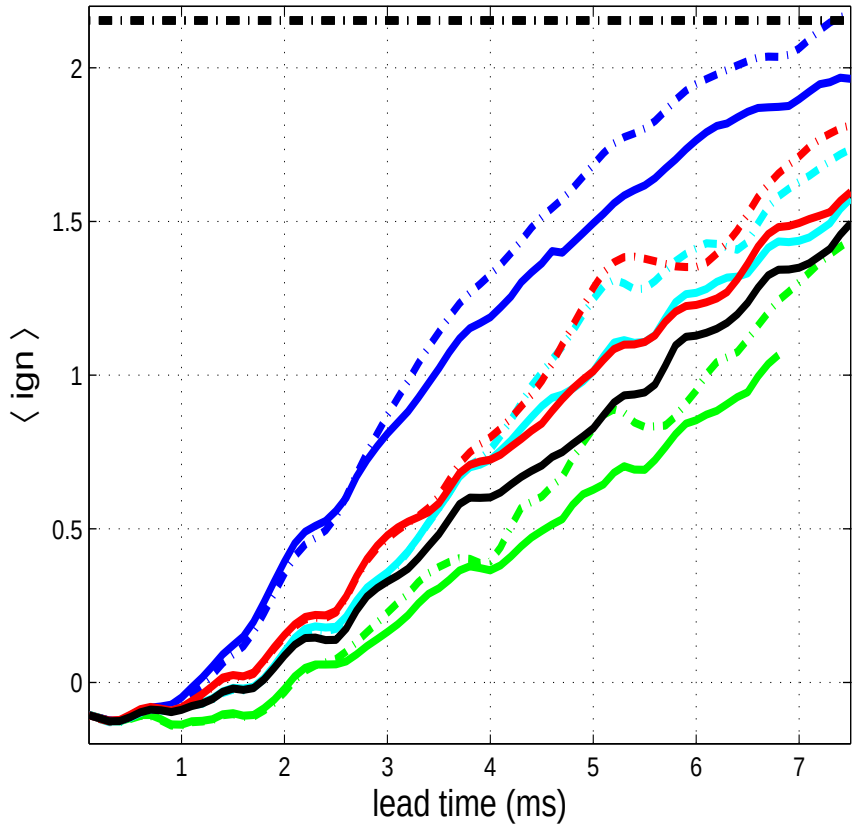


Figure 6.2: Graphs of average Ignorance against lead time on 1024 training data points for dressed ensemble predictions by four models of the circuit M_1 (blue line), M_2 (green), M_3 (cyan), M_4 (red) and the combined model forecasts (of M_1, M_3 and M_4), M_c (black). In each case the solid lines correspond to ensembles blended with climatology and the corresponding dashed lines to not blending with it. The black dashed line is the climatology.

testing data sets each consists of 1024 data points. Each model has $N = 32$ ensemble members. The training set is used to fit parameters and the testing data set is used to score the performance of the models. For all the models, blending with climatology out-performs not doing so as seen in figures 6.2 and 6.3. We note that model M_2 outperforms the other models in both the training period (see figure 6.2) and testing period (see figure 6.3) because it has the lowest score at all lead times. A combined model, M_c, is produced using only M_1, M_3 and M_4 by training over 1024 using the pre-trained individual models (blended with climatology) and setting only

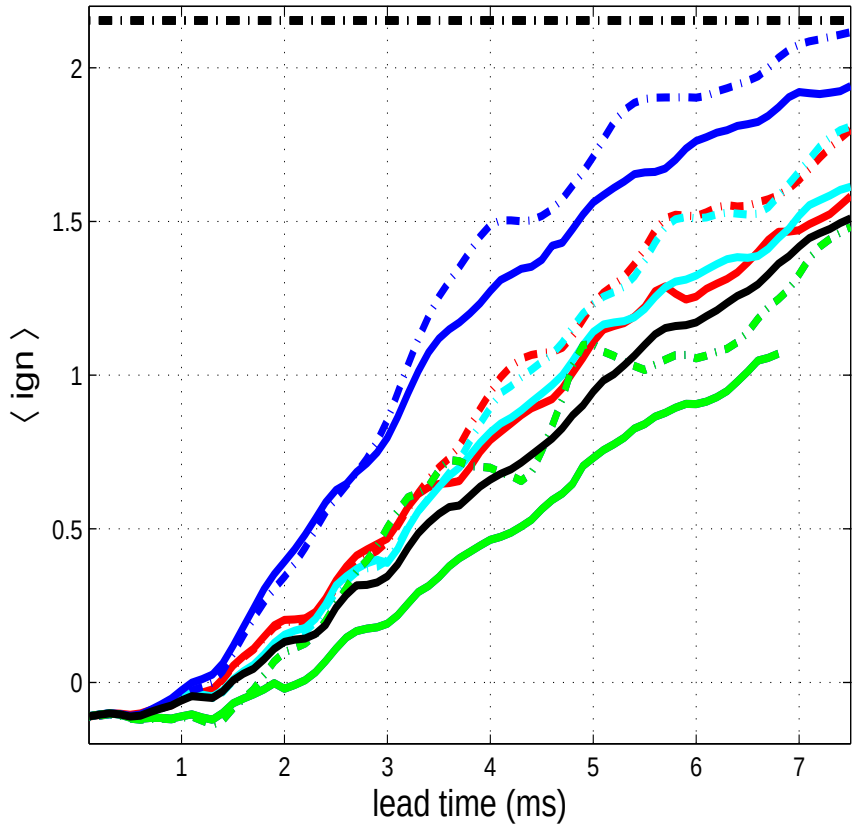


Figure 6.3: Graphs of average Ignorance over 1024 testing data against lead time on testing data for dressed ensemble predictions by four models of the circuit M_1 (blue line), M_2 (green), M_3 (cyan), M_4 (red) and the combined model forecasts (of M_1, M_3 and M_4), M_c (black) in the testing data. The black dashed line is the climatology.

the blending parameters. It is evident from figure 6.2 that model M_c out-performs the constituent models in the training period. It out-performs them again out-of-sample as shown in figure 6.3. At short lead times, all the models perform equally and they begin to separate after a lead time of 0.5 ms.

The weights accorded to the constituent models of model M_c are shown in figure 6.4. Notice that, on average, the worst model (M_1) is accorded a near zero weight, which is consistent with the results of figure 6.3. It is only within the 1 ms lead time that

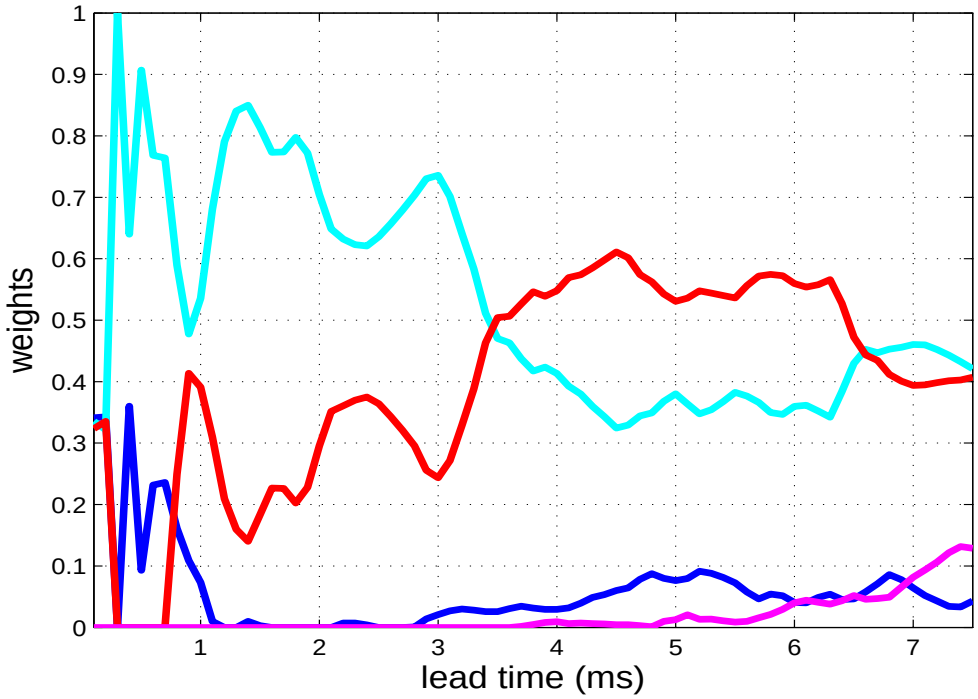


Figure 6.4: Graphs of the weights of the constituent models of the combined model, M_c, against lead time: M_1 (blue line), M_3 (cyan), M_4 (red) and the climatology (magenta).

it competes well with with the other models. Figure 6.4 indicates that blending combines the models according to how they perform individually.

The mean absolute errors of the associated deterministic forecasts obtained using equations (6.11) for the individual models and (6.12) for the combined model yielded the graphs shown in figure 6.5. The deterministic forecasts were computed out-of-sample. According to these graphs, the combined model performs better than the individual, constituent models. Again, the combined model is out-performed by the best model. Nevertheless, the results of this score do not agree with the those of Ignorance for models M_3 and M_4[1]. In fact, if we consider model M_i to be better than

[1]For instance, at lead times higher than 7 ms.

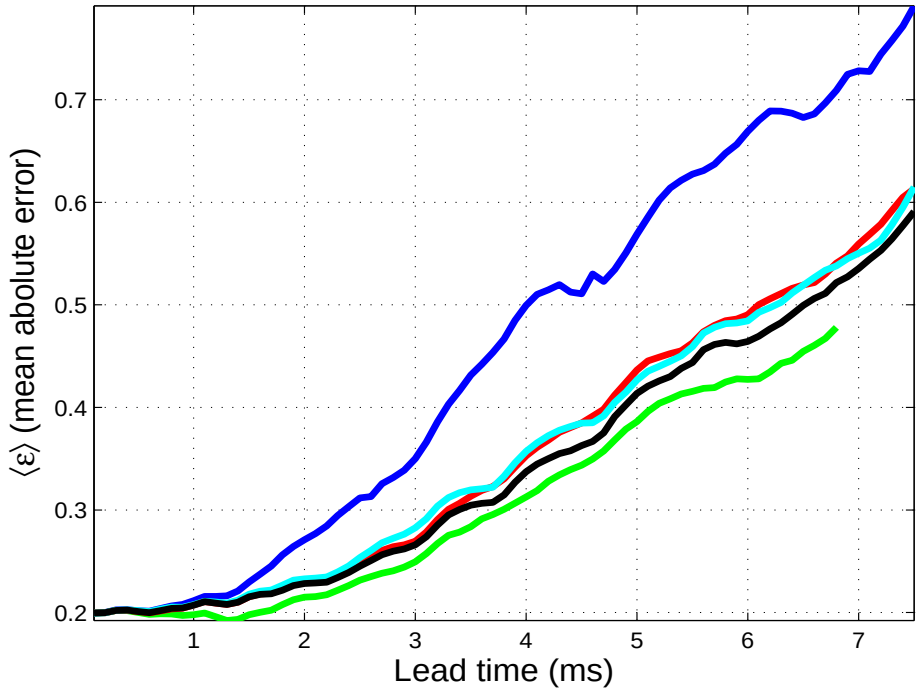


Figure 6.5: Graphs of the out-of-sample mean absolute errors of the deterministic forecasts of models M_1 (blue), M_2 (green), M_3 (red), M_4 (cyan) and M_c (black) obtained using equations (6.11) for the individual models and (6.12) for the combined model.

model M_j when

$$F_i(\varepsilon) \geq F_j(\varepsilon) \text{ for all } \varepsilon, \tag{6.13}$$

where $F_i(\varepsilon)$ is the cumulative distribution of absolute forecast errors of model M_i, then

$$\mu^{(i)} \leq \mu^{(j)} \tag{6.14}$$

does not imply that (6.13) holds[2]. A forecast that has a few very poor ensemble members will be heavily penalised by the mean score even if the vast majority of the ensemble members are close to the verification. Of course, if (6.13) holds then (6.14)

[2] $\mu^{(i)} = \int_0^\infty \varepsilon F_i'(\varepsilon) d\varepsilon$

holds because

$$\begin{aligned}\mu^{(i)} &= \int_0^\infty \varepsilon F_i'(\varepsilon)\mathrm{d}\varepsilon \\ &= \int_0^\infty \varepsilon \left[\frac{d}{\mathrm{d}\varepsilon}\left(F_i - 1\right)\right]\mathrm{d}\varepsilon \\ &= \left[\varepsilon(F_i(\varepsilon) - 1)\right]_0^\infty - \int_0^\infty (F_i(\varepsilon) - 1)\mathrm{d}\varepsilon.\end{aligned}$$

Hence

$$\mu^{(i)} = \int_0^\infty (1 - F_i(\varepsilon))\mathrm{d}\varepsilon. \tag{6.15}$$

Expressed another way,

$$\mu^{(i)} = \int_0^\infty P(X_i > \varepsilon)\mathrm{d}\varepsilon.$$

Similar arguments apply to the root-mean-square error. Nevertheless, we should not miss the point that averaging the ensembles in this way to obtain a deterministic forecast is superior to a simple average of the individual ensemble members.

6.4 Conclusions

In this chapter, it has been shown that additional forecasting skill, as quantified by comparing the average Ignorance scores, can be obtained by combining the output of inferior models. This has effectively afforded us a way of exploiting the diversity in our models. The weighted average deterministic forecast of the combined models out-performs those of the constituent models, but is out-performed by the best single model. To what extent these results have been affected by looking at marginal PDFs remains to be investigated.

The following is a list of things that are new in this chapter:

- The combined model out-performs the constituent models.

- Deterministic forecasts of the combined model out-perform those of the constituent models.

- The idea of blending the different model ensembles by first dressing the individual model ensembles and then selecting the weights of the pre-trained models.

Chapter 7

Conclusions and Further Work

We shall here draw the main conclusions of this thesis. Our goal was to explore and exploit model error in the models of the circuit. To explore model error, we computed and compared distributions of q-pling times for various models of the circuit. It was found that differences in the distributions of q-pling times were maximised by differences in coordinate spaces. For instance, delay space models exhibited more similarities among themselves than with measurement space models. It is, therefore, persuasive to model the circuit in various coordinate spaces to increase the diversity in our models. Probabilistic forecasts from these models can then be blended together to obtain superior forecasting skill to using the constituent models.

Whereas we obtained superior skill by averaging over the attractor, there is still need to weight the forecasts according to the various regions of the attractor. Such an approach has the potential to provide superior forecasts to averaging over the entire attractor. The results of this thesis may be readily applied to other areas of the applied sciences where chaos are assumed. In the case of weather forecasting, various regions of the attractor could correspond to the different seasons.

The following are things to be pursued in the future:

- It would be interesting to use some data assimilation techniques to determine optimum parameters in MS model of the circuit. It could then be enquired as to how well the resulting model performs?

- The probability density functions used in the computations were univariate (marginal). To what extent projection affected the results remains to be addressed by considering multivariate probability density functions.

- The method for determining the radius of the initial uncertainty ball was ad-hoc and I would like to formalise it in a mathematical framework.

- Explore the performance of global models based on the refined Kwasniok-Smith learning set. It would be useful if these models out-perform those based on the natural measure of the underlying attractor.

- The benefits of using models that are diverse with respect to q-pling times needs to be pursued investigating contrasting the performance of similar combined models with that of more diverse models.

- Measuring how reliable models are needs investigated closely.

Appendix A

Lorenz system

The Lorenz system [37, 68] is given by the following equations:

$$\begin{aligned} \dot{x} &= -\sigma x + \sigma y, \\ \dot{y} &= -xz + rx - y, \\ \dot{z} &= xy - bz, \end{aligned} \tag{A.1}$$

where, in its original formulation, x describes the intensity of the convective motion, y characterises the temperature difference between ascending and descending fluid elements, and z is proportional to the deviation of the temperature profile from its equilibrium value. The clasical parameter values, $\sigma = 10$, $r = 28$ and $b = \frac{8}{3}$ which yield the standard Lorenz attractor. The Jacobian of the vector field is given by

$$J = \begin{bmatrix} -\sigma & \sigma & 0 \\ (r - z) & -1 & -x \\ y & x & -b \end{bmatrix}.$$

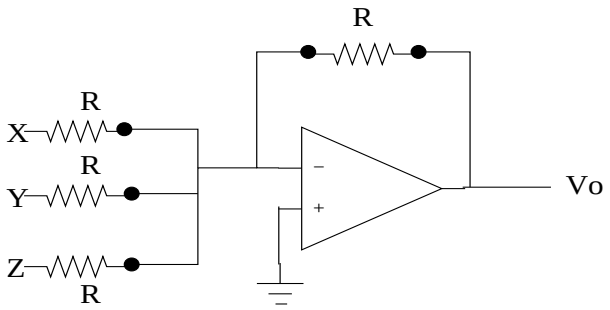


Figure B.1: Schematic diagram of an adder module.

Appendix B

Circuit Modules

An adder module is shown in figure B.1. For this adder

$$V_o = X + Y + Z. \tag{B.1}$$

To test the Op-Amp, one may input three signals that are phase. The ouput should then have an amplitute equal to the sum of the constituent signals.

A diagram of an integrator is shown in figure B.2. The transfer function is given by

$$V_o(t) = \frac{1}{RC}\int_0^t V(s)\mathrm{d}s. \tag{B.2}$$

RC is the time constant of integrator.

A schematic diagram of the multiplier module is shown in figure B.3. Each mul-

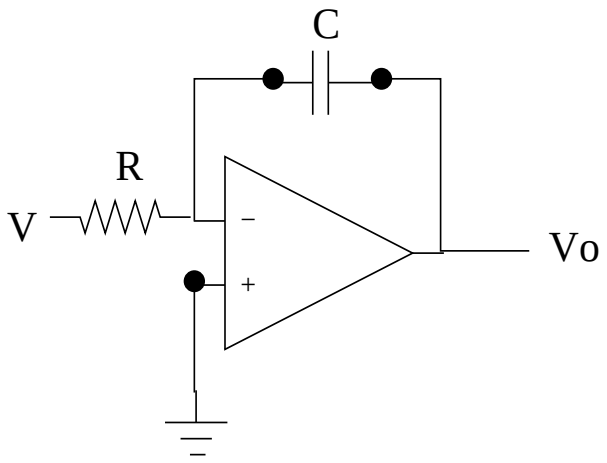


Figure B.2: Schematic diagram of an integrator module.

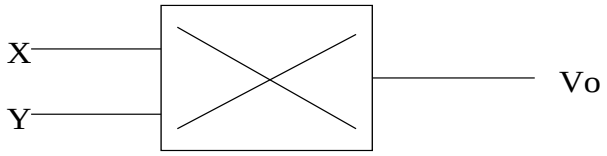


Figure B.3: Schematic diagram of a multiplier module.

tiplier was checked by inputing $X = Y$ plotting X versus V_o. The only flaw with this is that if the transfer function is $V_o = X^3/Y$, this malfunction would not be detected because the out-put would still be $V_o = X^2$.

Appendix C

Least Squares Solution

Different algorithms are available for solving the least squares problem (3.16), but the most numerically stable is the *singular value decomposition* (SVD). If $C = U\Sigma V^T$ is the SVD of C, then U and V are orthogonal matrices and Σ is a diagonal matrix of *singular values*, i.e $U^TU = I$, $V^TV = I$ and $\Sigma = \text{diag}(\sigma_i)$, in which case

$$a_{LS} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i, \tag{C.1}$$

where $r\text{rank}(C)$ and a_{LS} is the value of a that solves the least squares problem. r may be estimated from the computed singular values of C by choosing a tolerance $\delta > 0$ and using the convention that C has numerical rank $\hat{r}$ if the $\hat{\sigma}_i$ satisfy

$$\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_r > \delta \geq \hat{\sigma}_{r+1} \geq \cdots \geq \hat{\sigma}_{m+1} \tag{C.2}$$

and δ should be consistent with the machine precision, $\mathbf{u}$. For example, we may take $\delta = k\mathbf{u}||C||_\infty$. However, if the general relative error in the data[1], ϵ, exceeds $\mathbf{u}$, δ should be correspondingly bigger, say $\delta = k\epsilon||C||_\infty$.

[1]Relative error in the data may be defined as, $\epsilon = \frac{||\hat{s}-s||}{||s||}$.

Bibliography

[1] Final report on Mathematical techniques for improving forecasting of hazardous weather. University of Reading, 16-20, June, 2003.

[2] A. Aminian and M. K. Kazimierczuk. *Electronic Devices: a design approach.* Pearson Prentice Hall, 2004.

[3] 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, 1994.

[4] M. Balkanski and R. F. Wallis. *Semiconductor Physics and Applications.* Oxford University Press, 2000.

[5] N. J. Balmforth and R. V. Craster. Sychronizing Moore and Spiegel. *Chaos*, 7:738:752, 1997.

[6] J. Barrow-Green. *Poincaré and the Three Body Problem*, volume History of Mathematics Volume 11. American Mathematical Society, 1997.

[7] J. Brocher, Clarke L., Kilminster D., and Smith L. A. Scoring probabilistic forecasts. 2005.

[8] J. Broecker and L. A. Smith. From Ensemble Forecasting to Predictive Distribution Functions. Preprint, Tellus, 2007.

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

[10] J. M. Calvert and M. A. H. McCaulsland. *Electronic*. John Wiley and Sons, 1985.

[11] M. Casdagli. Nonlinear Prediction of Chaotic Time Series. *Physica D*, 35:335–356, 1989.

[12] P. Cvitanovic. Invariant Measurement of Strange Sets in Terms of Cycles. *Phys. Rev. Lett.*, 61:2729–2732, 1988.

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

[14] A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33:1134, 1986.

[15] M. I. Glavinovic. Comparison of Parzen density and frequency histograms as estimators of probability density functions. *Eur. J. Phsiol.*, 433:174–179, 1996.

[16] J. Gleick. *Chaos*. Vintage, 1998.

[17] P. Glendinning. *Stability, Instability and Chaos: an introduction to the theory of nonlinear differential equations*. Cambridge University Press, 1999.

[18] T. Gneiting, F. Balabdaoui, and A. E. Raftey. Probabilistic forecasts, calibration and sharpness. Technical Report no. 483, Deparment of Statistics, University of Washington, May, 2005.

[19] G. H. Golub and C. F. Loan. *Matrix Computations*. The John Hopkins University Press, third edition, 1996.

[20] P. Grassberger and I. Procaccia. Characterisation of strange attractors. *Phys. Rev. Lett.*, 50:346–349, 1983.

[21] J. Guckenheimer. *The Hopf Bifurcation and its Applications*. Springer-Verlag, New York, 1976.

[22] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer, New York, 1983.

[23] W. Hardle, M. Muller, S. Sperlich, and A Werwatz. *Nonparametric and Semiparametric Models*. Springer-Verlag, Berlin Heidelberg, 2004.

[24] McEliece R. J. *The Theory of Information and Coding*. Cambridge University Press, second edition, 2002.

[25] I. R. H. Jackson. Convergence Properties of Radial Basis Functions. *Constructive Approximation*, 4:243–264, 1988.

[26] K. Judd and Smith L. A. Inditinguishable states ii: The imperfect model scenario. *Physica D*, 196:224–242, 2004.

[27] K. Judd and A. Mees. On selecting models for nonlinear time series. *Physica D.*, 82:426–444, 1995.

[28] K. Judd and A. Mees. Embedding as a modelling problem. *Physica D*, 120:273–286, 1998.

[29] E. Kalnay, M. Corazza, and M. Cai. Are Bred Vectors the same as Lyapunov Vectors. University of Maryland, 2002.

[30] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge University Press, 1999.

[31] A. I. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, Inc., 1st edition, 1957.

[32] D. Kilminster. *Modelling Dynamical Systems via Behaviour Criteria*. PhD thesis, University of Western Australia, 2002.

[33] D. Kilminster and R. Machete. Prediction, behaviour, ignorance. *Nolta*, 2005.

[34] S. Kullback and R. A. Leiber. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[35] F. Kwasniok and L. A. Smith. Real-Time Construction of Optimised Predictors from Data Streams. *Phys. Rev. Let.*, 92:164101–1, 2004.

[36] J. M. Lee. *Introduction to Topological Manifolds*. Springer-Verlag, 2000.

[37] E. N. Lorenz. Deterministic Non-periodic flow. *J. Atmos. Sci.*, 20:130–141, 1963.

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

[39] D. G. Luchisky, P. V. E. McClintock, and M. I. Dykman. Analogue studies of nonlinear systems. *Reports on Progress in Physics*, 61:889–997, 1998.

[40] T. Lyons, editor. *The Interpretation and Solution of Ordinary Differential Equations Driven by Rough Signals*, volume 57. Proceddings of Symposia in Mathematics, AMS, 1995.

[41] D. Madigan and A. E. Raftery. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, 89:1535–1546, 1994.

[42] B. B. Mandelbrot. *Fractals: Form, Chance and Dimension*. Freeman, San Francisco, 1977.

[43] M. Martelli. *Introduction to Discrete Dynamical Systems and Chaos*. Wiley Inter-Science, first edition, 1999.

[44] R. May. A simple mathematical equation with very complicated dynamics. *Nature*, 261:459–469, 1976.

[45] P. E. McSharry and L. A. Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. *Physical Review Letters*, 83:4285–4288, 1999.

[46] C. A. Michelli. Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions. *Constructive Approximation*, 2:11–22, 1986.

[47] W. D. Moore and E. A. Spiegel. A thermally excited nonlinear oscillator. *The Astrophysical Journal*, 143:871–887, 1966.

[48] E. Ott, T. Sauer, and J. Yorke. *Coping With Chaos: Analysis of chaotic data and the exploitation of chaotic system*. John Wiley and Sons, Inc. New York, 1994.

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

[50] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill International Editions, New York, 1991.

[51] G Parker. *Introductory Semiconductor Device Physics*. Prentice Hall, 1994.

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

[53] W. H. Press and W. T. Vettering. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, 1988.

[54] A. E. Raftery, T. Gneiting, F. Balabdoui, and M. Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133:1155–1174, 2005.

[55] M. S. Roulston and L. A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130:1653–1660, 2002.

[56] M. S. Roulston and L. A. Smith. Combining dynamical and statistical ensembles. *Tellus*, 55A:16–30, 2003.

[57] D. Salmon. *Data Compression: the complete reference*. New York Springer, 2001.

[58] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *Journal of Statistical Physics*, 65:579–616, 1991.

[59] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

[60] C. E. Shannon. A Mathematical theory of communication. *Bell Systems Technology Journal*, 27:379–423,623–656, 1948.

[61] C. E. Shannon, editor. *Communication in the presence of noise*, volume 37. Pro. Institute of Radio Engineers, 1949.

[62] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, first edition, 1986.

[63] S. Smale. Mathematical Problems for the next century. *Math. Intelligencer*, 20:7–15, 1998.

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

[65] L. A. Smith. Local Optimal Prediction: Exploiting strangeness and variation of sensitivity of initial conditions. *Phil. Trans. Royal Soc. London. A*, 348:371–381, 1994.

[66] L. A. Smith. Maintenance of uncertainty. Mathematical Institute, University of Oxford, OX1 3LB, U.K., 1998.

[67] L. A. Smith, C. Ziehman, and K. Raedrich. Uncertainty in dymanamics and predictability in chaotic systems. *Q. J. R. Meteorol. Soc.*, 125:2855–2886, 1999.

[68] C. Sparrow. *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*. Springer-Verlag, New York, 1982.

[69] J. S. Steinhart and S. R. Hart. Calibration curves for thermistors. *Deep Sea Res.*, 15:497–503, 1968.

[70] F. Takens. *Detecting strange attractors in turbulence*. Springer-Verlag, 1981.

[71] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[72] W. Tucker. A Rigorous ODE Solver and Smale's 14th Problem. *Found. Comput. Math.*, 2:53–117, 2002.

[73] D. H. Van Campen, M. D. Lazurko, and W. P. J. M van den Oever, editors. *Recurrence Analysis of a Moore-Spiegel Electronic Circuit, pp1800*, ENOC PROCEEDINGS 2005.

[74] D. Welsh. *Codes and Cryptography*. Oxford University Press, 1989.

[75] H. Whitney. Differentiable manifolds. *Ann. Math.*, 37:645–680, 1936.

[76] D. S. Wilks. Comparison of ensemble-mos methods in the lorenz '96 setting. *Meteorol. Appl.*, 13:243–256, 2006.

[77] C. Ziehmann, L. A. Smith, and K. Fraedrich. Localised lyapunov exponents and the prediction of predictability. *Phys. Rev. Lett.*, 271:237–251, 1999.