The London School of Economics and Political Science

Improving Predictability of the Future by Grasping Probability Less Tightly

Edward Wheatcroft

A thesis submitted to the Department of Statistics
of the London School of Economics for the degree of Doctor of Philosophy.

London, November 14, 2015

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identied in it).

I declare that my thesis consists of 40,891 words.

Abstract

In the last 30 years, whilst there has been an explosion in our ability to make quantative predictions, less progress has been made in terms of building useful forecasts to aid decision support. In most real world systems, single point forecasts are fundamentally limited because they only simulate a single scenario and thus do not account for observational uncertainty. Ensemble forecasts aim to account for this uncertainty but are of limited use since it is unclear how they should be interpreted. Building probabilistic forecast densities is a theoretically sound approach with an end result that is easy to interpret for decision makers; it is not clear how to implement this approach given finite ensemble sizes and structurally imperfect models. This thesis explores methods that aid the interpretation of model simulations into predictions of the real world. This includes evaluation of forecasts, evaluation of the models used to make forecasts and the evaluation of the techniques used to interpret ensembles of simulations as forecasts. Bayes theorem is a fundamental relationship used to update a prior probability of the occurence of some event given new information. Under the assumption that each of the probabilities in Bayes theorem are perfect, it can be shown to make optimal use of the information available. Bayes theorem can also be applied to probability density functions and thus updating some previously constructed forecast density with a new one can be expected to improve forecast skill, as long as each forecast density gives a good representation of the uncertainty at that point in time. The relevance of the probability calculus, however, is in doubt when the forecasting system is imperfect, as is always the case in real world systems. Taking the view that we wish to maximise the logarithm of the probability density placed on the outcome, two new approaches to the combination of forecast densities formed at different lead times are introduced and shown to be informative even in the imperfect model scenario, that is a case where the Bayesian approach is shown to fail.

Acknowledgements

I would like to thank my supervisor Professor Leonard Smith for his invaluable contribution to the work in this thesis. His enthusiasm and wisdom has been an endless source of motivation and inspiration throughout.

I would also like to thank all members of CATS for their support and friendship over the years and making my time at LSE a thoroughly rewarding and enjoyable experience.

‘All models are wrong, but some are useful.’

George E. P. Box

Contents

Abstract **ii**

Acknowledgements **iii**

1 Introduction **1**

1.1 Thesis Structure 8

2 Background Theory **12**

2.1 Forecasting 13

2.1.1 Dynamical Systems 13

2.1.2 Maps and flows 14

2.1.3 System-model pairs 15

2.1.4 Chaos 15

2.1.5 Point and Probabilistic forecasting 17

2.1.6 Forecasting framework 17

2.1.7 Perfect and Imperfect model scenarios 20

2.2 Indistinguishable States 21

2.2.1 Indistinguishable states and the case for probabilistic forecasting 25

2.3 Data Assimilation 27
2.3.1 Pseudo-orbit Data Assimilation 27
2.3.2 4DVAR 29
2.4 Probabilistic forecasting 30
2.4.1 Forming ensembles 30
2.4.2 System Density 34
2.4.3 Forming Forecast Densities from Ensembles 34
2.4.4 Gaussian Dressing 35
2.4.5 Kernel Density Estimation 36
2.4.6 Climatology 37
2.5 Probabilistic forecast evaluation 38
2.5.1 Scoring rules 38
2.5.2 Properties of scoring rules 39
2.5.3 Ignorance Score 40
2.5.4 Other scoring rules 40
2.5.5 Cross validation 41
2.5.6 Constructing Climatology 42
2.5.7 Kernel Dressing 42
2.5.8 Blending with climatology 44

3 Limitations of Linear Analysis 46

3.1 ACC for Data with a Trend 51
3.1.1 Linear Underlying Trend 53

3.1.2 Nonlinear Underlying Trend Assumed to be Linear 56

3.2 Relationship between the dispersion of outcomes and the ACC 60

3.2.1 Monthly performance of forecasts of the Central England Temperature record 61

3.3 Influential observations 64

3.4 The perpetual failure of linear correlation as an indication of predictability 67

4 Shadowing Ratios **69**

4.1 Shadowing Ratios 70

4.1.1 Shadowing 71

4.1.2 Definition of shadowing ratios 75

4.1.3 Uses of Shadowing Ratios 77

4.1.4 Example - Using shadowing ratios to compare model performance 78

4.2 Ensemble Shadowing Ratios 79

4.2.1 Example - Using ensemble shadowing ratios to compare the performance of ensemble formation techniques 82

4.3 Boosted Probability 84

4.3.1 Example - Assessing forecast density formation techniques using boosted probability 87

5 Kernel Dressing methods **91**

5.1 Variability in the dispersion of ensembles 95

5.2 Shortcomings of simple kernel dressing 98

5.3 K groups kernel dressing 108

5.3.1 Choosing the value of K . . . 113

5.4 Fixed Window Kernel Dressing . . . 117

5.5 Dynamic Kernel Dressing . . . 122

5.5.1 Robustness of dynamic kernel dressing . . . 128

5.6 Perfect Forecast Scenario . . . 130

5.6.1 Ignorance and Kullback-Leibler Divergence . . . 133

5.6.2 The performance of dynamic and simple kernel dressing in the perfect forecast scenario . . . 134

6 Beyond Bayesian Updating of Forecasts 141

6.1 Combining forecasts . . . 144

6.2 Forecast densities from multiple lead time ensembles . . . 146

6.2.1 Multiple lead time ensembles . . . 146

6.2.2 Naive method of building forecast densities from multiple lead time ensembles . . . 147

6.2.3 A time-weighted approach to building forecast densities from multiple lead time ensembles . . . 150

6.3 A Bayesian approach to the combination of forecast densities . . . 152

6.3.1 Bayes' Theorem . . . 154

6.3.2 Bayesian Inference . . . 155

6.3.3 Bayesian Updating with Probability Density Functions . . . 155

6.3.4 Pure Bayes method . . . 157

6.3.5 Experimental design . . . 158

6.3.6 Imperfect Model Scenario . . . 159

6.3.7 Perfect Model Scenario 162

6.4 Sequential blending 163

7 Evaluating Data Assimilation with Shadowing Ratios 170

7.1 Using Shadowing to Compare Data Assimilation Techniques 171

7.2 Experimental Design 173

7.3 PDA with estimated gradient information 173

7.4 PDA for systems with large variation in the standard deviation of variables 176

7.4.1 Rescaling the variables 178

7.5 The effect of nonlinearity on PDA and 4DVAR 182

7.5.1 Measure of nonlinearity 183

7.5.2 PDA versus 4DVAR in an increasingly nonlinear scenario . . . 185

8 Improving Probabilistic Prediction with Imperfect Models 190

8.1 Improving forecast densities using dynamic kernel dressing 192

8.1.1 A low dimensional system - Henon Map 192

8.1.2 A higher dimensional experiment: Lorenz '96 196

8.2 Using Boosted Probability Times to Compare Forecasting Systems . . 199

8.3 Combining Multiple Lead Time Forecasts 202

8.3.1 Moore-Spiegel IMS 204

8.3.2 Moore-Spiegel PMS 205

8.4 Linking shadowing and forecast skill 207

9 Conclusion 216

Appendix **218**

A Dynamical Systems **223**

A.1 Maps 223

A.1.1 Logistic Map 223

A.1.2 Henon map 224

A.1.3 Duffing Map 225

A.1.4 Ikeda map 226

A.2 Flows 228

A.2.1 Lorenz '63 228

A.2.2 Moore-Spiegel system 229

A.2.3 Lorenz '96 231

A.2.4 PST system 232

A.2.5 Rössler Map 233

B Details of Experiments **235**

C Details of Logarithmic Spiral **244**

C.1 Logarithmic Spiral 244

Bibliography **244**

List of Tables

3.1 The ACC of the forecasts $x_1, .., x_n$, the slope parameter b from fitting a regression of $y_1, .., y_n$ on $x_1, .., x_n$, the error of the original forecast x_n and the error of the calibrated forecast $\tilde{x}_n$ for different values of n. 68

5.1 Summary of the origins of the methods used in this chapter. 95

5.2 The optimised values of the blending parameter α and the kernel width σ, the optimised mean ignorance score over the training set and the mean ignorance over the test set from applying simple kernel dressing to the ensembles in experiment 5.A formed using a perfect model of the Lorenz '63 system. 101

5.3 Mean ignorance scores relative to simple kernel dressing of forecast densities formed from the test set of experiment 5.A using K groups kernel dressing for different values of K. Also shown are resampling intervals of the mean relative ignorance. Since, for the first 6 values of K considered, zero does not fall into the resampling intervals, K groups kernel dressing yields significantly more skillful forecasts than simple kernel dressing for these values of K. 112

5.4 Mean ignorance scores relative to simple kernel dressing of the forecast densities formed from the test set of experiment 5.A using fixed window kernel dressing for different values of M. Also shown are 95 percent resampling intervals of the mean relative ignorance in each case. Since zero does not fall into these intervals, a significant improvement is made over simple kernel dressing for all values of M considered except $M = 16$. The most effective value of M out of those considered appears to be 256. 121

5.5 The optimised parameter values, the minimised mean ignorance over the training set and the mean ignorance over the test set obtained from applying dynamic kernel dressing to the ensemble-outcome pairs in experiment 5.A. 123

6.1 Summary of the origins of methods used in this chapter. 144

6.2 Weighting parameters of sequential blending for the Lorenz '63 PMS in experiment 6.B. 168

7.1 Climatological standard deviations of the the PST system defined in appendix A.2.4 178

7.2 Results of applying PDA with and without rescaling the observations first. Values in brackets represent 95 percent resampling intervals. . . 182

8.1 Mean boosted probability times using each combination of ensemble formation scheme and kernel dressing method in experiment 8.C. Standard deviations are shown in brackets in each case. 202

B.1 Details of experiment 4.A 235

B.2 Details of experiment 4.B 236

B.3 Details of experiment 4.C 236

B.4 Details of experiment 5.A 237

B.5 Details of experiment 6.A 238

B.6 Details of experiment 6.B 238

B.7 Details of experiment 7.A 239

B.8 Initial conditions for the system trajectory in experiment 7.B shown in figure 7.2 and 7.3. 239

B.9 Details of experiment 7.B 239

B.10 Details of experiment 7.C 240

B.11 Details of experiment 8.A 240

B.12 Details of experiment 8.B 241

B.13 Details of experiment 8.C 241

B.14 Details of experiment 8.D 242

B.15 Details of experiment 8.E 242

B.16 Details of experiment 8.F 243

List of Figures

2.1 The Lorenz attractor with the traditional parameter values defined in appendix A.2.1. In the long term, all system states will lie on the system attractor. 16

2.2 Two indistinguishable states on the Ikeda map attractor. The red cross represents the true state x_0 whilst the red circle surrounding it represents the bound of the observational uncertainty, that is the area in which an observation can fall. The green cross and the circle surrounding it represents another state y_0 and its bound of uncertainty. Since the observation s_0, represented with the red star, lies in the overlapping region of the two uncertainty bounds, x_0 and y_0 are indistinguishable. 23

2.3 States on the Ikeda attractor coloured according to whether they are distinguishable or indistinguishable from the true state x_0 given an observation s_0. x_0 and s_0 are represented with a red cross and a red star respectively. States that are indistinguishable from the true state are coloured green. 24

2.4 Sets of indistinguishable states (coloured green) at time $t = 0$ given the number of observations stated in each panel. As more past observations are taken into account, the number of indistinguishable states diminishes. 26

2.5 A demonstration of how PDA ensembles are formed for a perfect model of the logistic map. The blue line represents sets of 3 consecutive states that are consistent with the model. The blue diagonal cross represents the final 3 points of the system trajectory which consists of a total of 16 time steps (the first 13 steps are not shown on the plot). The red point represents the final three observations (the point where PDA starts) which are assimilated to obtain the reference pseudo-orbit represented with the red cross. The black points are the result of adding random perturbations to the reference pseudo-orbit which are assimilated using PDA to obtain the green points which form the final initial condition ensemble. A close up view of the points is shown in figure 2.6. 32

2.6 A close up view of the points in figure 2.5. 33

3.1 Upper panel: A system density in which the mean, median and mode all differ. Lower panel: The mean squared error (blue line corresponding to the left y axis) and the mean absolute error (green line corresponding to the right y axis) between 1024 random draws from the system density in the upper panel and the forecast values on the x axis. In this case, both measures favour forecasts in which there is only low system density. 50

3.2 Upper panel: A time series of outcomes (black points) formed using equation 3.11 with $\sigma = 0.4$ where the climatological mean $\mu(t)$, represented with the black line, is governed by equation 3.13. Lower panel: The same time series linearly detrended. Linearly detrending fails to remove all of the effects of changes in the climatological mean. 58

3.3 The mean of the correlation coefficent between estimated anomalies of 1024 realisations of time series of forecasts and outcomes formed using equations 3.12 and 3.11 respectively on the interval [0,4] with the underlying trend described by equation 3.13 for different values of σ. As σ increases, the relative impact of the underlying trend falls but, since its effect is never removed completely, the ACC is always non-zero. This gives the misleading impression that the forecasts are skillful when, in fact, this is not the case. 59

3.4 Scatter plot of the standard deviation of daily temperatures against the ACC of our artificial forecasts for each month in the CET record. The blue line shows the expected ACC as a function of the standard deviation of the outcomes as described in equation 3.19. Although the expected forecast error remains constant, the ACC tends to increase as a function of the standard deviation of the observed temperatures. 63

3.5 A scenario in which the relationship between a set of 16 forecasts and outcomes is governed according to the behaviour of a logarithmic spiral. The x and y coordinates of the blue crosses represent the forecast values and the outcomes respectively. The red crosses, linked to the blue crosses, represent forecasts calibrated using linear regression. 66

4.1 Two shadowing trajectories of a one dimensional flow. The red circles represent 'noisy' observations whilst the red crosses represent the bounds of the noise. If a trajectory falls within the bounds of the noise, it is said to shadow the observation. The trajectories coloured blue and green, shown as solid when they are considered to shadow and dashed thereafter, shadow for 7 and 3 time steps respectively. . . 74

4.2 Upper panel: Mean shadowing lengths with 95 percent bootstrap resampling intervals for different values of the imperfection parameter c in experiment 4.A. Lower panel: Shadowing ratios of forecasts formed using the model with imperfection parameter c and those formed with imperfection parameter 4 for different values of c with 95 percent resampling intervals. Each measure gives a different perspective on the performance of the model with different degrees of imperfection. . . . 80

4.3 For experiment 4.B. Top left: Mean shadowing times of ensembles formed using inverse noise (blue), PDA with an assimilation window of 8 (red) time steps and PDA with an assimilation window of 16 time steps (yellow) as a function of ensemble size. Top right: Mean differences between shadowing times of ensembles formed using PDA with an assimilation window of 8 steps and inverse noise (blue) and PDA with assimilation windows of 16 and 8 time steps (red) as a function of ensemble size with 95% resampling intervals. Bottom: ensemble shadowing ratios between PDA with a window of 8 time steps and inverse noise (blue) and PDA with windows of 16 and 8 time steps (red) with 95% resampling intervals of the proportion as a function of ensemble size. For each given ensemble size, both the mean ensemble shadowing time and the shadowing ratio suggest that PDA performs better, on average, than inverse noise. Doubling the assimilation window, however, appears to make little difference to the performance of PDA in this case. 85

4.4 As a function of lead time: the mean ignorance score of forecast densities formed using forecasting system A (red line, left axis) with 95 percent resampling intervals, the blending parameter α (blue line, right axis) of forecasting system A and the proportion of forecast densities (with 90 percent confidence intervals) that boost probability for forecasting system A (solid green line, right axis) and B (dashed green line, right axis). Here, boosted probability identifies how, although the forecast densities formed using forecasting system B perform worse than forecasting system A in terms of the mean ignorance score, those formed using the former can sometimes be informative, whilst, for the latter, when $\alpha = 0$, this can never be the case. 89

5.1 A demonstration of how, although, on average, the dispersion of ensembles increases with lead time, there is significant variation in the dispersion at each one. Each blue point represents the standard deviation at a given lead time of a 64 member inverse noise ensemble evolved forward using a perfect model of the Lorenz '63 system. The red line represents the mean standard deviation over all 32 ensembles at each lead time. 97

5.2 Two sets of ensemble members on the attractor of the Lorenz '63 system evolved forward 3.2 days in Lorenz time. The ensemble members coloured green evolve over a section of the attractor in which there is little uncertainty and thus stay close together. The ensemble members coloured in red, on the other hand, evolve over a section of the attractor on which there is significant uncertainty as to whether a trajectory will remain on the same wing or move to the other and thus the ensemble standard deviation is much larger. 99

5.3 Kernel density estimates from samples of size 256 drawn from the logistic distribution with, top left: $s = 1$ and $\sigma = 0.6$, top right: $s = 5$ and $\sigma = 0.6$. Bottom left: $s = 1$ and $\sigma = 2.5$, bottom right: $s = 5$ and $\sigma = 2.5$. In each case, the dashed line represents the true distribution from which the sample was drawn and the green line represents the estimated distribution. The black crosses at the bottom represent the positions of the sample members. Each of the underlying distributions can be estimated well with a good choice of kernel width but this cannot be done when the kernel widths are constrained to be equal. 100

5.4 Histogram of the ignorance scores of forecast densities formed from ensembles in the test set in experiment 5.A using simple kernel dressing. Although most of the forecasts perform much better than climatology, some have very high ignorance scores. 103

5.5 The forecast density with the worst ignorance score of all those formed from ensembles in the test set of experiment 5.A using simple kernel dressing. The ensemble and the outcome are represented by the black crosses and the green circle respectively. The shape of the forecast density suggests that the kernel width is too small and hence the level of smoothing is insufficient. The ignorance in this case is particularly high because of the relatively large distance between the outcome and the nearest ensemble member to it. 105

5.6 The forecast density formed using simple kernel dressing from the ensemble with the highest standard deviation of all those in the test set in experiment 5.A. The ensemble and the outcome are represented by the black crosses and the green circle respectively. As in figure 5.5, it appears that the kernel width is too small. In this case, however, the outcome happens to lie much closer to the nearest ensemble member and hence the ignorance score is much better. 106

5.7 Scatter plot of the standard deviation of ensembles in the test set against the ignorance of the forecast densities formed using simple kernel dressing in experiment 5.A. Ensembles with high standard deviation are more likely to yield forecast densities that perform worse than climatology. 107

5.8 The mean ignorance as a function of the kernel width for the entire training set (top), G_1 (bottom left) and G_2 (bottom right) in K groups kernel dressing for the case when $K = 2$ for the ensemble-outcome pairs defined in experiment 5.A. In each case, the optimal kernel widths, represented with black vertical lines, are different. 110

5.9 Top: The standard deviation of ensembles in the training set (short dark blue lines) of experiment 5.A along with lines indicating the boundaries that determine which set of parameters should be used to dress a new ensemble for $K = 2$ (light blue), $K = 4$ (green) and $K = 8$ (red) in the K groups method. Bottom: Kernel widths chosen to dress a new ensemble with a given standard deviation for $K = 2$ (light blue), $K = 4$ (green) and $K = 8$ (red). 111

5.10 Top: Fitted kernel widths for ensembles in the test set of experiment 5.A that fall into each group using K groups kernel dressing with $K = 16$ (red circles) and simple kernel dressing (blue crosses). Bottom: Mean ignorance of forecast densities formed from ensembles that fall into each group using K groups method (red circles) and simple kernel dressing (blue crosses). The increased skill achieved from using K groups rather than simple kernel dressing comes mainly from forecast densities derived from the most dispersed and least dispersed ensembles. 114

5.11 The optimised mean ignorance over the training set (blue), the mean ignorance obtained using 2 fold cross validation (green) and the mean ignorance over the test set (red) as a function of K for the ensemble-outcome pairs defined in experiment 5.A. Here, it is clear how 2 fold cross validation fails to identify the optimal number of groups. 116

5.12 An example of how the subset of the training set is selected in fixed window kernel dressing for the case $M = 32$. The long black line represents the standard deviation of the ensemble to be dressed while the smaller lines represent the standard deviation of the ensembles in the training set. Those that are coloured red represent the ensemble-outcome pairs over which the parameter values are optimised. 118

5.13 The kernel width used to dress an ensemble with a given standard deviation for $M = 16$ (blue), $M = 128$ (green) and $M = 1024$ (red) using fixed window kernel dressing with the ensemble outcome pairs defined in experiment 5.A. 120

5.14 Forecast densities formed using dynamic (green) and simple (blue) kernel dressing from the ensemble in the test set of experiment 5.A which produces the highest ignorance score when simple kernel dressing is applied (as shown in figure 5.5). The black crosses along the x axis show the positions of the ensemble members and the green circle the position of the outcome. By taking account of the ensemble standard deviation, dynamic kernel dressing is able to apply a larger kernel width and hence yield a more informative forecast density. . . 124

5.15 Forecast densities formed using dynamic (green) and simple (blue) kernel dressing from the ensemble in the test set of experiment 5.A with the lowest standard deviation. The black crosses along the x axis show the positions of the ensemble members and the green circle the position of the outcome. The forecast density formed using simple kernel dressing places a large amount of density outside of the range of the ensemble members suggesting that the kernel width is too large. The kernel width obtained from dynamic kernel dressing, on the other hand, is smaller and hence the resulting forecast density appears to reflect the distribution of the ensemble members better. 126

5.16 Scatter plot of the ignorance of forecast densities formed from the test set of experiment 5.A using simple and dynamic kernel dressing. Each point is coloured according to its ensemble standard deviation with warmer colours indicating those with high standard deviation. The blue line indicates the points at which the ignorance of both methods would be equal. It is clear that most of the improvement in skill yielded from dynamic kernel dressing comes from ensembles with relatively low or relatively high standard deviation since these points are much more likely to lie above the blue line. 127

5.17 The mean ignorance of forecast densities formed from 8 member ensembles using simple (blue solid lines) and dynamic (green solid lines) kernel dressing averaged over 64 repeats for each training set size. The error bars represent 90 percent bootstrap resampling intervals. Using the same colour scheme, the dashed lines show the mean optimised mean ignorance for each training set size. For ensembles of this size, dynamic kernel dressing yields a significant improvement in ignorance even when the training set is very small. 131

5.18 The mean ignorance of forecast densities formed from 64 member ensembles using simple kernel dressing (blue solid lines) and dynamic kernel dressing (green solid lines) averaged over 64 repeats for each training set size. The error bars represent 90 percent bootstrap intervals. Using the same colour scheme, the dashed lines show the mean optimised ignorance for each training set size. Dynamic kernel dressing yields a significant improvement in forecast skill for training sets consisting of 32 ensemble-outcome pairs and more. For the smallest training set sizes, there is still some improvement but this difference is not significant for this ensemble size. 132

5.19 The kernel width that optimises the KL divergence (red dots), the kernel width fitted using dynamic kernel dressing (green crosses) and the kernel width fitted using simple kernel dressing (blue crosses) for each ensemble against its standard deviation in the perfect forecast scenario. By taking the ensemble standard deviation into account, dynamic kernel dressing gets much closer to recovering the optimal kernel widths. 136

5.20 Forecast densities formed from the ensembles in the test set with the largest (upper panel) and smallest (lower panel) standard deviation using simple (blue) and dynamic (green) kernel dressing. Also shown is the forecast density obtained by applying the kernel width that minimises the KL divergence (red). The black dashed line represents the system density from which both the ensemble and outcome are drawn. In both cases, dynamic kernel dressing gets much closer to recovering the system density. 138

5.21 Top left:the mean ignorance score of forecast densities formed using simple (blue) and dynamic (red) kernel dressing along with that of the forecasts formed using the optimal kernel width (green) as a function of ensemble size. Top right: The mean relative ignorance between forecasts formed using dynamic and simple kernel dressing (green) and of forecasts formed using the optimal kernel widths and dynamic kernel dressing (red). Bottom left:the mean KL divergence of forecast densities formed using simple (blue) and dynamic (red) kernel dressing along with that of the forecasts formed using the optimal kernel width (green) as a function of ensemble size. Top right: The mean difference in KL divergence between forecasts formed using dynamic and simple kernel dressing (green) and of forecasts formed using the optimal kernel widths and dynamic kernel dressing (red). 140

6.1 The mean ignorance score, expressed relative to that of the standard method, of 8192 forecast densities formed using the naive method (green stars) on multiple lead time ensembles where 32 ensemble members were launched at a lead time of 96 hours and another 32 at $96 + h$ hours for different values of h. The error bars represent 95 percent resampling intervals of the mean relative ignorance. The dotted line represents the mean relative ignorance achieved by setting $h = 0$ (i.e. doubling the ensemble size 96 hours ahead.) Since zero does not lie within the resampling intervals, when h takes a value of 54 hours or less, there is significant evidence that the naive method of combining forecast densities yields improved skill over the standard method. For larger values of h, there is no significant evidence of an improvement in skill whilst, for values of h of 78 hours and longer, there is significant evidence that the forecast densities perform worse on average than the standard method. 149

6.2 The same as figure 6.1 with the results of using the time-weighted method added (red stars). Applying this approach results in skillful forecast densities from larger values of h. Unlike those formed using the naive method, forecast densities formed using the time-weighted method are never expected to perform worse than those formed using the standard method. 153

6.3 The results of applying the Pure Bayes method in the Lorenz '63 imperfect model scenario of experiment 6.B. The black stars show the mean ignorance of the original forecasts whilst the blue lines show the mean ignorance using the Pure Bayes method where the starting point of each line indicates where the updating process was first applied. The colour of the points on the blue lines indicate whether the Pure Bayes method performs significantly worse than (red), better than (green) or not significantly different to (yellow) the original forecasts. Although there appears to be some benefit in using the climatology as a prior, in general, the Pure Bayes method appears to be counterproductive in this scenario. 161

6.4 The results of applying the Pure Bayes method in the Lorenz '63 perfect model scenario of experiment 6.B. The black stars show the mean ignorance of the original forecasts whilst the blue lines show the mean ignorance using the Pure Bayes method where the starting point of each line indicates where the updating process was first applied. The colour of the points on the blue lines indicate whether the Pure Bayes method performs significantly worse than (red), better than (green) or not significantly different to (yellow) the original forecasts. In this scenario, in many cases, the Pure Bayes method yields significantly improved forecasts whilst using the climatology as a prior, on the other hand, appears to be counterproductive. 164

6.5 Mean ignorance scores of sequential blending expressed relative to that of the standard method with 95 percent bootstrap resampling intervals of the mean in the Lorenz '63 PMS. Since sequential blending yields a lower mean ignorance score and the resampling intervals do not contain zero, there is significant evidence that this approach performs better than the standard method in 8 out of 9 lead times considered. 167

6.6 Mean ignorance scores of sequential blending expressed relative to that of the standard method with 95 percent bootstrap resampling intervals of the mean in the Lorenz '63 IMS. Since sequential blending yields a lower mean ignorance score and the resampling intervals do not contain zero, there is significant evidence that this approach performs better than the standard method in 7 out of 9 lead times considered. 169

7.1 Top panel: Mean shadowing length of model trajectories formed using PDA with an exact (blue) and an approximate (red) Jacobian as a function of the assimilation window. The error bars represent 95% resampling intervals of the mean in each case. Since the means are so close, it is difficult to distinguish the lines. Middle panel: The mean pairwise difference between the shadowing lengths of model trajectories formed using PDA with an exact and an approximate Jacobian with 95% resampling intervals of the mean difference. Lower panel: Shadowing ratios of model trajectories formed using an exact and approximate Jacobian. The error bars represent 95% resampling intervals of the shadowing ratio. Since the mean difference between the shadowing lengths is not significantly different from zero and the shadowing ratio is not significantly different from 1, both measures suggest a lack of a difference between the performance of each approach. 175

7.2 The results of assimilating using PDA without rescaling the variables first. For each variable, the black lines are the true trajectory, the red dots the observations and the green lines are the analysis. Although PDA has found a pseudo-orbit that stays close to the observations of the x,y,X and Y variables, this is not the case for the Z variable. . . 179

7.3 The same as figure 7.2 but with the modified version of PDA in which the variables are rescaled. This time, the pseudo-orbit stays close to the observations of all of the variables. 181

7.4 Points on the attractor of the Rössler map coloured according to the length of the linear regime of a trajectory initialised at that point. Darker colours imply a shorter linear regime (as indicated on the colour bar) and hence higher nonlinearity. Note how the nonlinearity tends to be highest for trajectories initialised close to where the attractor splits, either continuing around the base of the attractor or up in the z direction. For trajectories initialised on this part of the attractor, the linear regime tends to be less than 100 time units. . . . 186

7.5 Top left panel:the mean length of the linear regime as a function of F. Top right: the mean shadowing lengths achieved using PDA (blue), 4DVAR (red) and when no data assimilation is used (orange) as a function of F. The error bars represent 95 percent resampling intervals of the mean. Bottom left: the mean difference in shadowing lengths between PDA and 4DVAR (green) and between 4DVAR and when no data assimilation is used (red) as a function of F. The error bars represent 95 percent resampling intervals of the mean difference in shadowing length. Botttom right: Shadowing ratios between PDA and 4DVAR (green) and between 4DVAR and the observations (red) as a function of F. The error bars represent 95 percent resampling intervals of the shadowing ratio in each case. 187

8.1 The standard deviation (blue points) along with the mean (black solid line) and the median (black dashed line) standard deviation at each lead time of the ensembles in experiment 8.A. Although, on average, the standard deviation of the ensembles increases with lead time, there is significant variation at each one. 194

8.2 The mean ignorance scores of forecast densities formed using simple (blue) and dynamic (green) kernel dressing for the 10 different lead times in experiment 8.A. The error bars represent 95 percent bootstrap resampling intervals of the mean ignorance. At most lead times considered, dynamic kernel dressing yields more skillful forecast densities, on average, than simple kernel dressing. 195

8.3 The standard deviation (blue points) along with the mean (black solid line) and the median (black dashed line) standard deviation of the ensembles at each lead time in experiment 8.B. Although, on average, the standard deviation of the ensembles increases with lead time, there is significant variation at each one. 197

8.4 The mean ignorance scores of forecast densities formed using simple (blue lines) and dynamic (green lines) kernel dressing for each of 10 different lead times in experiment 8.B. For the shortest lead times, dynamic kernel dressing improves forecast skill whilst for longer lead times, very little improvement is made over simple kernel dressing. . . 198

8.5 Mean ignorance of forecast densities formed using dynamic kernel dressing relative to the ignorance of forecast densities formed using simple kernel dressing as a function of lead time for experiment 8.A (blue), in which ensembles are formed using an imperfect model of the Henon map, and experiment 8.B (red), in which ensembles are formed using an imperfect model of the Lorenz '96 system. The fact that zero does not fall within the resampling bars suggests that dynamic kernel dressing performs significantly better than simple kernel dressing at many of the lead times considered in these two examples. 200

8.6 Proportion of instances in which probability is boosted as a function of lead time using dynamic (solid lines) and simple (dashed lines) kernel dressing with inverse noise (blue) and PDA ensembles (red) on the ensembles formed from the perfect model of the Lorenz '63 system in experiment 8.C. The error bars represent 95 percent confidence intervals of each proportion at a subset of lead times. 203

8.7 Top: The mean ignorance scores obtained by applying the Pure Bayes method (blue lines), the time-weighted method (red crosses) and sequential blending (cyan diagonal crosses) for the imperfect model of the Moore-Spiegel system in experiment 8.D. The black circles represent the mean ignorance scores of the original forecast densities. Bottom left: The mean ignorance scores of the Pure Bayes method with coloured dots indicating whether the forecasts are significantly than (better), worse than (red) or not significantly different (yellow) from the original forecasts. Bottom right: Mean relative ignorance scores between sequential blending and the original forecasts (blue), the time weighted method and the original forecasts (red) and the time weighted method and sequential blending (orange) with 95% resampling intervals of the mean difference. Whilst the Pure Bayes method tends to yield forecast densities that are significantly worse than the original forecasts, the time-weighted method and sequential blending yield improved forecast densities in around half of the lead times considered. 206

8.8 Top: The mean ignorance scores obtained by applying the Pure Bayes method (blue lines), the time-weighted method (red crosses) and sequential blending (cyan diagonal crosses) for the perfect model of the Moore-Spiegel system in experiment 8.D. The black circles represent the mean ignorance scores of the original forecast densities. Bottom left: The mean ignorance scores of the Pure Bayes method with coloured dots indicating whether the forecasts are significantly than (better), worse than (red) or not significantly different from (yellow) the original forecasts. Bottom right: Mean relative ignorance scores between sequential blending and the original forecasts (blue), the time weighted method and the original forecasts (red) and the time weighted method and sequential blending (orange) with 95% resampling intervals of the mean difference. In this case, for shorter lead times, the Pure Bayes method tends to yield forecast densities that are significantly better than the original forecasts whilst at longer lead times they tend to be worse. Both the time weighted method and sequential blending are able to yield significantly improved forecasts for all but the two shortest lead times. 208

8.9 The mean proportion of ensemble members that shadow the observations (blue line, left axis) and the mean ignorance score obtained by fitting Gaussian distributions (red starred line, right axis) to the ensemble members as a function of forecast lead time for the ensembles in experiment 8.E. The error bars represent 95% resampling intervals of the mean ignorance. Points that lie below the dashed line indicate forecast skill with respect to climatology. For lead times greater than 6 time steps, a large proportion of ensemble members shadow whilst the forecast densities formed using this approach perform worse, though not generally significantly, than climatology. This suggests that a better approach to forecast density formation could be taken. 210

8.10 The mean proportion of ensemble members that shadow the observations (blue line, left axis) and the mean ignorance score obtained by fitting Gaussian distributions with (magenta starred line, right axis) and without (red starred line, right axis) blending as a function of forecast lead time. The error bars represent 95% resampling intervals of the mean ignorance. Forecasts formed without blending only yield significant skill with respect to climatology up to 6 steps ahead whilst blending increases this time to 13 steps. 212

8.11 The mean proportion of ensemble members that shadow the observations (blue line, left axis), the mean ignorance score obtained by fitting Gaussian distributions with (magenta starred line, right axis) and without (red starred line, right axis) blending and the mean ignorance score obtained using dynamic kernel dressing (green starred line, right axis) as a function of forecast lead time in experiment 8.E. The error bars represent 95% resampling intervals of the mean ignorance. Dynamic kernel dressing makes no assumption about the underlying distribution and thus yields better forecast skill than the Gaussian approach. 213

8.12 The mean proportion of ensemble members that shadow the observations (blue line, left axis), the mean ignorance score obtained by fitting Gaussian distributions with (magenta starred line, right axis) and without (red starred line, right axis) blending and the mean ignorance score obtained using dynamic kernel dressing (green starred line, right axis) as a function of forecast lead time in experiment 8.F. The error bars represent 95% resampling intervals of the mean ignorance. For the first 9 lead times, the approach of fitting Gaussian distributions without blending yields forecast densities that perform significantly better than climatology. The fact that a small proportion of ensemble members shadow longer than this time suggests that informative forecast densities may be found. This is confirmed by the performance of the forecasts formed using dynamic kernel dressing which yield better skill than the climatology up to 15 steps ahead. . . 215

A.1 The Henon map attractor with parameter values $a = 1.4$ and $b = 0.3$. 224

A.2 The Duffing map attractor with parameter values $a = 2.75$ and $b = 0.2$. 226

A.3 The Ikeda map attractor with parameter values $\alpha = 6$, $\beta = 0.4$, $\gamma = 1$ and $u = 0.83$. 227

A.4 The Lorenz attractor with parameter values $\rho = 28$, $\sigma = 10$ and $\beta = \frac{8}{3}$ 229

A.5 The Moore Spiegel attractor with parameter values $R = 100$ and $T = 35$. 230

A.6 The Rössler attractor with parameter values $a = 0.1$ and $b = 0.1$ and $c = 14$. 234

Chapter 1

Introduction

It has long been desirable for human beings to make predictions of the future. Early attempts at predicting the weather are known to have been made by Aristotle in around 340 B.C. [158] whilst some suggest that important battles have been won or at least significantly shortened on the basis of successful weather predictions (Meteorologists successfully predicted a short break in the stormy conditions that were occurring as the allies were planning the D-Day landings which allowed the operation to go ahead [1]).

Although it was recognised early on that weather conditions in one area could sometimes be used to inform prediction of future weather patterns in another, such methods were severely limited due to the fact that communication, whether over land or by sea, was slow. This, however, changed with the invention of the electric telegraph in 1835 which meant that weather conditions could be communicated almost instantaneously [138] and thus the modern era of weather forecasting began. After a major storm in 1859, which caused the loss of the Royal Charter, 15 land stations

were set up so that reports of the weather could be submitted to experts and hence prior warnings of gales could be made to ships at sea. Over time, more stations were set up so that regular weather predictions, based on the knowledge of experts, could be disseminated [62].

The origins of numerical weather prediction techniques can be traced back to the works of Lewis Fry Richardson who took the (now very widely used) approach of representing the natural processes of the atmosphere as a set of differential equations and solving them using a straightforward numerical integration technique [102]. He used his method to make forecasts of changes in the pressure and the wind at two points over central Europe. In practice, however, numerical integration was highly laborious and time consuming. It was not until the advent of computing that forecasts could be made quicker than real time. The first numerical simulations of weather forecasts were made in 1950 [24] whilst the first regular, and hence operational, forecasts were made in Sweden in 1954 [64]. Subsequently, operational numerical forecasts became operational in countries around the world. [64]

Although Richardson's first forecast was inaccurate, this allowed him to identify some of the challenges of numerical weather forecasting. He found that calculating useful estimates of the initial state of the atmosphere[1] could be difficult. The implications of this were further demonstrated in 1963 when Ed Lorenz showed, using a simple 3 dimensional model, that two simulations of a system with slightly different initial conditions will quickly diverge from each other [99]. As a result, he suggested that informative point predictions of the future state of the atmosphere can only be made up to around two weeks ahead. The true initial state of the atmosphere can never be found due to measurement error. It therefore became clear that

[1]Throughout this thesis, we distinguish model states and model state space from those of the system.

a deterministic forecast does not adequately reflect this uncertainty. Epstein [45] recognised this and proposed a dynamical model to predict the mean and variance of the distribution of possible future states. Whilst such Monte Carlo approaches were shown to be useful, it was suggested by Leith [95] that adequate representations of the future states of the atmosphere could only be made using multiple point forecasts with initial states sampled from the uncertainty in observations of the current state. Such techniques are known as ensemble forecasts. Ensemble forecasts were introduced operationally both by the European Centre for Medium and Mid Range Forecasting (ECMWF) [105] and by the National Centers for Environmental prediction (NCEP) [151] in 1992 and soon many other forecasting centres followed suit [33, 2]. Whilst the original aim was merely to obtain a measure of likely forecast error, this paved the way for the formation and dissemination of probabilistic forecasts, in which the future states of a system are represented with probabilities or probability density functions. Such forecasts are desirable since they communicate information regarding the uncertainty in a forecast.

Probabilistic forecasts are in widespread use in many fields. For example, in weather prediction, precipitation forecasts expressed as probabilities have long been disseminated to the public [113, 51] whilst much emphasis is placed on the creation of probabilistic forecast densities of continuous variables such as the temperature. In climatology, projections of the future state of the climate are almost always probabilistic in nature [116] whilst probabilistic forecasting techniques are used in areas as diverse as population forecasting [159], economics [20] and ecology [9].

In forecasting, the aim is to predict the future states of a system. We will focus on cases where this is done by simulating its evolution over time using some mathematical representation called a model. In practice, the ability to make effective forecasts

is always limited since, in the real world, any model will contain structural errors and therefore can only be considered to approximate the processes inherent in the system.

In this thesis, we focus on the following four aspects of the forecasting process:

1. Initial state estimation
2. Ensemble formation
3. Forecast density formation
4. Forecast Evaluation

We now briefly describe these and the contributions made in this thesis to each.

Initial State Estimation

Chaotic dynamical systems, by definition, are sensitive to initial condition uncertainty. Since observations of system states are always subject to measurement error, a forecast trajectory will always diverge from the true trajectory even when the model describes the system perfectly. Data assimilation techniques combine sets of past and present observations with the model to attempt to find improved estimates of the initial state. In section 4.1 of this thesis, we utilise an existing approach to forecast evaluation called shadowing and introduce a new method called shadowing ratios. In section 7.1, we use these methods in a number of experiments in which we compare the performance of different approaches to data assimilation.

Ensemble Formation

Ensemble forecasts are formed by sampling additional states around an observation or a 'best guess' of the initial state to form a set of initial conditions and using the model to evolve them forward in time to form a set of distinct model simulations which reflect the uncertainty/sensitivity of that particular forecast. The approach taken to the sampling of the initial conditions, however, is important and can have a large effect on the quality of the final forecast density. Generally, initial states that are consistent both with the observation and the model dynamics can be expected to perform better than those that are consistent only with the observation [81]. In section 4.2.1 and 8.2, we compare the performance of two ensemble formation techniques both in the context of the performance of the raw ensembles and forecast densities derived from them.

Ensemble Interpretation

Forecast densities are formed by converting an ensemble into a probability density function. Although on the face of it, this appears like a simple density estimation problem, in fact, it differs significantly since model error and other imperfections in the forecasting framework imply that the outcome, the observed system value, will not be drawn from the same distribution as the ensemble. This means that fitting standard parametric or non-parametric distributions to the ensembles cannot usually be expected to be effective in estimating the distribution of the true state. An existing approach to this problem, called kernel dressing [131, 22], takes account of this difference by optimising the performance of a set of forecast densities over an archive of past ensembles and outcomes. Commonly, an approach, which we refer to as simple kernel dressing, is taken which assumes that each of the ensembles have

a similar level of dispersion [143, 146]. In chapter 5, however, we show that this can yield systematically inferior forecast densities when this is not the case. We propose two new approaches to kernel dressing and consider an existing approach in which the varying level of dispersion in a set of ensembles is accounted for. We demonstrate that taking such approaches can improve the performance of sets of forecast densities.

In weather forecasting, ensembles are usually launched every 6 or 12 hours. Often, when a new ensemble is launched, it is used to form a new forecast density and the previous ensemble and forecast density are discarded. In chapter 5, we investigate whether the most recently launched ensemble can be combined with those previously launched to yield better forecast densities. Whilst one might have expected a Bayesian approach to be appropriate in this setting, we demonstrate, using examples, that it can, in fact, be counterproductive. It appears that flaws in the forecasting system, including model error, violate the common assumptions underlying Bayes theorem; otherwise that approach would be effective. We introduce two new approaches to the combination of ensembles launched at different lead times and demonstrate that they can be used to construct forecast densities that perform significantly better than those formed with ensembles launched at a single lead time, even when the model is imperfect.

Forecast Evaluation

Forecast evaluation is the process of assessing the performance of a set of forecasts. At this stage we wish to gain an understanding of their value. In probabilistic forecasting, this is usually done using a function of the forecast and the outcome called a scoring rule, whilst similar scores exist for point forecasts. The choice of

scoring rule is extremely important since, if it fails to identify the best forecast, it is of little use. Scoring rules also play an important role in the formation of forecast densities. Given a slowly increasing forecast-outcome archive, it is common for kernel dressing to use an independent set of ensembles and outcomes, called a training set, to train a set of forecast system parameters to optimise the performance with respect to a scoring rule.

In this thesis, we propose alternative approaches to forecast evaluation. A forecast evaluation method is optimised when the forecast that yields its best possible score is found. For point forecasts, distance based measures such as the mean squared error and the mean absolute error say little about their value since they are often optimised by the mean or the median of the distribution from which the outcome is drawn even if the probability of drawing such values is low. In chapter 3, we show that attempting to apply another common approach to the evaluation of point forecasts, the anomaly correlation coefficient skill score [4, 112, 111] can be misleading in practice and therefore should be treated with caution. We discuss an approach to forecast evaluation called shadowing and, capitalising on this concept, introduce a new approach called shadowing ratios.

It is useful to be able to assess various aspects of the ensembles themselves separately from the forecast (density). This means they can be assessed independently of forecast density formation and other interpretation techniques. In section 4.2.1, we introduce a new approach to the evaluation of ensembles, called ensemble shadowing ratios, in which the length of time the best performing member of each ensemble stays close to the observations is compared.

Scoring rules provide a useful measure of the performance of a set of forecasts [19, 44, 58, 130]. The mean of a set of scores, however, does not tell the entire story.

In section 4.3.1, we introduce a broader view of evaluating the performance of sets of forecast densities, called boosted probability, in which the period of time into the future they can be considered to be more informative than the distribution of the long term behaviour of the system is measured. This is not intended as an alternative to scoring rules, rather that boosted probability can be used alongside such measures to inform a forecaster about properties of the distribution of such scores.

1.1 Thesis Structure

It should be noted that, although each the research questions relate to the prediction of nonlinear systems, this thesis can be considered a collection of papers related to this theme rather than tackling an overarching research question. This thesis is structured in the following way:

In chapter 2, we provide background information on the process of forecasting dynamical systems and an overview of the steps involved. We focus on the individual steps of the forecasting framework, detailing the methods drawn upon in this thesis. Whilst the presentation of some of the material is new, there are no novel contributions in chapter 2.

In chapter 3 we turn our attention to the evaluation of point forecasts. We review some shortcomings of common evaluation techniques before focusing on the anomaly correlation coefficient skill score (ACC). We argue that this measure should be treated with extreme caution and that a very specific set of assumptions should be satisfied for its deployment to be valid.

In chapter 4, we focus on the evaluation of the two types of forecast discussed in

this thesis: point forecasts, which consist of a single estimate of the future and probabilistic forecast densities, which take the form of predictive probability density functions. In addition, we consider the evaluation of ensembles, which consist of multiple point forecasts initialised with slightly different values. We introduce a new approach to the evaluation of forecast trajectories, called shadowing ratios, which measures forecast performance based on the length of time forecast trajectories stay close to or 'shadow' a set of observations. We extend this approach to introduce a new method, called ensemble shadowing ratios, aimed at evaluating the performance of ensembles. Finally, we introduce a new approach to the evaluation of a sequence of forecast densities, called boosted probability, which provides an alternative view to regular methods of probabilistic forecast evaluation.

In chapter 5, we focus on the formation of forecast densities from ensembles. We show that, when there is significant variation in the dispersion of ensembles resulting from nonlinearities in the system, simple kernel dressing fails to provide useful forecast densities in some cases. We thus propose two new methods and apply another existing method aimed at correcting this problem, showing that each one tends to yield, on average, more informative forecast densities. We then show that, when the ensemble and the outcome are drawn from the same distribution, one particular approach which we refer to as dynamic kernel dressing, performs much better in terms of how closely forecast densities approximate this distribution.

In chapter 6, we address the less commonly considered question of whether ensembles evolved forward to the same target, but launched at different times, can be combined to yield better forecasts. We argue that, as the time between ensemble launches decreases, it is necessarily the case that the ensembles can be combined to yield better forecast densities. We demonstrate this using a simple example and show that this

approach can be improved upon by allowing ensembles launched at different times to be weighted differently. We call this approach the time-weighted method. We then investigate how forecast densities, rather than ensembles, launched at different lead times can be combined to yield improved forecast skill. The Bayesian approach is the correct way of combining sequentially formed probabilities. We show, however, that, in practice, the Bayesian approach is always likely to be sub-optimal due to the effects of model error and other imperfections in the forecasting system.

Finally, we introduce a new approach called sequential blending, which, like the time-weighted method, is shown to be effective in improving forecast skill, even when the model is imperfect. With this method, forecast densities are formed by sequentially updating forecasts by extending the blending approach introduced in [22].

In chapter 7, we perform three experiments, using the approach of shadowing ratios introduced in chapter 3, to make comparisons between data assimilation techniques. In the first, we compare the performance of an approach called Pseudo-orbit data assimilation (PDA) when the required gradient information required for the algorithm is known and when it is replaced with an approximation. In the second experiment we show how PDA can fail when there are large differences in the variability of the variables it is applied to. We show how the algorithm can be altered to allow for this variability and that its performance is improved as a result. In the third experiment, we make a direct comparison between the performance of PDA and another data assimilation technique, called 4DVAR, in an environment in which the level of nonlinearity in the system is varied. We show that, the more nonlinear the system, the bigger the improvement that can be made by applying PDA rather than 4DVAR.

In chapter 8, we perform a number of experiments with imperfect models using the

approaches introduced in the preceding chapters. First, we turn our attention back to dynamic kernel dressing. In chapter 4, we show that this approach is capable of improving the skill of forecast densities in the perfect model scenario. In this chapter, we demonstrate, using two different imperfect models, that this approach can also be found to improve forecast skill in the imperfect model scenario.

Next, we make use of the boosted probability approach introduced in chapter 3 to compare the performance of forecast densities formed using different ensemble and density formation techniques. We show that PDA can outperform another approach called inverse noise. Moreover, we show that the dynamic kernel dressing method introduced in chapter 5 can yield further improved forecast densities compared with simple kernel dressing.

We then demonstrate that the results found in chapter 6, in which different approaches to combining forecasts are investigated, can also be found in a different dynamical system. We argue that this gives us confidence that our results are likely to apply more generally.

Finally, we demonstrate the link between shadowing and forecast skill. We show that the period of time in which an ensemble of forecast trajectories shadow the observations can indicate whether skillful forecast densities can be expected to be found. We demonstrate how this approach would lead us to look for better forecast density formation techniques in the case when our forecast densities consist of Gaussian distributions.

A list of the symbols used in this thesis is given on page 219. Details of dynamical systems and models are shown in appendix A and details of experiments are given in appendix B.

Chapter 2

Background Theory

In this chapter, we describe background theory to the processes used in forecasting and some of the concepts capitalised upon in this thesis. This chapter contains no novel material beyond its presentation.

In section 2.1, we define forecasting and describe the fundamental differences between point forecasts, ensembles and probabilistic forecasts. We then describe our forecasting framework, a series of steps that make up the forecasting process. Next, we explain the important distinction between two forecasting scenarios used in this thesis: the perfect model scenario [140, 40], in which the underlying system processes are understood perfectly by a forecaster, and the imperfect model scenario, in which the model contains some level of structural uncertainty.

In section 2.2, we describe the theory of indistinguishable states [140, 141] and explain how this leads to the conclusion that, in the presence of observational noise, the true initial state cannot be distinguished from other model states. This forms the basis of the argument that forecasts should take the form of probability density

functions rather than single values.

In section 2.3, we explain the need for current state estimation techniques, or data assimilation [161, 40, 48]. We explain how data assimilation can be expected to improve forecast performance. We then describe two particular data assimilation techniques, pseudo-orbit data assimilation [80, 82, 41, 40] and 4DVAR [67, 68, 125], the performance of which we compare in chapter 7.

In section 2.4, we describe some of the methods used in the formation of probabilistic forecasts. We explain the rationale behind ensembles and describe two approaches to their formation. We then describe a number of approaches to ensemble interpretation in which ensembles are used to form forecast densities[129, 22].

2.1 Forecasting

Predicting the future evolution of any physical system is restricted by various types of uncertainty [99, 141]. To make predictions of future states, we attempt to build mathematical models of the underlying behaviour of a system. Since the complex rules of nature can never be understood perfectly, however, our models can only approximate the processes inherent in a system. Nevertheless, provided the model reproduces at least some of the key processes, we can often gain some understanding of how states may evolve. This process is called *forecasting* [10].

2.1.1 Dynamical Systems

A *dynamical system* is a set of rules governing the time dependence of a set of *states* $\boldsymbol{x} \in \mathbb{S}^d$ in a geometric space where $\boldsymbol{x}_t$ is the state of the system at time t and $\mathbb{S}^d$ denotes a d dimensional *state space*. The rules that govern the evolution of such a

system are called the *dynamics*. A dynamical system can be written in the form $\boldsymbol{x}_t = F^t(\boldsymbol{x}_0)$, where F denotes the dynamics of the system. $\boldsymbol{x}_0$ is called the *initial state* or *initial condition* of the system.

Dynamical systems can be divided into two different types, stochastic and deterministic. In deterministic systems, a state and the dynamics define all future states unambiguously whilst stochastic systems include some random element in the dynamics which means that two trajectories with different initial conditions can evolve differently over time. We consider only deterministic dynamical systems. Throughout this thesis, we assume that the model and the system share the same state space thereby avoiding issues of subtractability [100, 142].

2.1.2 Maps and flows

Dynamical systems can evolve either in discrete or continuous time. A system that evolves in discrete time is called a *map* and is defined by

$$\boldsymbol{x}_{i+1} = F(\boldsymbol{x}_i), \tag{2.1}$$

where $i \in \mathbb{Z}$. The dynamics F govern the evolution from one discrete time step to the next. A dynamical system that evolves in continuous time is called a *flow* and is usually described by a set of ordinary differential equations in the form

$$\frac{d\boldsymbol{x}(t)}{dt} = F(\boldsymbol{x}), \tag{2.2}$$

defined for all $t \in \mathbb{R}$ giving an unbroken continuous trajectory $\boldsymbol{x}_t$ for $t \in (0, T)$. Often, there is no analytical solution to the equations and thus the system states are found using some numerical integration scheme. Throughout this thesis, we use

a 4th order Runge-Kutta [91] integration scheme to **define** the *system*.

2.1.3 System-model pairs

Prediction of dynamical systems requires a forecasting *model*. A model attempts to give a mathematical description F, called the *model dynamics*, of the *system dynamics* $\tilde{F}$. Since it is impossible to perfectly describe the complex laws of nature, in real world forecasting problems, F and $\tilde{F}$ are necessarily distinct. We refer to the pair of dynamical systems governed by $\tilde{F}$ and F as a system-model pair. Details of the system-model pairs used in this thesis can be found in appendix B. An observation of a forecasted system state is called an *outcome*. We refer to a set of forecasts and corresponding outcomes as a forecast-outcome archive.

2.1.4 Chaos

Many real world dynamical systems are highly sensitive to initial condition uncertainty. This sensitivity is known as *chaos* [99] and is most famously known to occur in weather and climate but is also claimed to exist in a diverse set of fields such as psychology [72], economics [86] and ecology [104]. A common description of chaos, coined by Edward Lorenz, is that 'the flap of a butterfly's wings in Brazil can cause a tornado in Texas' [43], the implication being that a small and seemingly insignificant difference in the initial condition can lead to large differences later on. Chaos makes the task of prediction more difficult even when the model and the system dynamics are identical since small errors in the initial conditions will **eventually** cause a model trajectory to diverge from the system trajectory. Observations are always clouded by measurement error and thus the true initial condition can only be estimated [140, 141], making perfect prediction of future states impossible.

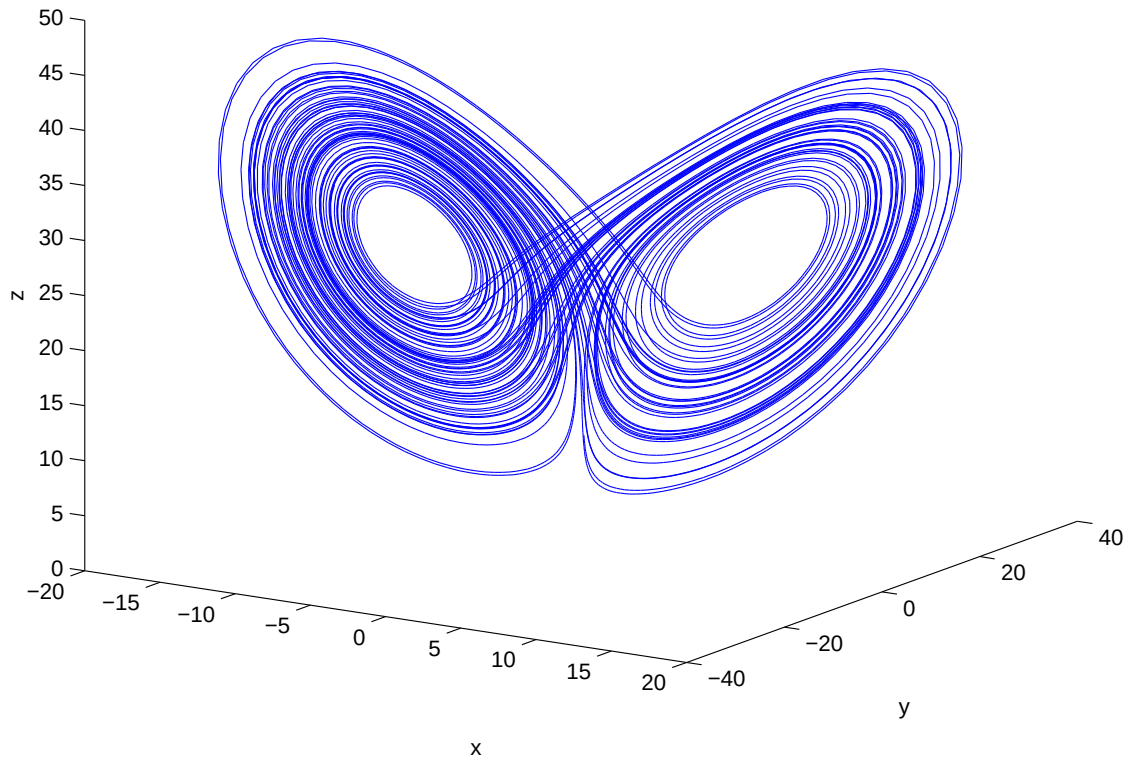

Figure 2.1: The Lorenz attractor with the traditional parameter values defined in appendix A.2.1. In the long term, all system states will lie on the system attractor.

Initial states of dynamical systems will tend to evolve over time towards a set of states called an *attracting set*. The collection of states in the attracting set is called the *attractor*. States that do not lie within the attracting set are called *transient* states. Transient states can be considered to be inconsistent with the long term behaviour of the system and therefore, in this thesis, the system states are assumed to be non-transient. The famous Lorenz '63 [99] attractor is shown in figure 2.1.

2.1.5 Point and Probabilistic forecasting

Forecasts can either consist of single values or be probabilistic in nature. For example, a forecaster might predict rain tomorrow or that the temperature will be 27 degrees Celsius. A *point* is defined as a particular single state. A forecast consisting of a single point is called a *point forecast*. Probabilistic forecasts, on the other hand, assign probabilities to different outcomes. For example, a forecaster might estimate a 90 percent chance of rain. Probabilistic forecasts of continuous variables consist of a probability density function called a *forecast density*.

2.1.6 Forecasting framework

The process of forecasting, as considered in this thesis, is conducted in a number of stages which we refer to as our *forecasting framework*. We briefly describe each of these stages below. A specific configuration of the forecasting framework is called a *forecasting system*.

Collection of observations

The collection of observations of past and present states is an important part of the forecasting framework. Observations play a crucial role both as inputs for current state estimation techniques and tools with which to evaluate our forecasts. In the real world, an observation can only be expected to approximate the underlying state due to the inevitable presence of observational noise. A *noise model* is the distribution from which realisations of observational error are drawn. Throughout this thesis, we define a *noise level* to be the standard deviation of the observational noise given as a percentage of the standard deviation of the long term behaviour of

the system[1].

Current state estimation

Current state estimation techniques combine sets of past and present observations with a model to attempt to find an improved estimate of the initial state. This process is called *data assimilation* [161, 40, 48]. Data assimilation techniques usually aim to find initial conditions that are consistent with both the observation(s) and the model dynamics. We discuss these in more detail in section 2.3.

Ensemble formation

Whilst data assimilation techniques can often be expected to find a more useful initial state for a model, we can never expect it to coincide with the underlying system state and thus some initial condition uncertainty will always be present. Therefore, instead of evolving a single model simulation, *ensemble* methods [96, 105, 151] attempt to account for initial condition uncertainty by running multiple simulations of the same model, each one with a slightly different initial condition. In point forecasting, this stage is usually omitted. We discuss ensemble formation in more detail in section 2.4.1.

Generation

At the generation stage, one or more initial conditions are evolved forward for a fixed period of time into the future called a *lead time*. If more than one initial condition is used, the resulting set of model trajectories is called an *ensemble*.

[1]The distribution of the long term behaviour of the system is referred to as the climatology. We formally define the climatology in section 2.4.6

Removal of biases

Forecasts often contain inherent systematic biases which can be detected using pairs of forecasts and outcomes. These biases can either come from the limitations of the model or from variations within the area being forecast. For example, in weather forecasting, predictions of variables such as the temperature usually apply to fixed grid boxes on or above the Earth's surface. Observed values are always likely to differ within each grid box depending on the topography of the local area and hence the forecast may consistently overestimate in some areas of the grid box and underestimate in others. At this stage, removal of such biases is attempted using sets of historical forecasts and outcomes.

Ensemble Interpretation

At this stage, ensembles are used to form predictive forecast densities. This stage is omitted when the desired outcome is a point forecast. We discuss ensemble interpretation in more detail in section 2.4.3.

Forecast evaluation

Forecast evaluation is the process of assessing the performance of a set of forecasts. This is a key part of the forecasting framework. If sets of forecasts are generated, each using a different forecasting system, a reliable way of evaluating them is imperative in choosing which, if any, is most useful in forming future forecasts. Forecast evaluation can also play an important role in forming forecast densities as we explain in section 2.5.7. We discuss forecast evaluation in more detail in section 2.5.

2.1.7 Perfect and Imperfect model scenarios

In this thesis, we make use of both the perfect model scenario (PMS) and the imperfect model scenario (IMS). These two concepts differ in the level of understanding a forecaster has of the underlying dynamics of a system. Both scenarios are defined below.

Imperfect model scenario

In the imperfect model scenario (IMS), the forecaster's model provides only an approximate representation of the underlying system dynamics. In fact, every real world forecasting problem falls into this category since, in practice, it is impossible for a model to reproduce all of nature's processes perfectly. In the IMS, the existence of initial condition uncertainty is also assumed since real world measurements are always clouded in observational error, whether this be from finite precision measuring devices, rounding error, noise, human error, other, or a combination of these factors.

Perfect model scenario

In the perfect model scenario (PMS), it is assumed that the mathematical structure of the model is identical to that of the system. In the PMS, however, a forecast trajectory can only be perfect if the exact initial condition is known. Arguably, all forecasting problems in the real world fall into the imperfect model scenario and hence the PMS can only be constructed by generating system states using the model.

2.2 Indistinguishable States

With a perfect model of the system dynamics and a perfect initial condition, an exact point forecast of any future state of a deterministic system can be made. When an observation of the initial condition is clouded in measurement error, however, multiple model states will exist that are consistent with an observation given its error distribution and thus are indistinguishable from the true state. Such states are known as *indistinguishable states* [140, 141]. Under the perfect model scenario, evolving any of these states forward in time represents a plausible, but not necessarily equally likely, scenario of the future given the information contained in the observation.

We now describe the theory of indistinguishable states as introduced in [140] and [141]. Although we do not directly use indistinguishable states in this thesis, the theory provides useful justification for why probabilistic forecasts are more appropriate than point forecasts for nonlinear systems.

Let s_t represent an observation of the true state x_t and let y_t represent some other state. Let the probability density function of the observational noise be $\rho(.)$. The joint probability density function that an observation s_t renders x_t and y_t indistinguishable is thus given by

$$\int \rho(s_t - x_t)\rho(s_t - y_t)ds_t. \tag{2.3}$$

Making the substitutions $b = x_t - y_t$ and $z = s_t - x_t$, so that b represents the distance

between the two states and z is the actual measurement error, define

$$g(b) = \int \rho(z)\rho(z-b)dz \tag{2.4}$$

and normalise to obtain

$$q(b) = \frac{g(b)}{g(0)}, \tag{2.5}$$

such that $q(b)$ is the probability that y_t is indistinguishable from x_t. Since $q(0) = \frac{g(0)}{g(0)} = 1$, a state is indistinguishable from itself with probability one.

The theory of indistinguishable states can be extended to a sequence of observed values. Given a time series of observations $\boldsymbol{s} = s_0, s_{-1}, s_{-2}, ...$ of the true trajectory $\boldsymbol{x}$, the probability that $\boldsymbol{x}$ and a trajectory $\boldsymbol{y}$ are indistinguishable from each other is

$$Q(\boldsymbol{x}, \boldsymbol{y}) = \prod_{t \leq 0} q(y_t - x_t) \tag{2.6}$$

and hence if $Q(\boldsymbol{x}, \boldsymbol{y}) = 0$, the trajectories are distinguishable with probability one. Note that $Q(\boldsymbol{x}, \boldsymbol{y})$ is the probability that any set of observations will render two trajectories indistinguishable.

In practice, whilst the system state can only be known if no observational noise is present, it is, however, possible to find states that are indistinguishable from the true state. For two states to be indistinguishable, each one must be consistent with the observation. Since the true state x_t is consistent with the observation s_t with probability one, any state that is consistent with the observation must also be indistinguishable from the true state. The set of trajectories indistinguishable from $\boldsymbol{x}$ is defined by

$$\mathbb{H}(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R} : Q(\boldsymbol{x}, \boldsymbol{y}) > 0\}. \tag{2.7}$$

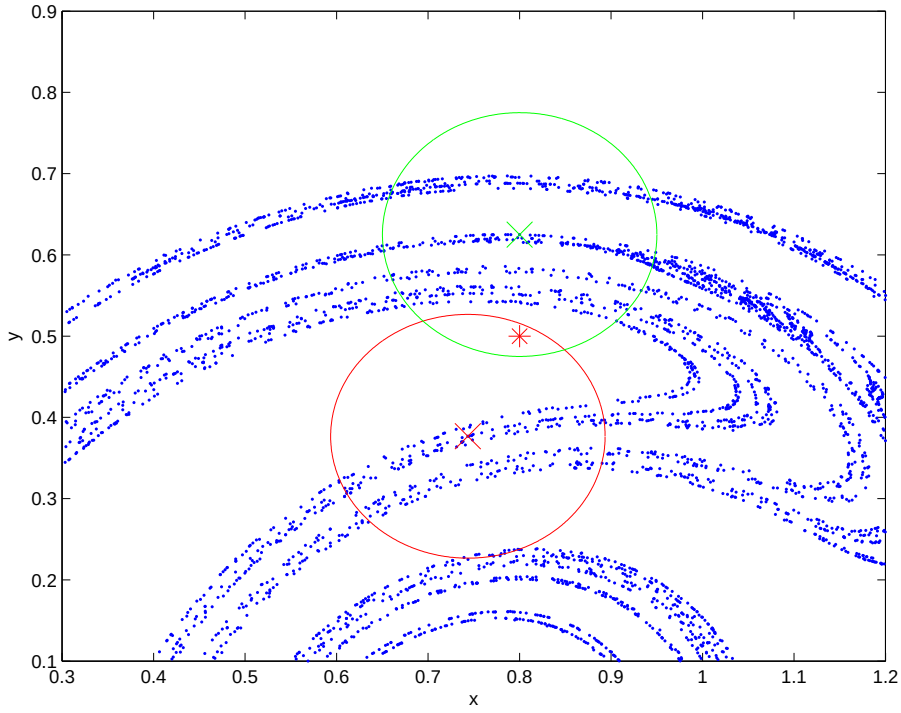

Figure 2.2: Two indistinguishable states on the Ikeda map attractor. The red cross represents the true state x_0 whilst the red circle surrounding it represents the bound of the observational uncertainty, that is the area in which an observation can fall. The green cross and the circle surrounding it represents another state y_0 and its bound of uncertainty. Since the observation s_0, represented with the red star, lies in the overlapping region of the two uncertainty bounds, x_0 and y_0 are indistinguishable.

An illustration of indistinguishable states is shown in figure 2.2. Here, the red cross represents a true state x_0 on the attractor of the Ikeda Map defined in appendix A.1.4 and the red circle surrounding it, the bound of the observational noise in which an observation of x_0 must fall. The green cross and the circle surrounding it represents another system state y_0 and the bound of its observational noise, that is the area in which an observation of y_0 must fall. The red star represents an observation s_0 of x_0. Since the observation falls within the bounds of uncertainty of both states, and hence could conceivably be an observation of either, y_0 is indistinguishable from the true state x_0.

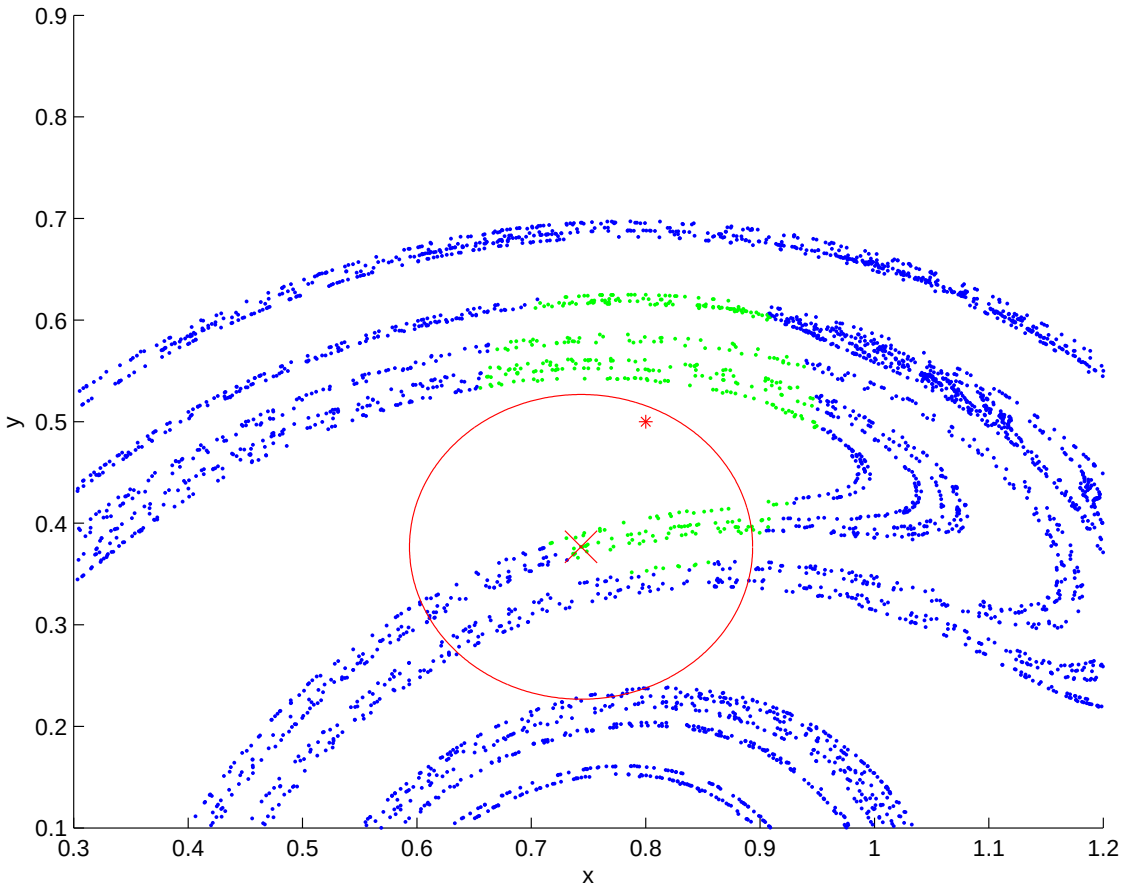

Figure 2.3: States on the Ikeda attractor coloured according to whether they are distinguishable or indistinguishable from the true state x_0 given an observation s_0. x_0 and s_0 are represented with a red cross and a red star respectively. States that are indistinguishable from the true state are coloured green.

The set of system states indistinguishable from the true state are coloured in green in figure 2.3 whilst those that are distinguishable are coloured blue. Note how the position of the observation influences the set of indistinguishable states.

Now, suppose that as well as the observation s_0, a set of observations $s_{-1}, s_{-2}, ...$ of past states $x_{-1}, x_{-2}, ...$ is available. Each state at time $t = 0$ uniquely defines a system trajectory stretching into the past. Suppose a state y_0 is found that is consistent with s_0 and is therefore indistinguishable from x_0. Based only on this

information, the two trajectories $\boldsymbol{x}$ and $\boldsymbol{y}$ cannot be distinguished. Now suppose that an observation from the previous time step s_{-1} of x_{-1} is known. If y_{-1} is not consistent with s_{-1}, the two states x_{-1} and y_{-1}, and thus the two trajectories $\boldsymbol{x}$ and $\boldsymbol{y}$, are distinguishable. By considering extra observations from the past, the set of indistinguishable states can be reduced. This is demonstrated in figure 2.4. In the top left panel, states that are indistinguishable from the true state given only a single observation s_0 are shown in green (as in figure 2.3). In the top right panel, an observation from the previous time step s_{-1} (not shown) is also taken into account and the number of indistinguishable states is reduced. In the bottom left panel, observations s_{-2}, s_{-1} and s_0 are taken into account whilst 4 observations s_{-3},s_{-2},s_{-1} and s_0 are taken into account in the bottom right panel. As the number of observations from the past increases, the number of trajectories that are consistent with $\boldsymbol{x}$ decreases.

2.2.1 Indistinguishable states and the case for probabilistic forecasting

That taking more past observations into account reduces the range of indistinguishable states might lead us to expect that as the number of past observations tends to ∞, the number of indistinguishable states narrows down to only the true state. It is shown in [140], however, that, in fact, this is not the case. This supports the view that, however many observations are available, no single best estimate of the true initial state can be made when observations are subject to measurement error. This has an important impact on our philosophy of forecasting. If the true initial condition can never be recovered in a chaotic system, even in the PMS; we can expect, at best, to find a collection of states that are indistinguishable from it. Thus any

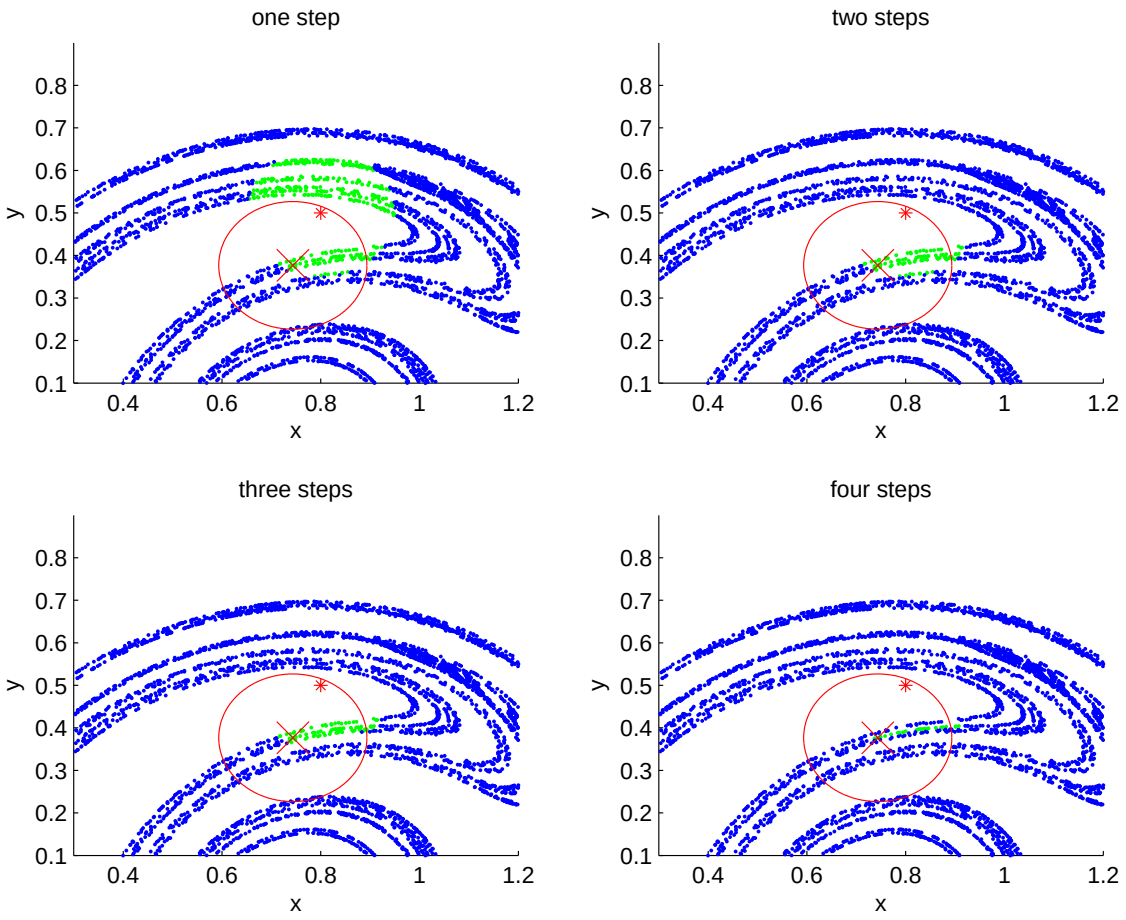

Figure 2.4: Sets of indistinguishable states (coloured green) at time $t = 0$ given the number of observations stated in each panel. As more past observations are taken into account, the number of indistinguishable states diminishes.

point forecast initialised using one of these states can only be considered a single draw from the distribution of possible point forecasts given the set of observations and the model. Sampling the uncertainty around the best guess of the initial state to form multiple forecast trajectories is thus expected to be more informative than forming a single model trajectory.

2.3 Data Assimilation

Data assimilation[161, 40, 48] is the process of combining noisy observations with a model to attempt to find improved estimates of a set of system states. In forecasting, this is usually done over a set of past and present observations with the aim of finding an initial state for a forecast that is consistent with both the model dynamics and the observations. The output of a data assimilation algorithm is called the *analysis*. If the data assimilation scheme and model are effective, a model trajectory initialised with states of this type can usually be expected to perform better than one initialised with noisy observations. Data assimilation is applied to a set of observations over a time period called an *assimilation window*. We now describe the two assimilation techniques used in experiments in this thesis.

2.3.1 Pseudo-orbit Data Assimilation

To describe Pseudo-orbit data assimilation (PDA), [140, 141, 82, 80] we first define the mismatch cost function. This is given by

$$L(\boldsymbol{u}) = \frac{1}{2}\sum_{i=1}^{n-1} ||\boldsymbol{u}_{i+1} - f(\boldsymbol{u}_i)||^2, \tag{2.8}$$

where $\boldsymbol{u} = \boldsymbol{u}_1, ..., \boldsymbol{u}_n \in \mathbb{R}^d$ denotes a vector of states and f denotes the model operator that evolves the model forward one discrete time step. $L(\boldsymbol{u})$ is equal

to zero if and only if $\boldsymbol{u}$ forms a deterministic model trajectory. Otherwise $\boldsymbol{u}$ is called a *pseudo-orbit*. The smaller the mismatch, the closer a pseudo-orbit is to a model trajectory. PDA uses the well known gradient descent algorithm to attempt to minimise $L(\boldsymbol{u})$ so that the pseudo-orbit represented by the observations tends towards a model trajectory. This is equivalent to solving the differential equation

$$\frac{dL(\boldsymbol{u})}{d\tau} = -\nabla L(\boldsymbol{u}), \tag{2.9}$$

which can be solved numerically using the simple Euler method [46] giving the iterative relationship

$$\boldsymbol{u}_{i,m+1} = \boldsymbol{u}_{i,m} - \delta \begin{cases} -A(\boldsymbol{u}_{i,m})(\boldsymbol{u}_{i+1,m} - f(\boldsymbol{u}_{i,m})), & \text{if } i = 1 \\ \boldsymbol{u}_{i,m} - f(\boldsymbol{u}_{i-1,m}) - A(\boldsymbol{u}_{i,m})(\boldsymbol{u}_{i+1,m} - f(\boldsymbol{u}_{i,m})), & \text{if } 1 < i < n \\ \boldsymbol{u}_{i,m} - f(\boldsymbol{u}_{i-1,m}), & \text{if } i = n \end{cases}$$

where $\boldsymbol{u}_{i,m}$ is the m_{th} iteration of the i_{th} state, $A(\boldsymbol{u}_{i,m})$ is the adjoint matrix (the transpose of the Jacobian) of f at $\boldsymbol{u}_{i,m}$ and δ is the euler step size. Larger values of δ tend to result in faster convergence but when this value is too large, the pseudo-orbit can fail to converge. Thus care should be taken to ensure that the mismatch function decreases with algorithmic time. If this does not happen, the value of δ can be reduced until convergence is achieved. It is suggested in [144] that a value of $\delta = 0.1$ provides roughly optimal convergence. We follow this convention, halving the value of δ when the mismatch does not decrease over time. The algorithm can either be applied for a set number of iterations or stopped when the mismatch function falls below some threshold value.

2.3.2 4DVAR

Four Dimensional Variational Assimilation (4DVAR) [67, 68, 125] is a common method of data assimilation used extensively in weather forecasting [17]. The algorithm attempts to find a model trajectory that best fits a set of past and present observations by minimising a cost function.

For a set of observations on the interval $t \in (-n, 0)$, the cost function is given by

$$C_{\text{4DVAR}} = \frac{1}{2}(\boldsymbol{x}_{-n} - \boldsymbol{x}^b_{-n})^T \boldsymbol{B}^{-1}_{-n}(\boldsymbol{x}_{-n} - \boldsymbol{x}^b_{-n}) + \frac{1}{2}\sum_{t=-n}^{0}(H(\boldsymbol{x}_t) - \boldsymbol{s}_t)^T \boldsymbol{\Gamma}^{-1}(H(\boldsymbol{x}_t) - \boldsymbol{s}_t), \tag{2.10}$$

where $\boldsymbol{x}_{-n}$ is a model initial condition, $\boldsymbol{x}^b_{-n}$ is the first guess of the initial condition, otherwise known as the background model state, $\boldsymbol{s}_t$ is an observation at time t, $\boldsymbol{B}^{-1}_{-n}$ is the inverse of the covariance matrix $\boldsymbol{x}^b$ and $\boldsymbol{\Gamma}^{-1}$ is the inverse of the covariance matrix of the observational noise. The first term in equation 2.10 is often known as the background term. The second term minimises the distance between the model trajectory and the observations. By minimising the cost function, initial conditions are found that define a model trajectory with minimal distance from the observations. As more observations are taken into account, more information is available and hence the model trajectory obtained by minimising equation 2.10 is expected to get closer to the system trajectory. With an increasing assimilation window, however, the number of local minima increases, making it harder to locate the global minimum [123]. Therefore, in practice, 4DVAR is usually applied to a moderate length assimilation window.

2.4 Probabilistic forecasting

To attempt to account for the uncertainty in an initial condition, ensembles sample nearby states and evolve them forward with the model. A probabilistic forecast density is a convenient way of interpreting the information in an ensemble. In this section, we describe approaches both to the formation of ensembles and the formation of forecast densities from ensembles.

2.4.1 Forming ensembles

Ensembles are formed by sampling m states consistent with an observation of the initial condition of a dynamical system to form an *initial condition ensemble*. Each member of the initial condition ensemble is then evolved forward in time, using the model, to form an ensemble of point forecasts. Throughout this thesis, we refer to each ensemble and corresponding outcome as an *ensemble-outcome pair*. In the experiments in this thesis, two different ensemble formation techniques are used which we describe below.

Inverse Noise Ensembles

A simple and computationally cheap approach to the formation of initial condition ensembles is simply to perturb the best guess of the initial condition (which may simply be a full observation of the state) with random draws from the inverse distribution of the observational noise. The inverse distribution is defined as the distribution obtained by reflecting the noise distribution in the y axis. If the noise distribution is symmetric and has mean zero, the noise distribution and the inverse distribution coincide. With this approach, the true initial condition always lies

within the range of values that can be drawn. Ensembles formed in this way are called *inverse noise ensembles*.

PDA ensembles

Although inverse noise ensembles are both simple and computationally cheap to run, the resulting initial conditions are consistent with the observations but not usually the model dynamics. Intuitively, to obtain an improved set of initial conditions, it makes sense to aim for consistency with both [79]. When an initial condition ensemble is not consistent with the model dynamics, a period of time elapses before the model simulation move towards the attractor. In meteorology, this time is often referred to as the 'spin up time' [83].

PDA ensembles [40, 41] are a way of finding initial condition ensembles that lie close to the model attractor. PDA ensembles use PDA to search for model trajectories consistent with past observations and lying close to the model attractor. In order to find PDA ensembles, first, a *reference pseudo-orbit* is found. This is done by applying the PDA algorithm to a set of past observations $s_{-n}, .., s_0$ over an assimilation window n. For each ensemble member, a new pseudo-orbit $(r_{-n} + \epsilon_{-n}), .., (r_0 + \epsilon_0)$ is formed where $\epsilon_{-n}, .., \epsilon_0$ are random *iid* draws from the inverse distribution of the noise. The resulting analysis at time zero is then propagated forward using the model to the lead time(s) of interest. This process is repeated m times to form an ensemble of model simulations $\boldsymbol{x}_1, .., \boldsymbol{x}_m$. As ever, the closeness of the ensemble to the model attractor depends on the number of iterations of the PDA algorithm. The process of obtaining PDA ensembles is demonstrated in figure 2.5 for a perfect model of the 1 dimensional logistic map with parameter $a = 4$. This demonstration of PDA ensembles is a novel contribution.

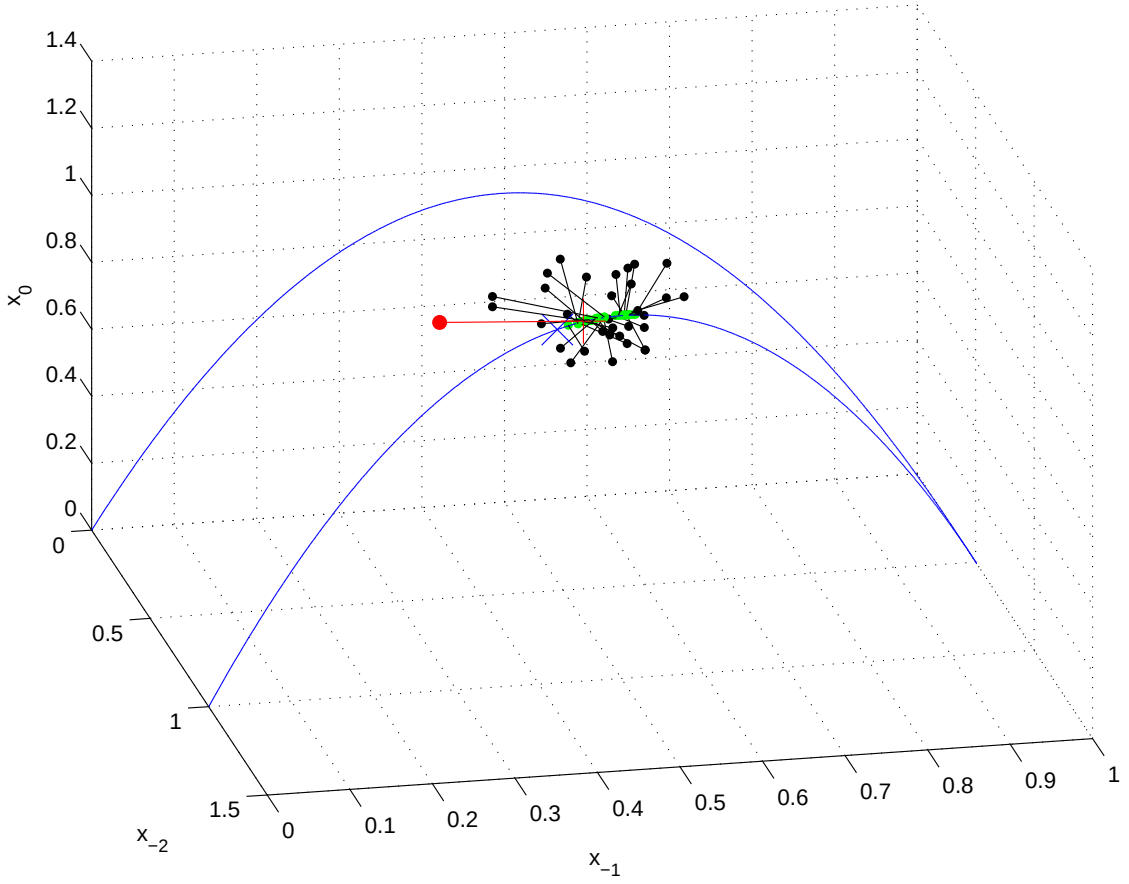

Figure 2.5: A demonstration of how PDA ensembles are formed for a perfect model of the logistic map. The blue line represents sets of 3 consecutive states that are consistent with the model. The blue diagonal cross represents the final 3 points of the system trajectory which consists of a total of 16 time steps (the first 13 steps are not shown on the plot). The red point represents the final three observations (the point where PDA starts) which are assimilated to obtain the reference pseudo-orbit represented with the red cross. The black points are the result of adding random perturbations to the reference pseudo-orbit which are assimilated using PDA to obtain the green points which form the final initial condition ensemble. A close up view of the points is shown in figure 2.6.

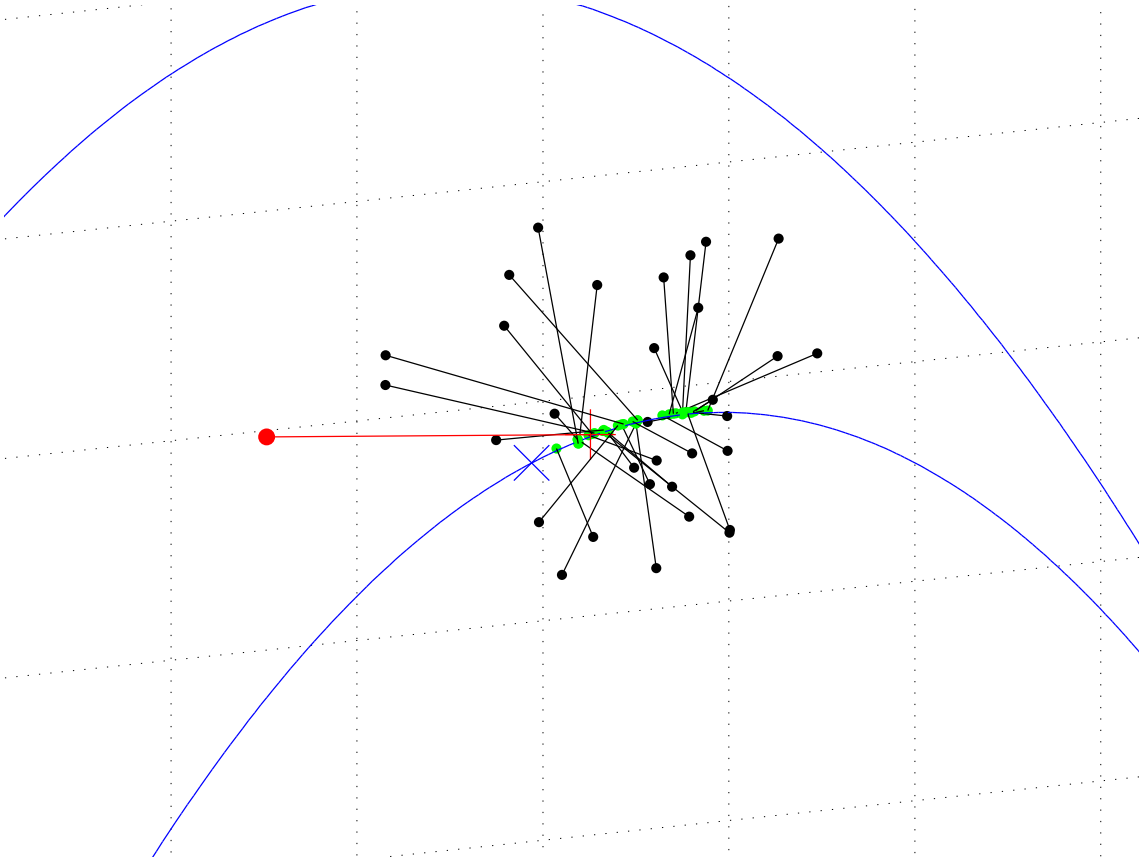

Figure 2.6: A close up view of the points in figure 2.5.

The blue line shows the relationship between 3 consecutive steps of the map and can be thought of as the model attractor. Any points that lie on this line thus represent a model trajectory whilst any points that lie off it are inconsistent with the model dynamics and thus represent pseudo-orbits. The black diagonal cross shows the position of the final 3 points of the system trajectory which consists of 16 steps in total (the first 13 steps are not represented on the plot). The red point represents a set of observations of the system trajectory which form a pseudo-orbit of the model. The observations are then assimilated to obtain the reference pseudo-orbit represented by the red cross which, whilst still not representing a model trajectory, lies much closer to the attractor than the observations. Random perturbations are then added to the reference pseudo-orbit to form new pseudo-orbits, represented by the black points, which are then assimilated using PDA to form the initial condition ensemble shown in green.

2.4.2 System Density

In the forecasting of dynamical systems, when observational noise is present, we can expect, at best, to obtain the exact distribution of possible future states given the distribution of possible initial conditions and the system dynamics. We call this distribution the *system density*. In practice, the ensemble is unlikely to be drawn from the system density even when the model is perfect due to shortcomings in the ensemble formation scheme.

2.4.3 Forming Forecast Densities from Ensembles

Ensemble forecasts consist of a set of point forecasts initialised with slightly different initial conditions. In principle, the variation between ensemble members allows for

a forecaster to understand the uncertainty stemming from the measurement error of the initial condition. In practice, however, it is usually more straightforward to manipulate a distribution function than a set of point forecasts. It is therefore convenient to use ensemble members to form forecast densities in the form of probability density functions. We now review a number of methods of forming forecast densities from ensembles.

2.4.4 Gaussian Dressing

A simple approach to the formation of forecast densities is to fit parametric distributions. Gaussian Dressing [129] assumes that forecast densities can be well described using Gaussian distributions and thus they take the form

$$p(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(y-\mu)^2}{2\sigma^2}}, \tag{2.11}$$

where the parameters μ and σ are chosen to be functions of the ensemble mean $m(\boldsymbol{x})$ and the ensemble standard deviation $s(\boldsymbol{x})$ respectively. This approach, however, has two potentially undesirable properties. The first being that the nonlinear dynamics of many systems are unlikely to yield ensembles that can be realistically represented using a Gaussian distribution. The second property, as noted in [75], is that forecast densities are based purely on the ensemble and no account is taken of forecast performance. This approach assumes that the outcome can be considered to be drawn from the same distribution as the ensemble. In reality, even when the model is perfect, due to the existence of other imperfections in the forecasting system, this is extremely unlikely to be the case. A remedy to the latter problem, suggested in [57], is to optimise the parameters μ and σ according to forecast performance over some training set of ensemble-outcome pairs.

2.4.5 Kernel Density Estimation

Kernel density estimation is a well known and widely used nonparametric method of estimating an underlying probability distribution from a finite sample [119, 127]. Let $\boldsymbol{x} = x_1, x_2, ..., x_n$ be an *iid* sample from some unknown distribution f. A *kernel density estimate* of f is given by

$$\hat{f}_\sigma(x) = \frac{1}{n\sigma} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{\sigma}\Big), \tag{2.12}$$

where $K(.)$ is called a *kernel*, a function which satisfies $\int_{-\infty}^{\infty} K(t)dt = 1$ and is usually but is not necessarily symmetric. In this thesis, we use Gaussian kernels in the form $K(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2}$. In this case, σ is a parameter to be chosen which influences the shape of the density. Kernel density estimation can be used to form forecast densities from ensembles. Like Gaussian dressing, however, kernel density estimation estimates the distribution of the ensemble rather than the outcome.

A vast body of research has been carried out regarding the choice of the bandwidth parameter σ in kernel density estimation [18, 78]. Generally, the aim is to find an estimator that minimises some measure of divergence $d(\hat{f}, f)$ between the true distribution f and the estimated distribution $\hat{f}$. Although f is unknown, it is often possible to choose a kernel width that, given certain assumptions, is optimal in some sense. A common approach to the choice of σ is to use Silverman's constant [137] given by

$$\sigma = (\frac{4s^5}{3n})^{\frac{1}{5}} \approx 1.06sn^{-\frac{1}{5}}, \tag{2.13}$$

where s is the standard deviation of the sample $x_1, .., x_n$. Under the assumption that the underlying distribution $f(x)$ is Gaussian it can be shown that Silverman's

constant is roughly optimal with respect to minimising the Mean Integrated Squared Error (MISE) defined by

$$MISE(\sigma) = E \int (\hat{f}_\sigma(x) - f(x))^2 \, dx. \tag{2.14}$$

A similar approach to forecast density estimation, called kernel dressing, in which the kernel width is trained according to forecast performance, is described in section 2.5.7.

2.4.6 Climatology

The *climatological distribution* or *climatology* is an unconditional probability distribution based on observed values of a system variable collected over some fixed time period. As well as providing a statistical description of the observed values, the climatology can be used as a forecasting distribution. For example, suppose we were interested in the temperature at Heathrow airport exactly a year from now. Since we can only expect weather forecasts to make informative predictions up to around 2 weeks ahead, the best information available is likely to be obtained from looking at the distribution of observed temperatures on that day for the previous, say, 50 years. When the climatology is used as a forecasting distribution, the forecast for any given lead time is fixed, i.e. $p_{t+l}(x) = p_{clim}(x)$ for any l, and so the forecasting density is independent of lead time. The climatology can generally be considered a robust forecast as long as it is based on a large enough number of observations and the system dynamics do not change significantly over time. We can usually, therefore, expect no big surprises in the observation of future states.

Although, on the face of it, the climatology doesn't seem a particularly informative forecast, it actually plays an important role in the treatment of other conditional forecasts. Since past observations are usually easy to obtain and thus the climatology is easy to construct, the skill of the climatology serves as a natural zero value. After all, if we construct forecast densities that perform worse, on average, than the climatology, we may as well issue the climatology (if available) as our forecast. We discuss how to construct climatological distributions in section 2.5.6.

2.5 Probabilistic forecast evaluation

In forecasting, the aim is usually to achieve forecasts of the best possible quality. This leads to the question of how to determine what constitutes a good forecast and how the performance of two or more competing forecasts should be compared. In point forecasting, a measure of skill often consists of some measure of the distance between the forecast and the outcome such as the mean squared error or the mean absolute error [135, 23]. Probabilistic forecast densities are usually assessed using scoring rules which we describe below.

2.5.1 Scoring rules

A *scoring rule* S is a function that measures the performance of a probabilistic forecast. Since it is impossible for an individual probabilistic forecast to be considered 'right' or 'wrong', forecast performance is measured by taking the mean score over multiple forecasts and outcomes. The empirical scoring rule over N ensemble-outcomes pairs is defined by

$$E(S) = \frac{1}{N} \sum_{i=1}^{N} S(p_i, Y_i), \tag{2.15}$$

where p_i and Y_i are the i_{th} forecast density and outcome respectively.

2.5.2 Properties of scoring rules

Scoring rules are an essential part of the forecasting framework since they provide a tool with which to assess the performance of sets of forecast densities. It is thus of vital importance that the scoring rule of choice can identify differences in the performance of two forecasting systems. We describe important properties of scoring rules below.

Propriety

A desirable property of scoring rules is *propriety*. A scoring rule is proper if, for a model density p and a system density q, the following inequality holds:

$$\int_{-\infty}^{\infty} S(p,y)q(y)dy \geq \int_{-\infty}^{\infty} S(q,y)q(y)dy. \tag{2.16}$$

A score is *strictly proper* if, when $p \neq q$, the left hand side of the inequality is strictly greater than the right. Equation 2.16 implies that the expected score achieved using an imperfect forecast p is greater than or equal to the expected score achieved using a perfect forecast q and therefore that a proper score will always, on average, favour the 'true' model. Clearly, this is a desirable property since, in the unlikely case of being in possession of both perfect and imperfect forecast densities, a proper scoring rule can be expected to favour the former.

Locality

A scoring rule has the property of *locality* if its value depends only on the probability or probability density placed on the outcome Y. A discussion of the issues

surrounding the property of locality can be found in [75].

2.5.3 Ignorance Score

The ignorance score [58, 130] is given by

$$S(p, Y) = -\log_2(p(Y)), \tag{2.17}$$

where $p(Y)$ is the probability density placed on the outcome Y. The ignorance is a local score and can be shown to be proper [56, 103]. In fact, the ignorance score is the only score that is both proper and local [15].

2.5.4 Other scoring rules

Naive Linear Score

The naive linear score is a local score for forecasts of continuous variables and is given by

$$S(p, Y) = -p(Y). \tag{2.18}$$

Although seemingly similar to the ignorance score, the naive linear score can be shown to be improper [132].

Proper Linear Score

The proper linear score, given by [75]

$$S(p, Y) = \int p^2(x)dx - 2p(Y), \tag{2.19}$$

is a modified version of the naive linear score. The additional term $\int p^2(x)dx$ renders the score strictly proper. The proper linear score is non-local because it is a function of the entire forecast density.

Continuous Ranked Probability Score

The continuous ranked probability score [44] is given by

$$\int_{\infty}^{\infty} (F(x) - I(x < Y))^2 dx \tag{2.20}$$

where $F(x)$ is the cumulative density function of the forecast density and $I(x < Y)$ is the Heaviside function, which takes a value of 1 when $Y > x$ and 0 otherwise. It is a non-local score because it is a function of the entire forecast density and can be shown to be proper [114].

2.5.5 Cross validation

In any problem in which the parameters of a model are optimised over a training set, the question of how their values generalise out of sample is an important one. If the parameter values perform well with respect to some measure over the training set but not over an independent sample, the parameter values are said not to be robust. Several approaches to this problem have been proposed. For example, information criteria apply a penalty for each extra fitted parameter [6, 133]. An alternative approach is cross-validation. With cross-validation, the parameter values are optimised over subgroups of the training set and tested over others. In this thesis, we use two types of cross-validation which we now describe.

Two Fold cross-validation

In two fold cross-validation, the training set is randomly divided into two sets d_0 and d_1. In the first stage, the parameter values are optimised over the values contained in d_0. The performance of the parameter values, with respect to some measure, is then tested on the values contained in d_1. In the second stage, the parameters are optimised over d_1 and tested on d_0.

Leave one out cross validation

In leave one out cross-validation, the parameters are optimised over all but one of the members of the training set and the performance tested over the remaining member. This process is repeated leaving each member of the sample out exactly once. This method is useful when the training set is small. It can, however, be computationally expensive.

2.5.6 Constructing Climatology

In this thesis, we construct the climatological distribution using kernel density estimation. The kernel width is selected using leave one out cross validation, minimising the mean ignorance score.

2.5.7 Kernel Dressing

Kernel dressing [131, 22] is a method similar to kernel density estimation (KDE) in which a continuous forecast density is formed from an ensemble $\boldsymbol{x} = x_1, x_2, .., x_m$. Like kernel density estimation, a forecast density constructed in this way takes the form of a linear combination of kernels centred around the ensemble members. In

its general form [22], a forecast density is given by

$$p(y|\boldsymbol{x}, a, o, \sigma) = \frac{1}{n\sigma}\sum_{i=1}^{n} K\Big(\frac{y - ax_i - \upsilon}{\sigma}\Big), \tag{2.21}$$

where $\boldsymbol{x}$ represents an ensemble, a is a scaling parameter, υ is an offset parameter and σ is the kernel width. The parameters a, υ, σ are optimised over a training set of ensemble-outcome pairs with respect to the empirical scoring rule given by

$$S_{N_{tr}} = \sum_{i=1}^{N_{tr}} S(p_i(x), Y_i), \tag{2.22}$$

where S defines some scoring rule and N_{tr} is the size of the training set. In this thesis, we use Gaussian kernels. We refer to the case in which the scaling parameter a is set to 1, leaving only the bandwidth parameter offset parameter υ and the bandwidth parameter σ to estimate, as *simple kernel dressing*. Equation 2.21 thus reduces down to

$$\hat{f}_\sigma(y|\boldsymbol{x}, \upsilon, \sigma) = \frac{1}{n\sigma}\sum_{i=1}^{n} K\Big(\frac{y - x_i - \upsilon}{\sigma}\Big). \tag{2.23}$$

Although kernel dressing and kernel density estimation seem very similar in their formulation, they are actually quite different in their aims. In kernel density estimation, the aim is to estimate the underlying distribution of a sample. With a sensible choice of the bandwidth, the estimated density will usually tend towards the density from which the sample was drawn as the sample size n tends to infinity. In kernel dressing, on the other hand, the ensemble can not usually be assumed to be drawn from the same distribution as the outcome and thus the aim is to find a forecast density that is as 'useful' as possible. A useful forecast density is defined as one that performs as well as possible with respect to the chosen scoring rule. Throughout

this thesis we optimise the parameters with respect to the empirical ignorance score given by

$$S_{N_{tr}}(\sigma) = \sum_{i=1}^{N_{tr}} -\log_2[p_i(Y_i, \boldsymbol{x}_i, \sigma)]. \tag{2.24}$$

The optimised parameter values are then used to convert future ensembles into forecast densities. Each forecast lead time is treated distinctly and different parameter values are found in each case.

2.5.8 Blending with climatology

The climatological distribution of a system provides a natural zero value for the skill of any system. After all, if any set of forecasts perform worse, on average, than the climatology, they are of little value since better results could be obtained by issuing the climatology, if available, as a forecast. *Blending* [22] ensures that a set of forecasts are never expected to perform worse than the climatology, in expectation, by producing final forecasts that are a linear combination of the model based forecast and the climatology. Blended forecasts take the form

$$p(x) = \alpha p_m(x) + (1 - \alpha)p_{clim}(x), \tag{2.25}$$

where $p_m(x)$ is a model based forecast, $p_{clim}(x)$ is the climatological distribution and $\alpha \in [0, 1]$ is a parameter to be found by optimising with respect to some scoring rule over a training set of forecast-outcome pairs. In simple kernel dressing, the blending parameter is optimised simultaneously with the kernel width σ. If the optimised value of α is found to be close to 0 then the model forecast has little value and the blended forecast and the climatology will coincide. As well as performing the useful role of ensuring that the performance of the forecasts is bounded below by that of

the climatology, blending can also significantly improve forecast skill [22]. In this thesis, unless otherwise stated, all forecast densities are blended with climatology.

Chapter 3

Limitations of Linear Analysis

Forecast evaluation is an extremely important part of the forecasting framework. After all, without a reliable way of assessing the performance of different forecasting systems, it is difficult to know which, if any, are useful. Proposed forecasting systems are usually assessed by measuring their performance over sets of past states. The approach taken to the evaluation of a forecast depends on its form. In this chapter, we consider the evaluation of point forecasts.

The simplest and most commonly used method of evaluating point forecasts is to take some measure of distance between the forecast and its corresponding outcome such as the mean squared error or the mean absolute error [135, 23]. Whilst this approach seems straightforward and intuitive, it is known to have serious shortcomings when the underlying system dynamics are nonlinear [110]. Another commonly used method of evaluating point forecasts is the anomaly correlation coefficient skill score (ACC) which measures Pearson's correlation coefficient between the forecast and outcome anomalies, that is the forecasts and outcomes after their climatologi-

cal means have been subtracted [112, 121, 77]. In this chapter, we demonstrate a number of misgivings of this approach.

In section 3.1, we demonstrate the fact that, although the ACC, by definition, can just as easily be applied to data in which the climatological mean varies with time, in practice, its exact value may be unattainable since a forecaster is unlikely to know or to be able to estimate the climatological mean perfectly at any given point in time. In practice, a forecaster can only calculate the correlation between estimated anomalies, that is the original forecasts and outcomes with an imperfect estimate of the climatological mean subtracted. We show that this can give misleading results and, more specifically, that linear trend removal [52, 97, 14, 29, 88], a common remedy to this problem, is unlikely to yield a solution.

In section 3.2, we demonstrate how, even if the climatological mean can be accurately estimated, the dispersion of a set of outcomes can have a strong effect on the ACC giving potentially misleading results.

In section 3.3, we demonstrate, using an artificial example, how the issues arising from influential observations [27, 28] can have a strong effect on the ACC. We suggest that great care should be taken to ensure that certain observations do not have a misleading effect on the perceived skill of a set of forecasts.

Whilst the presentation of the observations made in sections 3.1 and 3.2 are new to this thesis, we claim only the those in section 3.3 to be novel.

In chapter 2, we described the theory of indistinguishable states [140, 141] and how this leads to the conclusion that, with the presence of observational noise, the true initial condition cannot be distinguished from other model states. The result of this is that the point forecast obtained by evolving forward a ‘best guess’ of the initial

condition can only be considered to be a single draw from the conditional forecast density given an observation and the model. For this reason, single valued point forecasts can be considered to be incomplete. Nonetheless, point forecasts continue to be used in many fields and thus their evaluation is of some importance. This, however, is fraught with difficulties as we now explain.

The mean squared error (MSE) between a set of point forecasts and outcomes is defined by

$$\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}, \tag{3.1}$$

where $\boldsymbol{x} = x_1, .., x_N$ represents a vector of forecasts and $\boldsymbol{y} = y_1, .., y_N$ a corresponding vector of outcomes. Using the same notation, the mean absolute error (MAE) is defined by

$$\mathrm{MAE}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{N}|x_i - y_i|}{N}. \tag{3.2}$$

Whilst both of these measures are commonly used to evaluate sets of forecasts [135, 23], neither are necessarily expected to reward finding values with a high probability of occurrence [55]. In fact, it can be shown that the expected MSE is minimised by choosing the point forecast to be the mean of the system density whilst the expected MAE is minimised by the median [110]. It could be argued, however, that an evaluation method should reward finding the *mode* of the system density, its most likely value [110] or, in other words, the point forecast that maximises the likelihood. In linear systems, the error distribution keeps its shape when evolved forward in time and therefore if the distribution of the observational noise is Gaussian, the mode of the system density will be equal to both the mean and the median. In nonlinear systems, however, this will almost certainly not be the case and hence, if we assume our aim is to find the mode of the system density, the aforementioned measures

are likely to be misleading. A demonstration of this is shown in figure 3.1. The bimodal distribution in the top panel represents a non-symmetric system density in which the mean, median and mode all differ. In the lower panel, the MSE (blue line, corresponding to the left y axis) and the MAE (green line, corresponding to the right y axis) between the point forecast value on the x axis and 1024 random draws from the system density are shown. As expected, these measures favour point forecasts close to the mean and the median respectively rather than the peaks of the system density. Point forecasts like these are not very informative since they lie in areas of the system density in which there is little chance of an outcome falling.

Another common approach to the evaluation of point forecasts is the Anomaly Correlation Coefficient skill score (ACC) [4, 52]. This score attempts to measure forecast skill by finding the correlation coefficient between a set of forecasts and corresponding outcomes. If the correlation coefficient is high, the forecasts are considered to have high skill [71]. When data show seasonality, i.e. the climatological distribution changes over some cycle, there will naturally be correlation between the outcomes and the seasonalised climatological mean, suggesting a higher level of performance than is actually present. Therefore, instead of finding the correlation between the original forecasts and the outcomes, the ACC uses anomalies. Anomalies are simply the original values of the forecasts and the outcomes with their climatological means subtracted from them. For a set of point forecasts $\boldsymbol{x} = x_1, ..., x_n$ of $\boldsymbol{y} = y_1, ..., y_n$, anomalies are defined by $\boldsymbol{x'} = \boldsymbol{x} - \mu(\boldsymbol{t})$ and $\boldsymbol{y'} = \boldsymbol{y} - \mu(\boldsymbol{t})$ respectively where $\mu(t)$ is the climatological mean which can be a function of time t.

The ACC is therefore defined by

$$\text{ACC} = \frac{S_{\boldsymbol{x'y'}}}{\sqrt{S^2_{\boldsymbol{x'}} S^2_{\boldsymbol{y'}}}}, \tag{3.3}$$

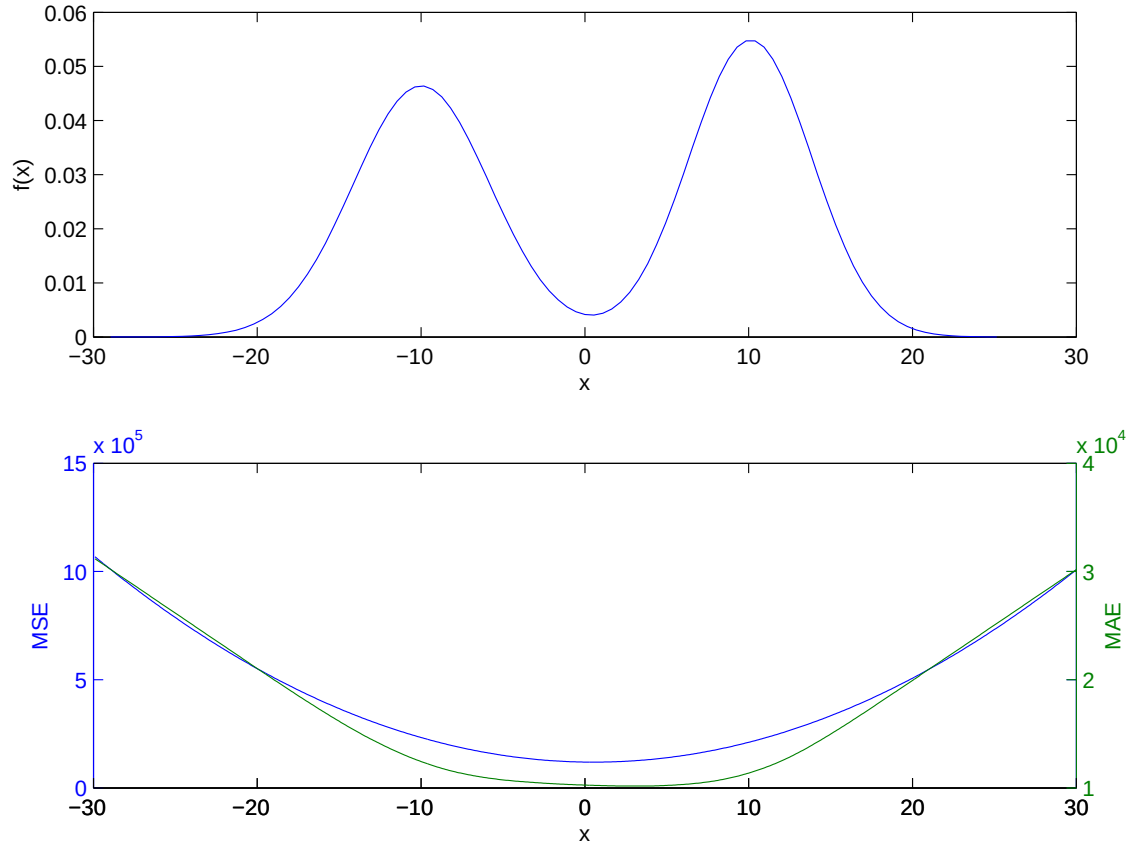

Figure 3.1: Upper panel: A system density in which the mean, median and mode all differ. Lower panel: The mean squared error (blue line corresponding to the left y axis) and the mean absolute error (green line corresponding to the right y axis) between 1024 random draws from the system density in the upper panel and the forecast values on the x axis. In this case, both measures favour forecasts in which there is only low system density.

where $S_{\boldsymbol{x}'\boldsymbol{y}'}$ is the sample covariance between $\boldsymbol{x}'$ and $\boldsymbol{y}'$, $S^2_{\boldsymbol{x}'}$ is the sample variance of $\boldsymbol{x}'$ and $S^2_{\boldsymbol{y}'}$ is the sample variance of $\boldsymbol{y}'$.

In this section, we focus on a number of scenarios that expose the shortcomings of the ACC. This is not the first time weaknesses of the score have been pointed out. It was noted by Brier and Allen in 1951 [8] that the correlation coefficient is insensitive to bias or error in scale. As an example, they point out that the score wouldn't differentiate between a forecast measured in Fahrenheit and a forecast measured in Celsius since these scales are perfectly correlated. Of course, it would not make sense to make a forecast of a set of values measured on one of these scales with the other. A similar point is made in [112] and [111] where common skill scores such as the mean squared error are decomposed into 3 separate terms, the first, a function of the correlation coefficient and the other two, functions of the conditional and unconditional bias respectively, thus showing that the latter two are not taken into account when the correlation coefficient alone is used to measure forecast skill. They thus suggest that rather than being treated as a measure of *actual* skill, the correlation coefficient should be treated as a measure of *potential* skill in the event that conditional and unconditional biases can be removed. Nevertheless, despite these warnings, the ACC continues to be heavily used to measure forecast skill [136, 89, 14, 29, 88]. In this section, we point out three potential weaknesses of the ACC.

3.1 ACC for Data with a Trend

Commonly, a set of outcomes of a variable show a trend over time. A well known example of this is the global mean temperature (GMT) which appears to have shown an upward trend since industrialisation of the western world in the 19th century [66].

Another example of this can be seen in measurements of the sea ice extent in the Arctic which has shown a sharp decline since observations were first taken using satellites in 1978 [85, 34]. As a result, if the correlation coefficient were applied directly between the forecasts and observations, a high positive correlation could be found between the outcomes and any straight line with a positive gradient. If the trend is caused by a shift in the climatological mean, however, by definition, since the climatological mean is subtracted from the forecasts and outcomes, this effect is removed from the ACC. In practice, however, in many cases, the forecaster is unlikely to know the climatological mean in advance and thus it must be estimated. This means that the correlation coefficient found by the forecaster may not resemble the true ACC.

To attempt to remove the effects of an underlying trend in the climatological mean, it is common to assume it is linear. The linear line of best fit is thus subtracted from each of the raw forecast and outcome values [52, 97, 14, 29, 88] before the correlation coefficient is calculated. This, however, assumes firstly that the trend is indeed linear and secondly that the parameters of the linear fit can accurately be found. We demonstrate that, if either of these assumptions fail, the true ACC will not be found and a potentially misleading correlation coefficient will be found instead.

In this section, we show that, although in theory, the ACC can be used to evaluate forecasts of sets of outcomes in which there is an underlying trend in the climatological mean, in practice, since, in many cases, the forecaster is unlikely to know the nature of the trend *a priori*, the correlation coefficient between the *estimated* anomalies of the forecasts and outcomes can bear little resemblance to the correlation between the *actual* anomalies. We consider two scenarios. In the first, we

assume that the trend is a linear function of time and that the forecaster *knows* this but not the true values of the intercept and gradient parameters. In the second scenario, we assume that the trend is nonlinear but is incorrectly assumed to be linear by the forecaster. In both cases, we design scenarios in which the expected correlation between the forecast and outcome anomalies is zero. We show, however, that, the expected correlation between the estimated anomalies (that is, the forecasts and outcomes with an imperfect estimate of the climatological mean subtracted) may be far from zero.

3.1.1 Linear Underlying Trend

Define a vector of one dimensional forecast outcomes $\boldsymbol{y}$ by

$$\boldsymbol{y} = \mu(\boldsymbol{t}) + \boldsymbol{\epsilon}, \tag{3.4}$$

where $\boldsymbol{t}$ is a vector of discrete times $1, .., n$, $\mu(\boldsymbol{t})$ is a function that governs the mean of the climatological distribution at time t and $\boldsymbol{\epsilon}$ is a vector of *iid* draws from a Gaussian distribution with mean zero and standard deviation σ_ϵ. The underlying trend of the climatological mean is linear and therefore takes the form $\mu(\boldsymbol{t}) = \alpha + \beta\boldsymbol{t}$ where α and β are intercept and gradient terms respectively unknown to the forecaster. When β is non-zero, the climatological mean increases or decreases over time.

Let a set of one dimensional point forecasts $\boldsymbol{x}$ of $\boldsymbol{y}$ be defined by the equation

$$\boldsymbol{x} = \mu(\boldsymbol{t}) + \boldsymbol{\delta}, \tag{3.5}$$

where $\boldsymbol{\delta}$ is a vector of *iid* draws from a Gaussian distribution with mean zero and standard deviation σ_δ which is independent of $\boldsymbol{\epsilon}$. The anomalies of both the fore-

casts and the outcomes consist only of the noise terms and thus, at a time t, $x'_t = \delta_t$ and $y'_t = \epsilon_t$. Since $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are defined to be independent, the expected correlation between the forecast and outcome anomalies, and thus the expected ACC, is zero.

In this case, the forecaster is required to estimate the parameters α and β that govern the climatological mean over time. Define $\hat{\mu}(t)$ to be the climatological mean estimated by the forecaster at time t and $\hat{x_t} = x_t - \hat{\mu}(t)$ and $\hat{y_t} = y_t - \hat{\mu}(t)$ to be the forecaster's estimated anomalies of the forecasts and outcomes respectively. Since the forecaster knows the climatological mean is a linear function of t, $\hat{\mu}(t) = \hat{\alpha} + \hat{\beta}t$ where $\hat{\alpha}$ and $\hat{\beta}$ are parameters to be estimated. The forecaster's estimated anomalies at time t are thus:

$$\hat{x}_t = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})t + \delta_t$$
$$\hat{y}_t = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})t + \epsilon_t.$$

For notational reasons, define the differences between the true and estimated parameters of the linear trend to be $\bar{\alpha} = \alpha - \hat{\alpha}$ and $\bar{\beta} = \beta - \hat{\beta}$. The correlation coefficient between $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ is given by:

$$r_{\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}} = \frac{\text{cov}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})}{\sqrt{\text{var}(\hat{\boldsymbol{x}})\text{var}(\hat{\boldsymbol{y}})}}, \tag{3.6}$$

where $\text{cov}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is the covariance between $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, $\text{var}(\hat{\boldsymbol{x}})$ is the variance of $\hat{\boldsymbol{x}}$ and $\text{var}(\hat{\boldsymbol{y}})$ is the variance of $\hat{\boldsymbol{y}}$. To find the correlation coefficient between $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, we derive each of these separately.

The covariance between $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ is

$$\begin{aligned}
\text{cov}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) &= \text{E}((\hat{\boldsymbol{x}} - \text{E}(\hat{\boldsymbol{x}}))(\hat{\boldsymbol{y}} - \text{E}(\hat{\boldsymbol{y}}))) \\
&= \text{E}((\bar{\alpha} + \bar{\beta}\boldsymbol{t} + \boldsymbol{\delta} - (\bar{\alpha} + \bar{\beta}\text{E}(\boldsymbol{t}) + E(\boldsymbol{\delta})))(\bar{\alpha} + \bar{\beta}\boldsymbol{t} + \boldsymbol{\epsilon} - (\bar{\alpha} + \bar{\beta}\text{E}(\boldsymbol{t}) + \text{E}(\boldsymbol{\epsilon})))) \\
&= \text{E}((\bar{\beta}(\boldsymbol{t} - \text{E}(\boldsymbol{t})) + \boldsymbol{\delta})(\bar{\beta}(\boldsymbol{t} - \text{E}(\boldsymbol{t})) + \boldsymbol{\epsilon})) \\
&= \text{E}((\bar{\beta}(\boldsymbol{t} - \text{E}(\boldsymbol{t})))^2 + \boldsymbol{\delta}\bar{\boldsymbol{\beta}}(\boldsymbol{t} - E(\boldsymbol{t})) + \boldsymbol{\epsilon}\bar{\boldsymbol{\beta}}(\boldsymbol{t} - \text{E}(\boldsymbol{t})) + \boldsymbol{\delta}\boldsymbol{\epsilon}) \\
&= \bar{\beta}^2\text{var}(\boldsymbol{t}).
\end{aligned} \tag{3.7}$$

The variance of $\hat{y}$ is

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{y}}) &= \text{E}((\hat{\boldsymbol{y}} - \text{E}(\hat{\boldsymbol{y}}))^2) \\
&= \text{E}(\bar{\alpha} + \bar{\beta}\boldsymbol{t} + \boldsymbol{\epsilon} - (\bar{\alpha} + \bar{\beta}\text{E}(\boldsymbol{t}) + \text{E}(\boldsymbol{\epsilon}))) \\
&= \text{E}((\bar{\beta}(\boldsymbol{t} - \text{E}(\boldsymbol{t})) + \boldsymbol{\epsilon} - \text{E}(\boldsymbol{\epsilon}))^2) \\
&= \text{E}((\bar{\beta}(\boldsymbol{t} - \text{E}(\boldsymbol{t}))^2) + \text{E}((\boldsymbol{\epsilon} - \text{E}(\boldsymbol{\epsilon}))^2) \\
&= \bar{\beta}^2\text{var}(\boldsymbol{t}) + \text{var}(\boldsymbol{\epsilon}) \\
&= \bar{\beta}^2\text{var}(\boldsymbol{t}) + \sigma_\epsilon^2
\end{aligned} \tag{3.8}$$

and the variance of $\hat{x}$ is

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{x}}) &= \text{E}((\hat{\boldsymbol{x}} - \text{E}(\hat{\boldsymbol{x}}))^2) \\
&= \text{E}(\bar{\alpha} + \bar{\beta}\boldsymbol{t} + \boldsymbol{\delta} - (\bar{\alpha} + \bar{\beta}\text{E}(\boldsymbol{t}) + \text{E}(\boldsymbol{\delta}))) \\
&= \text{E}((\bar{\beta}(\boldsymbol{t} - \text{E}(\boldsymbol{t})) + \boldsymbol{\delta} - \text{E}(\boldsymbol{\delta}))^2) \\
&= \text{E}((\bar{\beta}(\boldsymbol{t} - \text{E}(\boldsymbol{t}))^2) + \text{E}((\boldsymbol{\delta} - \text{E}(\boldsymbol{\delta}))^2) \\
&= \bar{\beta}^2\text{var}(\boldsymbol{t}) + \text{var}(\boldsymbol{\delta}) \\
&= \bar{\beta}^2\text{var}(\boldsymbol{t}) + \sigma_\delta^2.
\end{aligned} \tag{3.9}$$

Putting equations 3.7, 3.8 and 3.9 into equation 3.6 yields

$$r_{\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}} = \frac{\bar{\beta}^2 \text{var}(\boldsymbol{t})}{\sqrt{(\bar{\beta}^2 \text{var}(\boldsymbol{t}) + \sigma_\delta^2)(\bar{\beta}^2 \text{var}(\boldsymbol{t}) + \sigma_\epsilon^2)}}, \tag{3.10}$$

where var($\boldsymbol{t}$) is the variance of $\boldsymbol{t}$ which is simply the variance of the sequence of whole numbers from 1 to n and can easily be shown to be $\frac{(n^2-1)}{12}$. This means that, assuming $\hat{\beta}$ is constant over time[1], since var($\boldsymbol{t}$) is strictly increasing, the effect of σ_δ and σ_ϵ is reduced, and thus the correlation between $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ will approach one as n approaches infinity. Only when the forecaster is able to estimate the parameters perfectly is the true ACC expected to be recovered. When the mean of the climatology is not perfectly recovered, the correlation coefficient between the estimated anomalies is likely to be larger than the true ACC which risks overconfidence in the performance of the forecasts.

3.1.2 Nonlinear Underlying Trend Assumed to be Linear

In this example, we consider a case in which the function governing the climatological mean is nonlinear but is wrongly assumed to be linear. We show that incorrectly making this assumption can be highly misleading and can therefore give the impression that a set of point forecasts are skillful when, in fact, this is not the case.

Suppose that, for a set of times $\boldsymbol{t} = 1, .., n$, a set of outcomes $\boldsymbol{y}$ can be expressed in the form

$$\boldsymbol{y} = \mu(\boldsymbol{t}) + \boldsymbol{\epsilon}, \tag{3.11}$$

[1]In practice,the forecaster may refine these as the forecast-outcome archive grows. The effect of doing this is not considered here.

whilst a set of point forecasts $\boldsymbol{x}$ of $\boldsymbol{y}$ take the form

$$\boldsymbol{x} = \mu(\boldsymbol{t}) + \boldsymbol{\delta}, \tag{3.12}$$

where

$$\mu(t) = t^3 - 6t^2 + 12t - 6 \tag{3.13}$$

governs the behaviour of the underlying trend and both $\boldsymbol{\delta} = \delta_1, .., \delta_n$ and $\boldsymbol{\epsilon} = \epsilon_1, .., \epsilon_n$ are vectors of *iid* random numbers drawn from $N(0, \sigma^2)$ and are independent of each other. Here, there is a clear nonlinear trend in the climatological mean.

Given the assumption that the underlying trend in the climatological mean is linear, the estimated anomalies are formed by subtracting the linear line of best fit. This, however, will not completely remove the effect of the underlying trend. A time series of outcomes $\boldsymbol{y}$, formed using equation 3.11 on the interval $[0, 4]$ with $\sigma = 0.4$, are shown in the top panel of figure 3.2 whilst the linearly detrended version is shown in the lower panel. In the former case, the black line represents the climatological mean whilst, in the latter, it represents the climatological mean with the linear trend subtracted. Unsurprisingly, the climatological mean is still a function of time even with the linear trend removed.

By construction, since $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are independent, the expected ACC is zero. We now show, however, that the correlation between the estimated anomalies is not expected to be zero and therefore does not provide an accurate estimate of the ACC. The mean correlation between 1024 time series of outcomes and forecasts formed in this way is shown for different values of σ in figure 3.3. By simply linearly detrending, even when the anomalies are large compared to the effect of the underlying trend in the climatological mean, the ACC and the correlation between the estimated anomalies

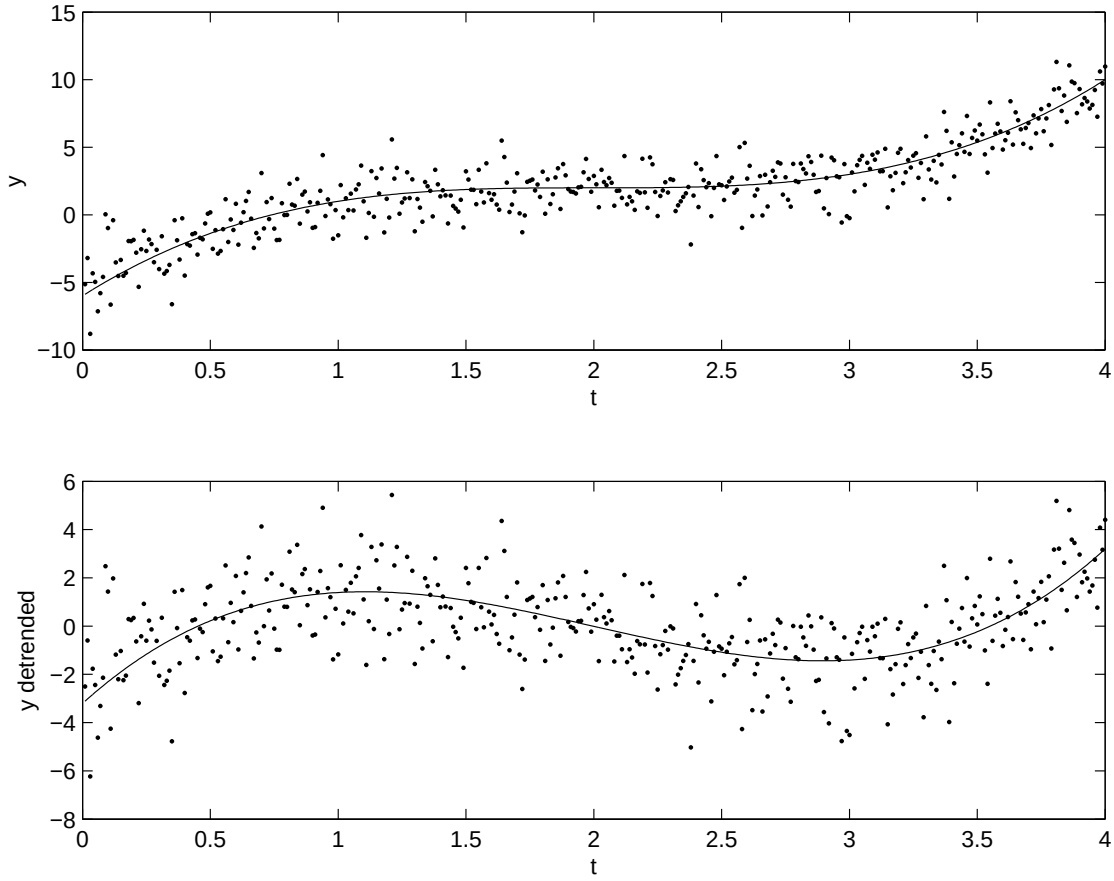

Figure 3.2: Upper panel: A time series of outcomes (black points) formed using equation 3.11 with $\sigma = 0.4$ where the climatological mean $\mu(t)$, represented with the black line, is governed by equation 3.13. Lower panel: The same time series linearly detrended. Linearly detrending fails to remove all of the effects of changes in the climatological mean.

bear little resemblance. It is clear then that simply removing a linear trend and finding the correlation between the estimated anomalies does not necessarily result in an accurate estimate of the true ACC. This can lead to highly misleading results. Moreover, a forecaster may be incentivised not to find an accurate estimate of the climatological mean in order for the forecasts to seem more effective.

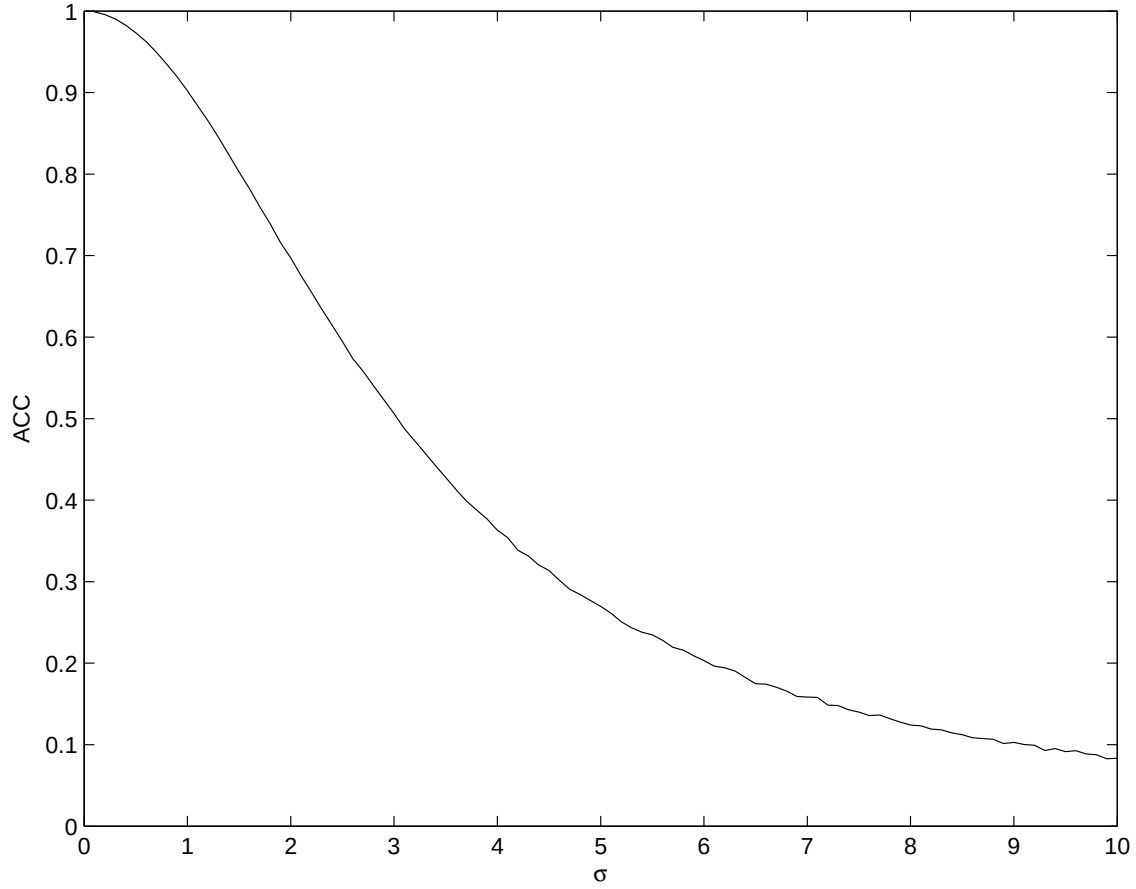

Figure 3.3: The mean of the correlation coefficent between estimated anomalies of 1024 realisations of time series of forecasts and outcomes formed using equations 3.12 and 3.11 respectively on the interval [0,4] with the underlying trend described by equation 3.13 for different values of σ. As σ increases, the relative impact of the underlying trend falls but, since its effect is never removed completely, the ACC is always non-zero. This gives the misleading impression that the forecasts are skillful when, in fact, this is not the case.

3.2 Relationship between the dispersion of outcomes and the ACC

In this section we show that the variability of a set of outcomes can strongly affect the ACC and yield misleading results. Specifically, we show that, as the standard deviation of the outcomes increases, the correlation between a set of forecasts and outcomes is also expected to increase even when forecast error remains constant. We argue that this renders the ACC a poor measure of absolute skill because its value will be heavily influenced by properties of the outcomes rather than the performance of the forecasts. This property makes it difficult to make meaningful comparisons of the performance of forecasting systems on different sets of outcomes.

Define a vector of outcomes $\boldsymbol{y}$ by

$$\boldsymbol{y} = \boldsymbol{\epsilon}, \tag{3.14}$$

where $\boldsymbol{\epsilon} = \epsilon_1, .., \epsilon_n$ is a vector of *iid* random draws from some distribution with mean zero and standard deviation σ_ϵ.

Define a vector of forecasts $\boldsymbol{x}$ of $\boldsymbol{y}$ by

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{\delta}, \tag{3.15}$$

where $\boldsymbol{\delta} = \delta_1, .., \delta_n$ is a vector of *iid* draws from some distribution with mean zero and standard deviation σ_δ. The expected forecast error is thus exactly σ_δ.

The variance of $\boldsymbol{y}$ is

$$\text{var}(\boldsymbol{y}) = \text{var}(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 \tag{3.16}$$

and the variance of $\boldsymbol{x}$ is

$$\text{var}(\boldsymbol{x}) = \text{var}(\boldsymbol{\epsilon} + \boldsymbol{\delta}) = \text{var}(\boldsymbol{\epsilon}) + \text{var}(\boldsymbol{\delta}) = \sigma_\epsilon^2 + \sigma_\delta^2. \tag{3.17}$$

The covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\begin{aligned} \text{cov}(\boldsymbol{x}, \boldsymbol{y}) &= E((\boldsymbol{y} - E(\boldsymbol{y}))(\boldsymbol{x} - E(\boldsymbol{x}))) \\ &= E(\boldsymbol{\epsilon}(\boldsymbol{\epsilon} + \boldsymbol{\delta})) \\ &= E(\boldsymbol{\epsilon}^2 + \boldsymbol{\epsilon}\boldsymbol{\delta}) \\ &= \sigma_\epsilon^2. \end{aligned} \tag{3.18}$$

The analytical correlation coefficient $r_{\boldsymbol{x},\boldsymbol{y}}$ is therefore given by

$$r_{\boldsymbol{x},\boldsymbol{y}} = \frac{\sigma_\epsilon^2}{\sqrt{\sigma_\epsilon^2(\sigma_\epsilon^2 + \sigma_\delta^2)}}. \tag{3.19}$$

Assuming the forecast error σ_δ is constant, the correlation coefficient is thus an increasing function of σ_ϵ, the standard deviation of the outcomes. This means that, when the mean forecast error remains constant, the ACC will reward forecasts of systems with higher dispersion. This is potentially misleading when using the ACC to compare the performance of forecasts as we now explain with a number of examples.

3.2.1 Monthly performance of forecasts of the Central England Temperature record

In this example, we consider the monthly performance of forecasts of daily maximum temperature in the Central England Temperature record [118] from the beginning

of 1878 to the end of 2014. We artificially create point forecasts for each day by adding random Gaussian perturbations with mean zero and standard deviation σ_δ to each observed daily value, resulting in a set of forecasts with constant expected forecast error σ_δ.

Over the duration of the data set, due to natural variability, we can expect there to be variation in the standard deviation of observed values in different months. To demonstrate the effect of this, the standard deviation of the observed values is plotted against the ACC for each month in the data set in figure 3.4. The blue line represents the expected ACC as a function of the standard deviation of the outcomes according to equation 3.19. Here, it is clear that, as expected, the ACC tends to reward sets of forecasts when the outcomes are more highly dispersed. This means that if a forecaster were to use this measure alone to evaluate a set of forecasts, they could be misled into believing that forecast performance is better during months with higher variability when, in fact, this is not the case.

We now describe two more examples of other scenarios in which the above result could mislead a forecaster:

- There is some evidence that climate change can affect not only the mean of a weather variable but also its variability [126, 36, 50]. If the climatological standard deviation increases over time whilst the mean forecast error remains constant, the ACC of the forecasts will increase giving the impression that forecast skill is improving over time.

- Studies often use the ACC to make comparisons between weather forecast performance in different parts of the world [153]. There is, however, often significant variation in the variability of the outcomes [5]. For example, in

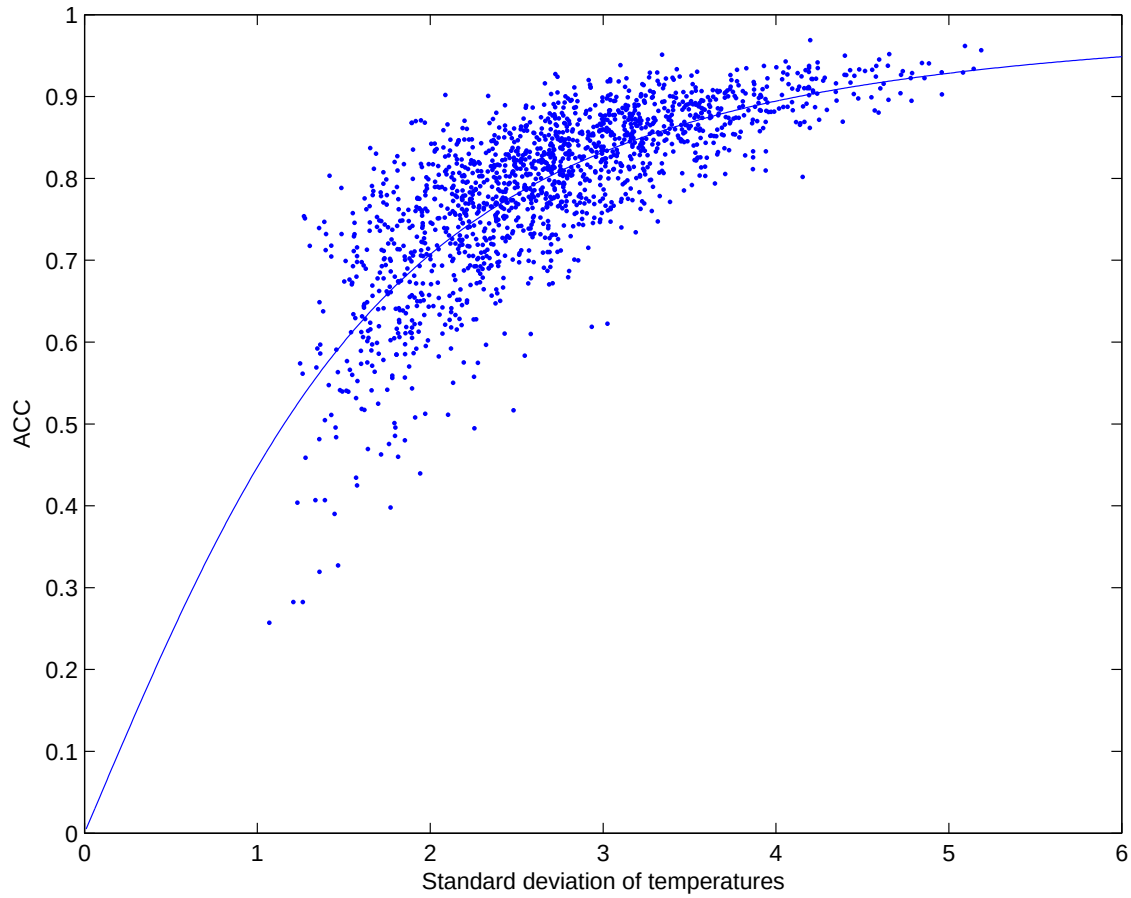

Figure 3.4: Scatter plot of the standard deviation of daily temperatures against the ACC of our artificial forecasts for each month in the CET record. The blue line shows the expected ACC as a function of the standard deviation of the outcomes as described in equation 3.19. Although the expected forecast error remains constant, the ACC tends to increase as a function of the standard deviation of the observed temperatures.

Honolulu, Hawaii, the daily maximum temperature rarely falls below 30 or above 35 degrees Celsius during the summer months. Great Falls, Montana on the other hand, can experience both temperatures exceeding 40 degrees Celsius and significant snowfall in the month of August [3]. This means that, over a year, even if the mean forecast error for each city were identical, the ACC would almost always give a better score to the forecasts of the temperature in Great Falls since the standard deviation of the outcomes is much higher.

3.3 Influential observations

It is well known [106] that the values of both the correlation coefficient and the parameters in linear regression can be disproportionately affected by one or more observations [27, 28]. These are often known as *influential observations*. We now demonstrate that they can have a large effect on the ACC, potentially leading to highly misleading conclusions.

To demonstrate the effect of influential observations on the ACC, we consider an artificial scenario in which a set of deterministic point forecasts and outcomes are related by way of a logarithmic-spiral. A logarithmic-spiral is described by the equations

$$\begin{aligned} x(t) &= a\mathrm{e}^{bt}\sin(\theta) \\ y(t) &= a\mathrm{e}^{bt}\cos(\theta) \end{aligned} \tag{3.20}$$

where a, b and θ are parameters that govern its behaviour.

Consider a set of 16 point forecasts $x_1, .., x_{16}$ and corresponding outcomes $y_1, .., y_{16}$ obtained by sampling from a logarithmic-spiral with parameters $a = 1$ and $b = \frac{\pi}{18}$

on the range $\theta \in (-16\pi + \frac{\pi}{4}, 7\pi - \frac{\pi}{4})$ at intervals of $\frac{3\pi}{2}$. Each forecast-outcome pair is represented with a blue cross in figure 3.5 where the values on the x and y axes represent the values of the forecasts and the outcomes respectively.

The climatological mean of both the forecasts and the outcomes are zero and hence the anomalies and the raw values are identical. The correlation coefficients between the first n points are shown for different values of n in table 3.3. The notable feature here is that each extra point included in the calculation is influential enough to cause the sign of the correlation coefficient to change. A common rule of thumb with the use of the ACC is that values of $r > 0.6$ represent significant forecast skill [71]. Therefore, in this example, if a forecaster assessed an odd number of forecast-outcome pairs, she would, according to this convention, conclude that these forecasts are skillful. If she considered an even number of pairs, the ACC would be negative and her conclusion would be that the forecasts do not have skill. This is clearly not a desirable trait since, as with all statistics, removing or adding a single sample member should not make such a large difference to the results.

Linear regression is often used to attempt to 'correct' future forecasts by treating the outcomes as the dependent variable and the forecasts as the independent variable. Future forecasts are then linearly transformed using these parameters with the aim of improving their accuracy [150]. This process is a form of calibration. We now show, however, that in the scenario outlined above, this process can be highly damaging to the performance of the forecasts.

Let $\tilde{x}_n$ denote the n_{th} calibrated forecast formed using the relation

$$\tilde{x}_n = a + bx_n, \tag{3.21}$$

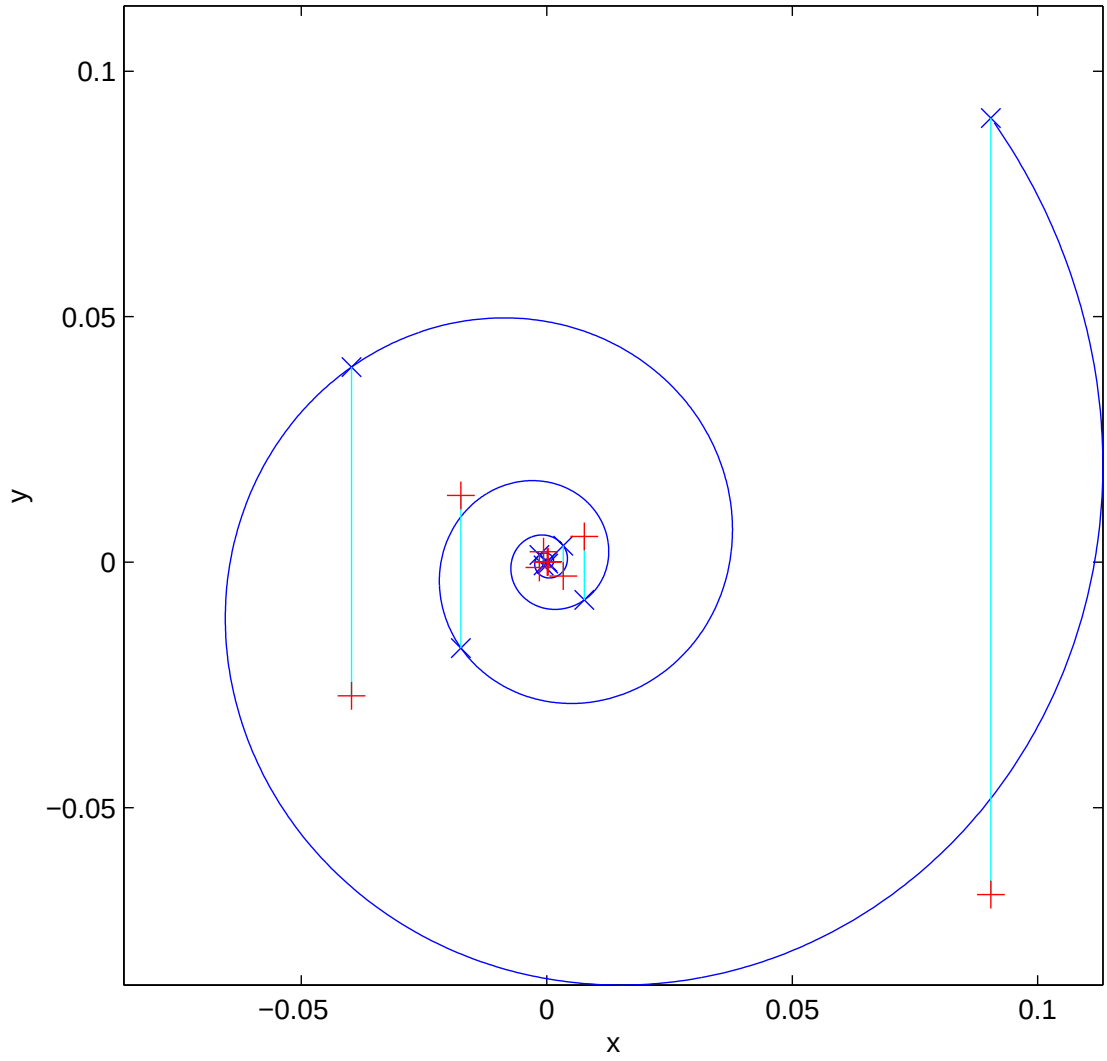

Figure 3.5: A scenario in which the relationship between a set of 16 forecasts and outcomes is governed according to the behaviour of a logarithmic spiral. The x and y coordinates of the blue crosses represent the forecast values and the outcomes respectively. The red crosses, linked to the blue crosses, represent forecasts calibrated using linear regression.

where a and b are regression parameters found using least squares estimation over the set of forecasts $x_1, .., x_{n-1}$ and outcomes $y_1, .., y_{n-1}$.

The red crosses linked to the blue crosses in figure 3.5 represent the relative positions of the calibrated forecasts $\tilde{x}_1, .., \tilde{x}_n$ and the outcomes. In addition to the ACC of the original forecasts, table 3.3 shows the value of the parameter b in equation 3.21, the forecast error between x_n and y_n and the forecast error between $\tilde{x}_n$ and y_n for each value of n. The calibrated forecasts have higher forecast error than the original forecasts and thus calibration is shown to be counterproductive. The reason for this, is that, like the ACC, the value of the parameter b is highly affected by influential observations. This means that calibration, rather than moving the forecasts closer to the outcomes, tends to move them further away and the result is that forecast error is increased.

In this example, although the correlation between the forecast and the outcomes remains high, the in-sample forecast error badly underestimates the error in the next forecast. Clearly, this is not a desirable trait since it could encourage a forecaster to be overconfident in their forecasts.

3.4 The perpetual failure of linear correlation as an indication of predictability

The shortcomings demonstrated in this chapter suggest that use of the ACC should be treated with extreme caution. For its use to be valid, any non-stationarity in the climatological mean should be removed. In practice, in many applications, such as in weather forecasting, it is debatable whether the underlying climatological mean can be estimated accurately enough to have confidence in the results and hence in

n	r	b	$\lvert x_n - y_n \rvert$	$\lvert \tilde{x}_n - y_n \rvert$
5	0.6774	0.6054	0.0007	0.0062
6	-0.6795	-0.7475	0.0025	0.0129
7	0.6791	0.6272	0.0042	0.0311
8	-0.6780	-0.7256	0.0129	0.0670
9	0.6775	0.6382	0.0236	0.1581
10	-0.6774	-0.7146	0.0668	0.3464
11	0.6772	0.6453	0.1278	0.8117
12	-0.6771	-0.7075	0.3461	1.7937
13	0.6775	0.6501	0.6813	4.1802
14	-0.6770	-0.7027	1.7923	9.2881
15	-0.6769	0.6537	3.6024	21.5679
16	0.6768	-0.6991	9.2827	48.1023

Table 3.1: The ACC of the forecasts $x_1, .., x_n$, the slope parameter b from fitting a regression of $y_1, .., y_n$ on $x_1, .., x_n$, the error of the original forecast x_n and the error of the calibrated forecast $\tilde{x}_n$ for different values of n.

most such cases, use of the ACC is not appropriate.

The ACC can be misleading when used to compare the performance of forecasts of sets of outcomes that vary in dispersion. This is because, the more dispersed the outcomes, the higher the ACC for a given level of forecast error. Use of the ACC to make comparisons of the performance of forecasts of different sets of outcomes should thus, in general, be avoided .

As we have demonstrated, the ACC is also highly prone to influential observations. Care should be taken, therefore, to ensure that certain forecast-outcome pairs do not have a disproportionately large effect. This, however, places severe limitations on the ACC. When certain forecast-outcome pairs influence the ACC disproportionately, a forecaster has two choices. They can either remove the influential forecast-outcome pair or abandon the use of the ACC entirely. The former is not a satisfactory solution since it is likely to bias the results and therefore, arguably, the only option is to abandon the use of the ACC in favour of an alternative measure of performance.

Chapter 4

Shadowing Ratios

In this chapter, we propose methods of forecast evaluation that take a different approach to those already considered. Instead of comparing the forecasts and outcomes at one particular lead time, these approaches measure the length of time into the future a set of forecasts satisfy some predefined criteria. In section 4.1, we discuss an existing approach to the evaluation of point forecasts called shadowing that measures the length of time a model trajectory formed using some forecasting system remains sufficiently close to a set of future observations. We then propose a new method called shadowing ratios that aim to give an indication of the relative ability of one forecasting system to shadow longer than another. We further demonstrate the use of shadowing ratios in chapter 7 in the context of comparing data assimilation schemes.

In section 4.2, we extend the approach of shadowing ratios to propose another new method called ensemble shadowing ratios aimed at evaluating the performance of ensembles rather than point forecasts. We demonstrate how shadowing times and

ensemble shadowing ratios can be used to evaluate ensemble formation schemes and give an example in which three different schemes are compared.

Finally, in section 4.3, we introduce boosted probability, a new approach to the evaluation of sequences of probabilistic forecast densities. We demonstrate the use of boosted probability as an evaluation method using a simple example.

The new contributions of this chapter are:

1. Shadowing ratios, a new method of comparing the performance of forecasting systems (section 4.1).
2. Using shadowing ratios to compare the performance of imperfect models (section 4.1.4).
3. Ensemble shadowing ratios, a new method of comparing the performance of ensembles (section 4.2).
4. Using ensemble shadowing ratios to compare the performance of PDA and inverse noise ensembles with different ensemble sizes (section 4.2.1).
5. Boosted probability, a new approach to measure the performance of forecast densities (section 4.3).
6. Comparing the performance of forecast densities using boosted probability (section 4.3.1).

4.1 Shadowing Ratios

In chapter 3, we explained how common point forecast evaluation techniques, such as the mean squared error and mean absolute error, can give misleading results

when the system dynamics are nonlinear. These approaches can often be expected to rank an imperfect model over a perfect one, particularly when observational noise is present [53]. Just as propriety is a desireable property in probabilistic forecasting, however, favouring a perfect model in point forecasting is also desireable. In this section, we discuss the use of an existing approach to evaluating point forecasts called *shadowing* [53] and introduce a new method based upon shadowing called *shadowing ratios*. Shadowing takes a different perspective to the above mentioned techniques in that instead of attempting to measure the performance of forecasts at one or more given lead times, the length of time forecast trajectories stay consistent with future observations is measured. Under this approach, unlike the mean squared error and the mean absolute error, by comparing the period of time a set of model trajectories shadow the observations, it is always expected, on average, for a perfect model to be favoured over an imperfect one [53]. Shadowing is an intuitive technique of measuring the performance of one or more forecasting systems and can be thought of as a measure of how long into the future one can expect a forecast trajectory to be reliable. Shadowing ratios, which estimate the probability that a trajectory formed using one forecasting system stays close to the observations longer than that of another, allow a user to understand better the nature of any differences in shadowing performance between two competing forecasting systems.

4.1.1 Shadowing

A model trajectory is said to *shadow* a set of observations for as long as it stays consistent (as defined below) with them given the observational noise model[1] [49, 53, 61]. We define the period of time for which a model trajectory shadows a particular set of observations as a *shadowing length*. We make an important distinction between the

[1]It should be noted that this differs from the most common definition of shadowing

shadowing *length* and the shadowing *time*. Whilst the former measures the length of time a single model trajectory shadows the observations, the latter attempts to measure the maximum shadowing length over all possible initial conditions. In reality, this is done by measuring the maximum shadowing length over a finite number of model trajectories and thus the true shadowing time is usually underestimated.

Shadowing provides us with a measure of the quality of a forecasting system. In the perfect model scenario, trajectories exist that can shadow the observations for any given period of time (of course the system trajectory will stay consistent with the observations indefinitely) although in reality finding them is usually difficult. In the real world, all models are imperfect and, as such, only finite shadowing trajectories exist. This observation brings up a useful property of shadowing, however. By comparing either shadowing lengths or shadowing times, a perfect model is always, on average, expected to outperform any imperfect model. Shadowing lengths also provide a useful measure of how far into the future we can rely on our forecasts. In order to find shadowing lengths in practice, a distinction is required between two different types of noise, bounded and unbounded. The two cases are described below.

Bounded noise

When the observational noise follows a bounded distribution, an observation $\boldsymbol{s}_t \in \mathbb{R}^m$ and the true state $\boldsymbol{x}_t \in \mathbb{R}^m$ can only lie a finite distance apart from each other. When this is the case, it is straightforward to determine whether a trajectory shadows an observation. A model state is consistent with and thus shadows an observation if $|\boldsymbol{s}_t - \boldsymbol{x}_t| < \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a vector of the bounds of the noise for each observed variable. For example, consider a case in which the observational noise of a one

dimensional system follows a continuous uniform distribution on the range $(-1, 1)$ and thus the noise is bounded by $\epsilon = 1$. The shadowing length is said to be the period of time during which the model trajectory stays within a distance of 1 unit of the observations. In figure 4.1, for the case in which only one variable is observed, the blue and the green lines represent two shadowing trajectories of the observations, represented by the red circles. For the period of time in which the trajectories fall within the bound of the noise (between the red crosses), they are considered to shadow the observations. Each trajectory is shown as solid when considered to shadow and dashed thereafter.

Unbounded noise

It is common for observational noise to follow an unbounded distribution. For example, as a likely consequence of the central limit theorem [13, 148], measurement error is often assumed to follow a Gaussian distribution. When this is the case, in theory, the observational noise can take any value between $-\infty$ and ∞. A different approach to that of bounded noise is thus required. Since unbounded distributions can take any value on the real number line, it is difficult to determine whether a trajectory shadows a single observation. Instead, one of several approaches could be taken. For example, the distribution of the residuals of the forecast trajectory can be compared to the known distribution of the observational noise and tested using a measure of the goodness of fit such as the Kalmogorov-Smirnov test. Alternatively, in [139], a less general approach is taken that applies to Gaussian noise only. In this thesis, for simplicity, we only consider bounded noise distributions.

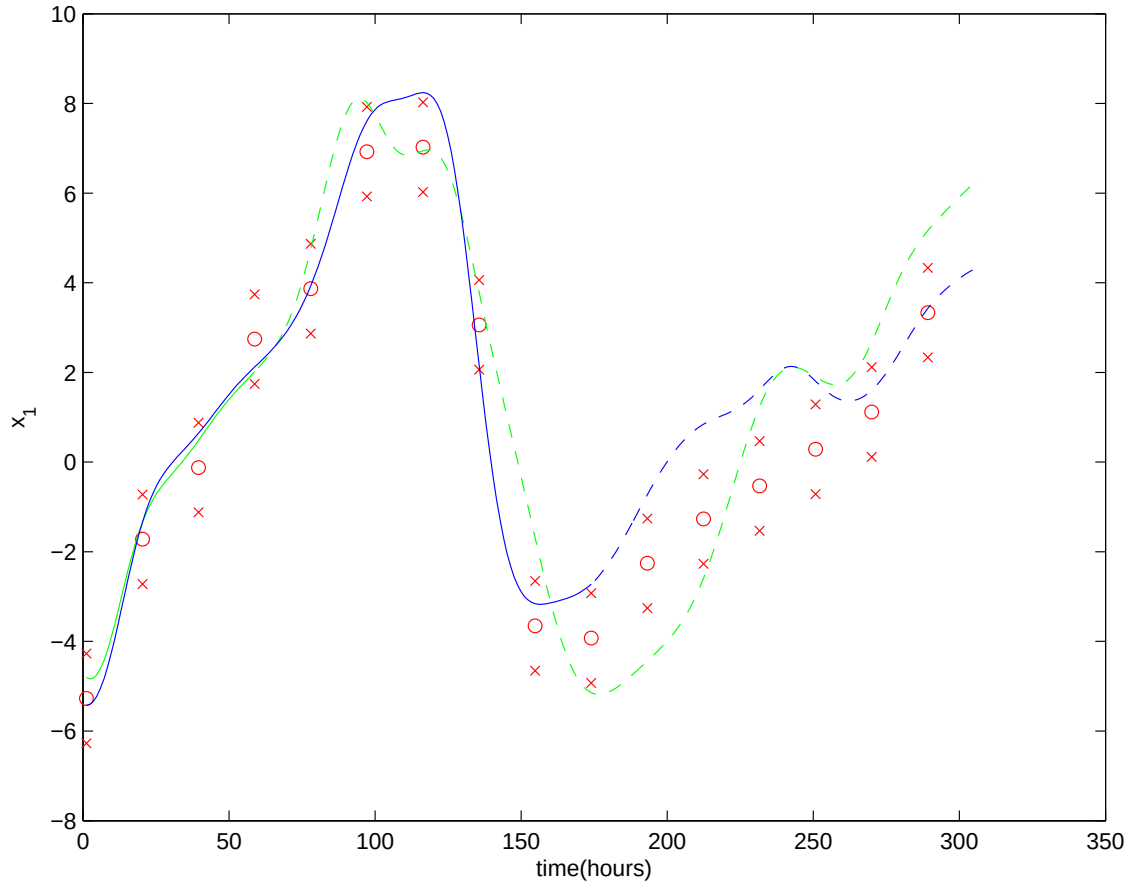

Figure 4.1: Two shadowing trajectories of a one dimensional flow. The red circles represent 'noisy' observations whilst the red crosses represent the bounds of the noise. If a trajectory falls within the bounds of the noise, it is said to shadow the observation. The trajectories coloured blue and green, shown as solid when they are considered to shadow and dashed thereafter, shadow for 7 and 3 time steps respectively.

4.1.2 Definition of shadowing ratios

In the forecasting of dynamical systems, it is desirable for model trajectories to stay close to the observations for as long as possible. Mean and median shadowing lengths are an intuitive way of measuring this length of time and therefore assessing forecast performance. These measures alone, however, do not tell the whole story. Suppose that, over a set of different forecasts and observations, one forecasting system has a longer mean shadowing length than another. With this information alone, it is impossible to differentiate a situation in which one forecasting system yields consistently longer shadowing lengths from one in which there is a larger improvement but only in a small number of cases. It is possible, for example, in a comparison of two forecasting systems, for one to yield a longer mean shadowing length but for the other to yield longer shadowing lengths in the majority of cases.

The proportion of forecasts in which some method yields a longer shadowing length than another provides a potentially useful measure of relative performance. This forms the basis of shadowing ratios which we formally define shortly. In practice, this is complicated slightly due to the potential existence of ties, i.e. occasions in which the observed lengths are the same for both forecasting systems. This problem arises in the non-parametric sign test of paired samples [26] in which the proportion of cases in which values in one sample exceed those in another is calculated. In its most common form, ties are simply ignored. This can be misleading, however, since, arguably, ties provide evidence for the null hypothesis that there is no difference between the two populations from which the samples were drawn. An alternative approach is, in the event of a tie, to attribute equal weight to both populations. This seems most reasonable in cases in which the underlying sample is on a continuous scale but observations are rounded. In the comparison of shadowing lengths, this

corresponds to a case in which the underlying system is on a continuous scale but observations are taken at discrete time steps. In practice, this is likely to be the case for all systems evolving in continuous time and it is therefore reasonable to assume that, given a high enough density of observations, we could usually find distinct shadowing lengths for two truly different forecasting systems. In the case of ties, we therefore place equal weighting on each forecasting system.

We now formally define shadowing ratios. Let $\tau_{i,new}$ and $\tau_{i,ref}$ be the shadowing length of model trajectories formed using some 'new' forecasting system and a 'reference' forecasting system respectively for the i_{th} set of observations. Define the *indication function*[2] to be

$$I(l_1, l_2) = \begin{cases} 0 & \text{if } l_1 < l_2 \\ \frac{1}{2} & \text{if } l_1 = l_2 \\ 1 & \text{for } l_1 > l_2 \end{cases} \tag{4.1}$$

The *shadowing proportion* of a 'new' forecasting system and some 'reference' forecasting system is defined as

$$\hat{p} = \frac{1}{N} \sum_{i}^{N} I(\tau_{i,new}, \tau_{i,ref}) \tag{4.2}$$

and the shadowing ratio is

$$\phi = \frac{\hat{p}}{1 - \hat{p}} \tag{4.3}$$

The shadowing ratio can be considered an estimate of the ratio of the number of

[2]Note that the indication function differs from the well known indicator function in its treatment of ties.

events in which the shadowing length is longer for the new forecasting system to when the shadowing length of the reference forecasting system is longer. If the aim of the 'new' forecasting system is to improve on the reference forecasting system, a shadowing ratio greater than 1 is desirable since this suggests that model trajectories formed using the former are likely to shadow longer than those generated using the latter. Since we include ties rather than just discarding them, finding the sampling distribution of $\hat{p}$ is slightly more difficult than if we were using the sample proportion. Bootstrap resampling can be used, however, to infer whether a shadowing ratio greater than 1 results from a genuine difference in performance or simply from sampling error.

4.1.3 Uses of Shadowing Ratios

We have discussed shadowing and defined shadowing ratios as a way of comparing two forecasting systems. Below, we provide a (not necessarily exhaustive) list of possible changes to the forecasting system that could be tested using either or a combination of these approaches.

1. model structure.
2. model parameter values
3. data assimilation scheme.
4. numerical integration scheme.
5. numerical precision.

Although shadowing lengths and shadowing ratios compare the performance of forecast trajectories, they can also be used to test a specific aspect of a probabilistic

forecasting system. Consider an example in which a forecaster wishes to build probabilistic forecast densities but has the choice of using one of two different data assimilation schemes in the process. To assess the performance of each technique, it is potentially very computationally expensive both to run sets of ensembles and then use them to form forecast densities. Instead, he can simply compare the performance of the deterministic trajectories found using each assimilation method. This approach also has the significant advantage of allowing him to assess the performance of the assimilation technique independently of subsequent aspects of his forecasting system. In section 7.1, we use shadowing to compare the performance of various data assimilation techniques.

4.1.4 Example - Using shadowing ratios to compare model performance

In this example, we show how shadowing ratios can be used to evaluate the performance of imperfect models with varying degrees of structural imperfection. In Appendix A.2.1, we define an imperfect model of the Lorenz '63 system where c is a parameter that controls the level of imperfection. As c tends to infinity, the model equations tend towards the system equations and thus small values of c result in high imperfection in the model. In this experiment, we use shadowing and shadowing ratios to assess the benefit to a forecaster of improving the model by increasing the value of c. As the reference method, we use the model with a value of $c = 4$ which gives the model a high level of imperfection. Details of the experiment, which we call experiment 4.A, are listed in table B.1. The observational noise is Gaussian but trimmed at 3 standard deviations. Each model is tested over 2048 different sets of observations. The high value of the imperfection parameter used for the reference

method causes it to perform poorly, only shadowing the observations for an average of 0.54 days in Lorenz time. Mean shadowing lengths for different values of c are shown in the upper panel in figure 4.2 along with 95 percent bootstrap resampling intervals of the mean. In the lower panel, the shadowing ratio of the imperfect model with varying values of c and the imperfect model with $c = 4$ are shown with 95 percent bootstrap resampling intervals. Whilst both measures indicate, as expected, that reducing the imperfection in the model tends to improve model performance, the shadowing ratios give a slightly different perspective. As c is increased and thus the model imperfection is reduced, the shadowing ratios level off at around $c = 14$ suggesting that, for larger values of c, increases in the mean shadowing length tend to be concentrated in sets of observations in which the shadowing length is already longer than for the reference method. If the increase in shadowing lengths were distributed over all sets of observations, we would expect the shadowing ratio to continue to increase with c since, eventually, all of the shadowing lengths would be longer than for the reference method. This demonstrates the value of shadowing ratios in determining the nature of the difference in shadowing lengths between two forecasting systems.

4.2 Ensemble Shadowing Ratios

Sometimes, it is of interest to compare the performance of ensembles rather than individual model trajectories. This can be a useful thing to do since it allows a forecaster to evaluate the performance of the ensembles independently of any interpretation techniques such as density formation. Diagnostic tools, such as rank histograms [63] and reliability diagrams [115], although not evaluation techniques per se, are often used to assess the consistency of a set of ensembles. It is often sug-

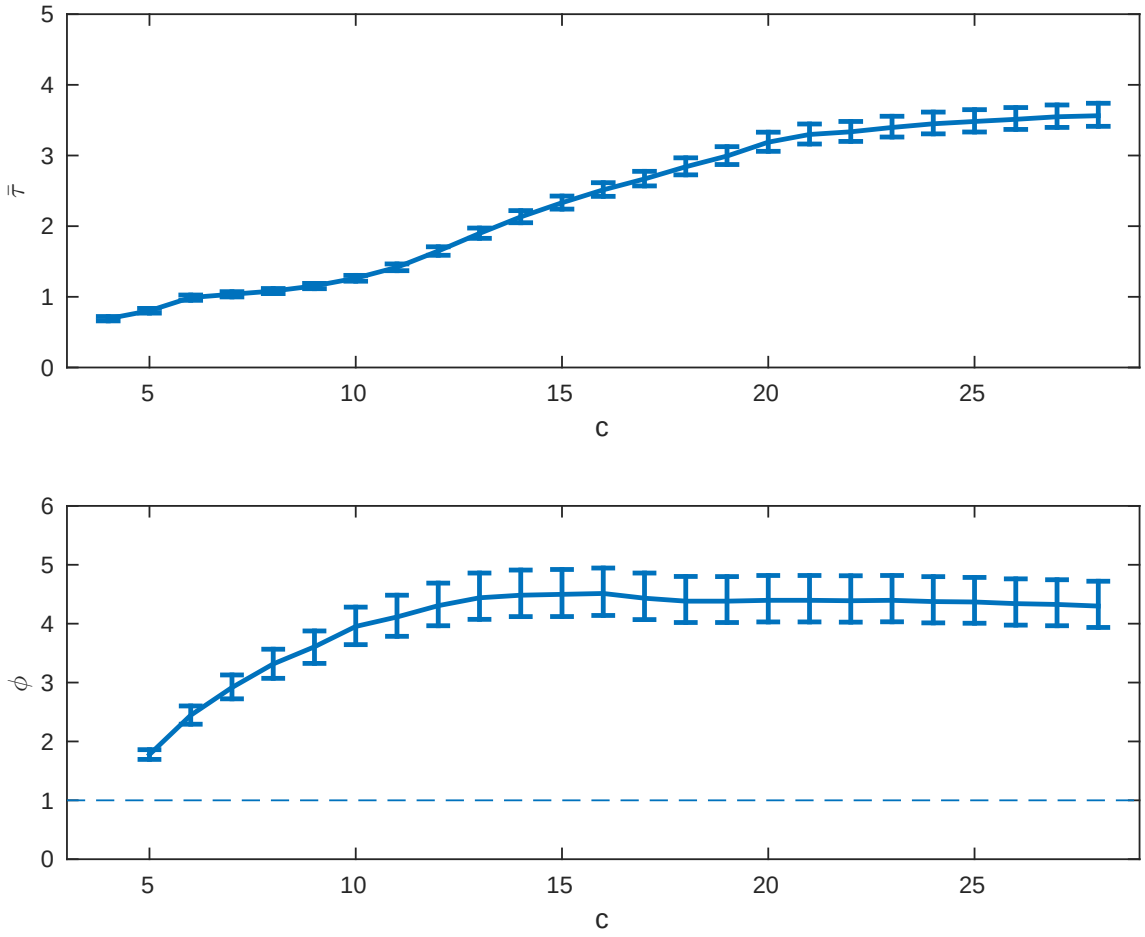

Figure 4.2: Upper panel: Mean shadowing lengths with 95 percent bootstrap resampling intervals for different values of the imperfection parameter c in experiment 4.A. Lower panel: Shadowing ratios of forecasts formed using the model with imperfection parameter c and those formed with imperfection parameter 4 for different values of c with 95 percent resampling intervals. Each measure gives a different perspective on the performance of the model with different degrees of imperfection.

gested that the continuous ranked probability score (CRPS) provides a convenient method of evaluating raw ensembles. It can be shown, however, that the score has the undesirable property that it is not expected to be optimised when the ensemble members are drawn from the system density [21]. The Brier score [19] can be used to evaluate ensembles but is only applicable when the outcome is a binary event.

In section 4.1, we discussed the use of shadowing lengths and introduced a new method called shadowing ratios which can be used to compare the performance of two forecasting systems. We discussed how this approach could be used to assess a wide variety of factors in a forecasting system. These approaches, however, are not appropriate to compare the performance of ensembles since they are designed to compare individual model simulations. In this section, we discuss the use of shadowing times and introduce *ensemble shadowing ratios*, a slightly modified version of shadowing ratios, to compare the performance of ensembles.

Consider the set of states, which may or may not lie on the model attractor, consistent with an observation $\boldsymbol{s}_0$ of an initial state $\boldsymbol{x}_0$. Each of these states, when evolved forward in time, will have a shadowing length associated with it. Any ensemble formation scheme should sample from this set of states but we suggest that, the more effective the scheme, the higher the probability of picking states in which the resulting model trajectory shadows for a long period of time. Recall that we defined a shadowing time to be the maximum shadowing length over an ensemble of model trajectories. The relative performance of two ensemble formation schemes can be assessed by comparing their shadowing times.

Define $\tilde{\tau}_{i,new}$ and $\tilde{\tau}_{i,ref}$ to be the shadowing time of an ensemble formed using some 'new' and some 'reference' technique respectively for the i_{th} set of observations.

Define the ensemble shadowing proportion to be

$$\tilde{p} = \frac{1}{N} \sum_{i}^{N} I(\tilde{\tau}_{i,new}, \tilde{\tau}_{i,ref}) \tag{4.4}$$

where I represents the indication function defined in 4.1.2. We define an ensemble shadowing ratio to be

$$\eta = \frac{\tilde{p}}{1 - \tilde{p}} \tag{4.5}$$

In the previous section, we described how shadowing lengths and shadowing ratios can be used to compare the performance of specific aspects of a forecasting system such as data assimilation techniques. Some data assimilation schemes, however, such as the ensemble kalman filter [47] and particle filtering [60], yield an ensemble of initial states rather than a single 'best guess' and hence these methods cannot be used to assess their performance. Shadowing time and ensemble shadowing ratios, on the other hand, provide an appropriate means of comparison for techniques such as these by directly assessing the performance of the ensembles.

4.2.1 Example - Using ensemble shadowing ratios to compare the performance of ensemble formation techniques

In this example, we demonstrate how shadowing times and ensemble shadowing ratios can be used to compare the performance of sets of ensembles formed using different techniques. In section 2.4.1, we described two ensemble formation schemes, inverse noise and PDA. The former approach has the advantage of being simple and computationally cheap to run but usually yields initial condition ensemble members inconsistent with the model dynamics. PDA ensembles, on the other hand, are expected to be initialised close to the model attractor. Consequently, we expect

PDA ensembles to perform better than inverse noise ensembles.

In this experiment, we compare the performance of inverse noise and PDA ensembles for different ensemble sizes using a perfect model of the Duffing map which is defined in appendix A.1.3. Sets of observations are formed by adding random Gaussian noise at regular intervals to 1024 randomly chosen trajectories from the system attractor. Ensembles are then formed using 3 approaches: inverse noise, PDA with an assimilation window of 8 steps and PDA with a window of 16 steps. Details of the experiment, which we call experiment 4.B, are listed in table B.2.

The results of the experiment are shown in figure 4.3. In the top left panel, the mean shadowing times of ensembles formed using inverse noise (blue), PDA with an assimilation window of 8 steps (red) and PDA with a window of 16 steps (yellow) are shown as a function of ensemble size. In the top right panel, the mean difference between the shadowing times of ensembles formed using PDA with a window of 8 steps and those formed using inverse noise (blue) and the mean difference between the shadowing times of PDA with a window of 16 and 8 time steps (red) are shown with 95 percent resampling intervals of the mean. In the lower panel, the ensemble shadowing ratios between PDA with a window of 8 steps and inverse noise (blue) and PDA with windows of 16 and 8 steps (red) are shown with 95% resampling intervals of the proportion. Here, for all techniques, mean ensemble shadowing times increase with ensemble size. This is unsurprising, given that, as the number of ensemble members increases, the probability of finding a long shadowing length improves and thus a longer shadowing time is likely to be found. Since the resampling intervals of the mean difference in shadowing time do not contain zero, PDA yields significantly longer shadowing times. There is no significant gain in shadowing time by increasing the assimilation window to 16, however. In terms of ensemble shadowing ratios,

similar conclusions can be made. This is because the resampling intervals of the ensemble shadowing ratios between PDA with 8 steps and inverse noise do not include one and thus the former approach yields longer shadowing times significantly more often. This is not true of the resampling intervals between PDA with 16 steps and PDA with 8 steps, however. In the former case, the ensemble shadowing ratio can be interpreted as the PDA ensemble shadowing longer around 1.4 times more often than than the inverse noise ensemble. In both cases, the ensemble shadowing ratios do not show large variation with the ensemble size. In this example, using PDA, rather than inverse noise ensembles, improves both the mean shadowing time and the ensemble shadowing ratio for each ensemble size tested. This suggests that PDA may be better than inverse noise at finding initial conditions that shadow for long periods of time. Increasing the assimilation window from 8 to 16 steps, however, appears to make no significant difference to either the shadowing times or ensemble shadowing ratios in this case..

4.3 Boosted Probability

Scoring rules are a useful tool both for the evaluation of probabilistic forecasts and for the formation of forecast densities in techniques such as kernel dressing. The ignorance score, defined in section 2.5.3, is a good choice of scoring rule since it has desirable properties such as propriety [56, 103]. Since a single probabilistic forecast cannot be shown to be right or wrong, the performance of a forecasting system is usually assessed using the mean score over a set of forecast densities. The mean, however, does not tell the whole story since it says nothing about other aspects of the distribution of scores. Suppose we calculate that the mean ignorance of a set of forecasts is zero. From the mean alone, it is impossible to distinguish a case in

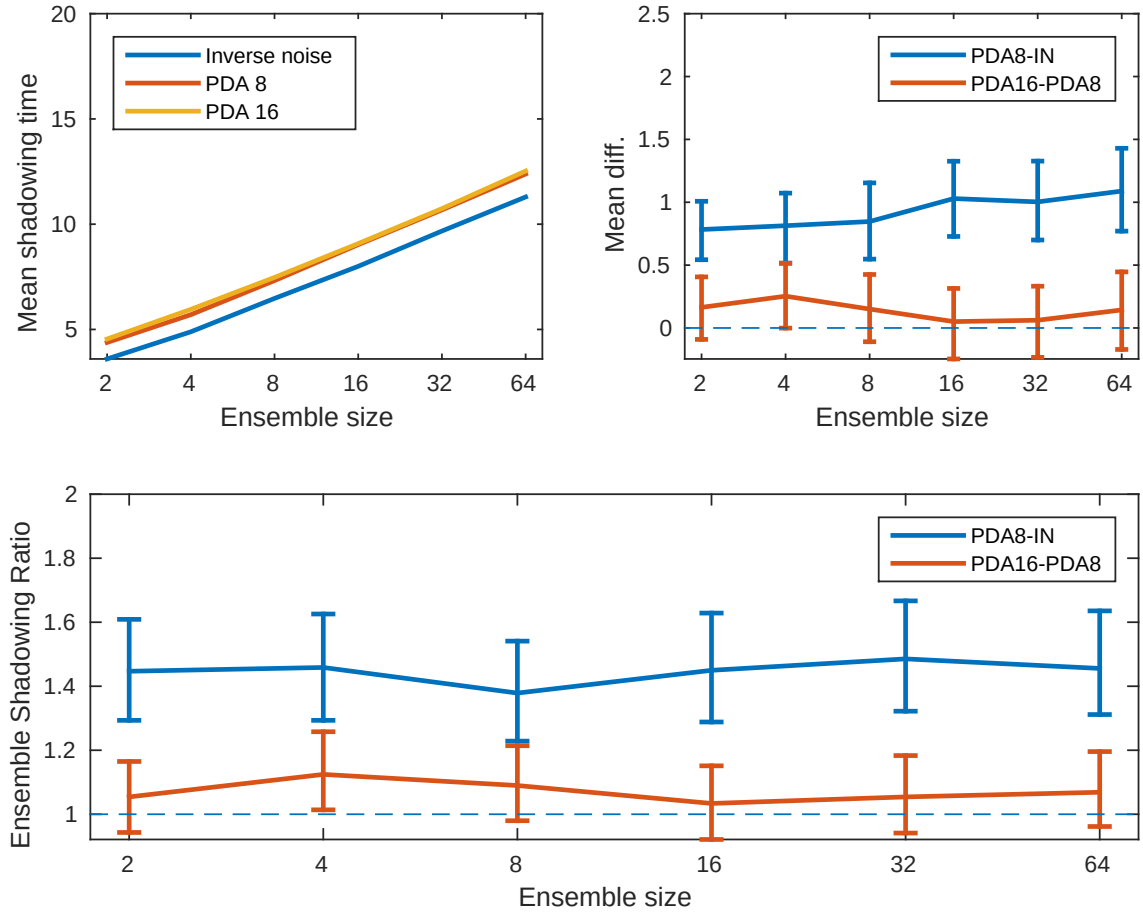

Figure 4.3: For experiment 4.B. Top left: Mean shadowing times of ensembles formed using inverse noise (blue), PDA with an assimilation window of 8 (red) time steps and PDA with an assimilation window of 16 time steps (yellow) as a function of ensemble size. Top right: Mean differences between shadowing times of ensembles formed using PDA with an assimilation window of 8 steps and inverse noise (blue) and PDA with assimilation windows of 16 and 8 time steps (red) as a function of ensemble size with 95% resampling intervals. Bottom: ensemble shadowing ratios between PDA with a window of 8 time steps and inverse noise (blue) and PDA with windows of 16 and 8 time steps (red) with 95% resampling intervals of the proportion as a function of ensemble size. For each given ensemble size, both the mean ensemble shadowing time and the shadowing ratio suggest that PDA performs better, on average, than inverse noise. Doubling the assimilation window, however, appears to make little difference to the performance of PDA in this case.

which each of the forecast densities has zero skill from one in which some of the forecasts have skill that is cancelled out by the poor skill of other, badly performing, forecasts. We now introduce a new measure, that may or may not be used alongside the ignorance score, aimed at differentiating such cases.

To define boosted probability, we first define enhanced probability. We say that, at a time t, a forecast density p_t *enhances probability* when $p_t(y_t) > p_{clim}(y_t)$, i.e. the forecast density places more density on the outcome than the climatology. We say that a sequence of forecast densities formed using the same ensemble *boosts probability* at a given lead time if it enhances probability at all observed times up to and including that time. We define a boosted probability time B_p to be the longest lead time at which a sequence of forecast densities boosts probability.

Formally, a boosted probability time is defined as

$$B_p = \max(T | p_t(y_t) > p_{clim}(y_t) \ \forall t \leq T) \tag{4.6}$$

where $p_t(y_t)$ is the probability placed on the outcome y_t at time t by a conditional forecast density and $p_{clim}(y_t)$ is the density placed on the outcome by the unconditional climatological distribution. Boosted probability is closely related to the ignorance score. A forecast density can be said to boost probability at a lead time of τ if

$$\text{ign}(p_t, y_t) < 0, \ \forall t \leq \tau \tag{4.7}$$

where $\text{ign}(p_t, y_t)$ is expressed relative to the skill of the climatology.

Boosted probability gives a different perspective on the performance of a set of forecasts. The ignorance score (relative to climatology) can be interpreted as the

expected return, in bits, of an investor in possession of a set of forecast densities compared to a house in possession of only the climatology. This means that, if an investor uses forecast densities that perform worse than the climatology according to this measure, she would expect to lose money in the long term. Suppose, however, that the investor is able to insure against all losses (and pays a favourable rate to do so). She should now be less concerned about poorly performing forecasts and thus may be inclined to invest since she is only worried about how much she profits from those densities that perform better than climatology. Boosted probability allows her to identify situations in which she can profit.

Boosted probability could also play a role in identifying state dependency. For example, suppose that, out of a set of forecast densities at a given lead time, some proportion boost probability, but that, overall, the performance according to the mean ignorance score is worse than that of the climatology. Although a punter investing based on such forecast densities would expect to lose money, the fact that some of the forecast densities still boost probability might lead them to try to identify some state dependency in forecast performance which, if successful, could lead to the identification of informative forecast densities before they are evaluated.

4.3.1 Example - Assessing forecast density formation techniques using boosted probability

In this example, we demonstrate how boosted probability can be used to identify cases in which, whilst, on average, a set of forecasts perform worse or no better than climatology, some forecast densities can be expected to outperform it. In this experiment, 2048 ensembles, formed using inverse noise and consisting of 32 members each, are launched using the imperfect model of the Lorenz '63 system defined in

appendix A.2.1 with imperfection parameter $c = 8$. Forecast densities are formed every 6 hours in Lorenz time using two different approaches. In one, which we call forecasting system A, forecast densities are formed using simple kernel dressing whilst, in the other, which we call forecasting system B, they are formed using kernel density estimation with each kernel width found using leave one out cross validation minimising the ignorance score. Boosted probability is then used to compare their relative performance. Details of the experiment, which we call experiment 4.C, are listed in table B.3.

The results of the experiment described above are shown in figure 4.4. Here, the red line, corresponding to the values on the left axis, shows the mean ignorance score of the kernel dressed forecasts as a function of lead time with the error bars representing 95 percent resampling intervals of the mean. The thin blue line, corresponding to the right axis, shows the blending parameter α, also as a function of lead time. The thick solid and dashed green lines represent the proportion of forecast densities that boost probability up to the lead time shown on the x axis formed using forecasting systems A and B respectively. The error bars on each of these represent 90 percent confidence intervals of the proportion. Due to model imperfection, in some cases, the forecast densities formed using kernel density estimation place extremely low density on the outcome and thus the ignorance score is treated (by a computer) as infinite. For this reason, the mean ignorance score of forecast densities formed using forecasting system B can not be represented on the plot.

Boosted probability identifies the difference between forecast densities formed in each way. In forecasting system A, the aim is to optimise the forecasts with respect to the mean of the ignorance score. This means that, for some lead times, whilst it may be possible to form skillful forecast densities from certain ensembles, a value

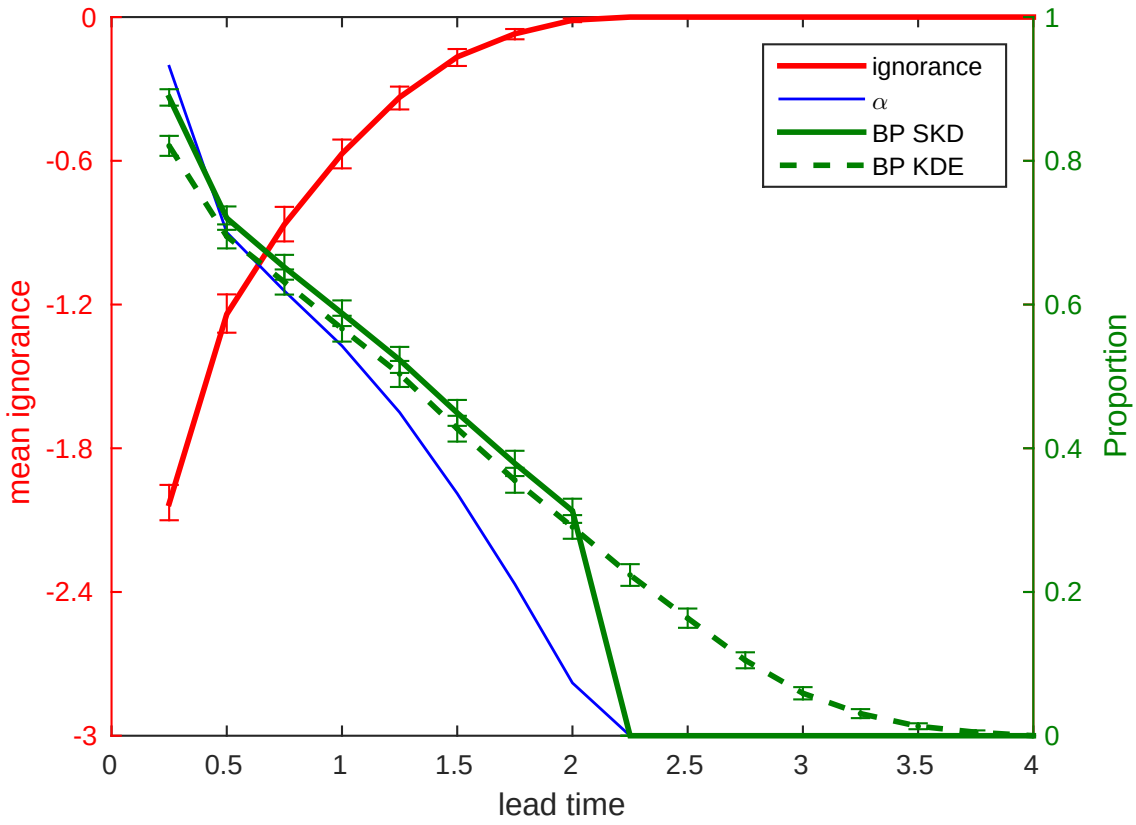

Figure 4.4: As a function of lead time: the mean ignorance score of forecast densities formed using forecasting system A (red line, left axis) with 95 percent resampling intervals, the blending parameter α (blue line, right axis) of forecasting system A and the proportion of forecast densities (with 90 percent confidence intervals) that boost probability for forecasting system A (solid green line, right axis) and B (dashed green line, right axis). Here, boosted probability identifies how, although the forecast densities formed using forecasting system B perform worse than forecasting system A in terms of the mean ignorance score, those formed using the former can sometimes be informative, whilst, for the latter, when $\alpha = 0$, this can never be the case.

of $\alpha = 0$ will be found and hence all forecast densities will have zero skill. Kernel density estimation, on the other hand, takes no account of the skill of the forecasts. This means that some of the forecast densities may actually be informative with respect to climatology, but that, overall, they will perform worse according to the mean ignorance score. This is reflected in the proportion of densities that boost probability. The blending parameter eventually reaches zero and hence, at this point and beyond, none of the forecast densities formed using forecasting system A boost probability. In the case of forecasting system B, on the other hand, a significant proportion of forecast densities continue to boost probability beyond this lead time even though the skill of such forecast densities is cancelled out by the poor performance of other forecast densities. This suggests that there may be some state dependency in the performance of the forecast densities.

Chapter 5

Kernel Dressing methods

In the simple kernel dressing approach to forecast density formation outlined in section 2.5.7, a single kernel width σ is sought which minimises an empirical scoring rule over a training set of ensemble-outcome pairs [131, 22, 143, 146]. If the forecast is blended with climatology, σ, the offset parameter v and the blending parameter α are optimised simultaneously. These parameters are then used to form forecast densities from newly launched ensembles. As long as the parameter values are robust and no major changes have been made to the way in which ensembles are formed and outcomes are collected, forecast densities formed out of sample can usually be expected to perform similarly, on average, to those in the training set. Generally, in simple kernel dressing, a different set of parameters are found for each forecast lead time. This is sensible since, as the time since launch increases, the ensemble members tend to move further apart both from each other and from the system trajectory due to the effects of chaos and/or model error. The parameter values σ and α are thus strongly related to the forecast lead time. Generally, the wider the distribution from which the ensemble members are drawn, the larger the kernel

width that is required to make a good approximation.

In this thesis, we address the fact that ensembles launched at the same lead time can also have significantly different properties. Consider the following scenario. A forecaster launches an ensemble of model simulations of the temperature at Heathrow airport in 3 days time. A large warm front is moving towards the UK and, as a result, the weather is likely to be settled and predictable. Now suppose she makes another 3 day ahead forecast a week later. This time there are two possible weather systems that could move in, one that will likely cause the temperature at Heathrow to be very warm and the other which will result in a relatively mild temperature. As long as the model is able to capture this uncertainty, in the former case, the ensemble members will tend to agree on similar temperatures whilst in the latter, it is likely that there will be significant differences between ensemble members stemming from which of the two possible weather systems will be dominant. The dispersion of the ensemble members in the former case is thus likely to be much lower than in the latter even though both ensembles are evolved forward for the same period of time. Since, in simple kernel dressing, a single kernel width is fitted to all ensembles launched at a given lead time, no account is taken of the information (available from the ensemble itself) which distinguishes this variation. This means that highly dispersed ensembles are fitted with the same kernel width as those with relatively low dispersion. If there is little variation in the dispersion of ensembles at a given lead time, this is unlikely to be a major problem. In section 5.1, we demonstrate, using the Lorenz '63 system, that the dispersion of ensembles can vary significantly, with the starting position on the attractor proving highly influential. As a result, the skill of a set of forecast densities formed using simple kernel dressing is limited by the fact that only a fixed level of smoothing can be applied, even when some ensembles

are far more dispersed than others. In this chapter, we present and investigate three approaches to kernel dressing in which the variation in the dispersion of ensembles is accounted for.

In section 5.3, we present a new method called K groups kernel dressing in which the training set is divided into K distinct groups, each containing ensemble-outcome pairs with similar ensemble standard deviation. The kernel width and blending parameters are then optimised separately for each group resulting in K different sets of parameters in total. A new ensemble is then dressed using the parameters optimised over the group containing ensembles with the most similar standard deviation to its own. We find that, in terms of the ignorance score, this method can outperform simple kernel dressing. We discuss 2 problems that arise with the use of this method. Firstly, each group needs to be large enough such that robust parameter estimates can be found and thus a large training set is usually required. Secondly, assuming that a large training set is available, choosing the number of groups is a non-trivial problem which requires either extensive cross-validation or the use of some other similar approach. This causes the method to be computationally intensive.

In section 5.4, we introduce a second new method called fixed window kernel dressing. With this approach, M ensemble-outcome pairs with similar standard deviation to that of the ensemble to be dressed are selected from the training set. The kernel width and blending parameters are then found by optimising over the M pairs and used to dress the ensemble. This method has an advantage over K groups kernel dressing in that subsets of the training set are selected based on the properties of the ensemble to be dressed and hence the parameters are expected to be more relevant. With this approach, a large number of sets of parameter values generally need to be found whilst choosing the window size, similarly to choosing the number of

groups in K groups kernel dressing, is a non-trivial problem and likewise involves computationally intensive[1] cross-validation.

In section 5.5, we describe a third method called dynamic kernel dressing in which the kernel width is chosen to be a linear function of the ensemble standard deviation. This is a special case of affine kernel dressing [22]. As with simple kernel dressing, the parameters are optimised by minimising an empirical scoring rule over a training set. We show that this method can outperform both K groups and fixed window kernel dressing with a smaller training set. Moreover, no decisions regarding the division of the training set need to be made saving significant computational time. Consequently, we argue that, when the dispersion of a set of ensembles launched at the same lead time varies significantly, out of the methods considered in this thesis, dynamic kernel dressing is the most practical.

In section 5.6, we compare the performance of simple and dynamic kernel dressing in a perfect forecast scenario in which the ensemble is drawn from the same known distribution as the outcome. This approach has a major advantage. Usually, the only available information about the underlying system density[2] is a single draw from it, the outcome. This means that we can't make meaningful evaluations of the performance of a single forecast density because it is impossible to differentiate a case when the approximation of the system density is poor from a case when the outcome falls in the tail of the system density (that is, the occurrence of a truly low probability event). By comparing the forecast density to the system density, we remove this uncertainty and can thus make better judgements about the performance of our

[1] given the computational cost of modern ensemble numerical weather prediction, the added "expense" of all methods suggested in this thesis may be small.

[2] Recall that the system density is defined to be the distribution of the true state given the uncertainty in the initial condition of a forecast. A forecast density can be considered an attempt to recover this distribution. A more detailed discussion of this can be found in section 2.2

Method	Origin σ
Simple Kernel Dressing	Introduced in [22]
K Groups Kernel Dressing	Introduced in the thesis
Fixed Window Kernel Dressing	Introduced in this thesis
Dynamic Kernel Dressing	Special Case of Affine Kernel Dressing [22]

Table 5.1: Summary of the origins of the methods used in this chapter.

forecast densities. We show that, not only does dynamic kernel dressing provide more skillful forecast densities than simple kernel dressing, but that the Kullback-Leibler divergence between the forecast and system densities is smaller, indicating a better approximation to the distribution of the outcome. We also show that forecast densities formed using dynamic kernel dressing can achieve the same level of skill as those formed using simple kernel dressing with a smaller ensemble size. This potentially has significant benefits for applications such as numerical weather prediction in which the computational cost of running each ensemble member can be very high.

In this chapter, only the perfect model and perfect forecast scenarios are considered. Since the ensembles are not expected to have an inherent bias in these scenarios, the offset parameter υ is set to zero. The performance of simple and dynamic kernel dressing in the imperfect model scenario is compared in section 8.1.

In table 5.1, we summarise the origins of the methods used in this chapter.

5.1 Variability in the dispersion of ensembles

Due to the existence of chaos in many nonlinear systems [99], model simulations with slightly different initial conditions are eventually expected to diverge from each other. This means that the further into the future an ensemble is evolved, the more dispersed, on average, the members are expected to be. The level of

dispersion of an ensemble, however, is also usually dependent on the section of the model attractor from which it was launched. This is demonstrated in figure 5.1 in which the standard deviation of 64 member inverse noise ensembles launched around randomly chosen points on the Lorenz '63 attractor are shown as a function of lead time. For each lead time represented on the x axis, the blue dots represent the standard deviation of individual ensembles whilst the red line represents the mean standard deviation over all ensembles. Although, on average, the standard deviation of the ensembles increases with time, there is significant variation between the standard deviation of ensembles at each individual time. For example, at the shortest lead time represented on the plot, although the standard deviation of the majority of the ensembles falls below 2 units, in one case, this value exceeds 12 units.

In the Lorenz 63 system, ensemble members initialised on some parts of the attractor will diverge onto opposite 'wings' whilst on others, all of the ensemble members will remain on the same wing. This will usually be reflected in the dispersion of an ensemble. Two ensembles that illustrate this are shown in figure 5.2. The ensemble coloured in green evolves over a part of the attractor in which there is little or no chance of a transition to the opposite wing. There is, therefore, little disagreement between ensemble members and thus the dispersion is low. The ensemble coloured red, on the other hand, lies on a part of the attractor on which there is substantial uncertainty surrounding this occurrence and so the dispersion is much larger. We can draw an analogy here with the example given at the beginning of the chapter concerning the temperature at Heathrow airport. The green ensemble is comparable to the scenario in which a single weather system is present and thus there is high agreement between ensemble members whilst the red ensemble resembles the alternative scenario in which two competing weather systems cause substantial

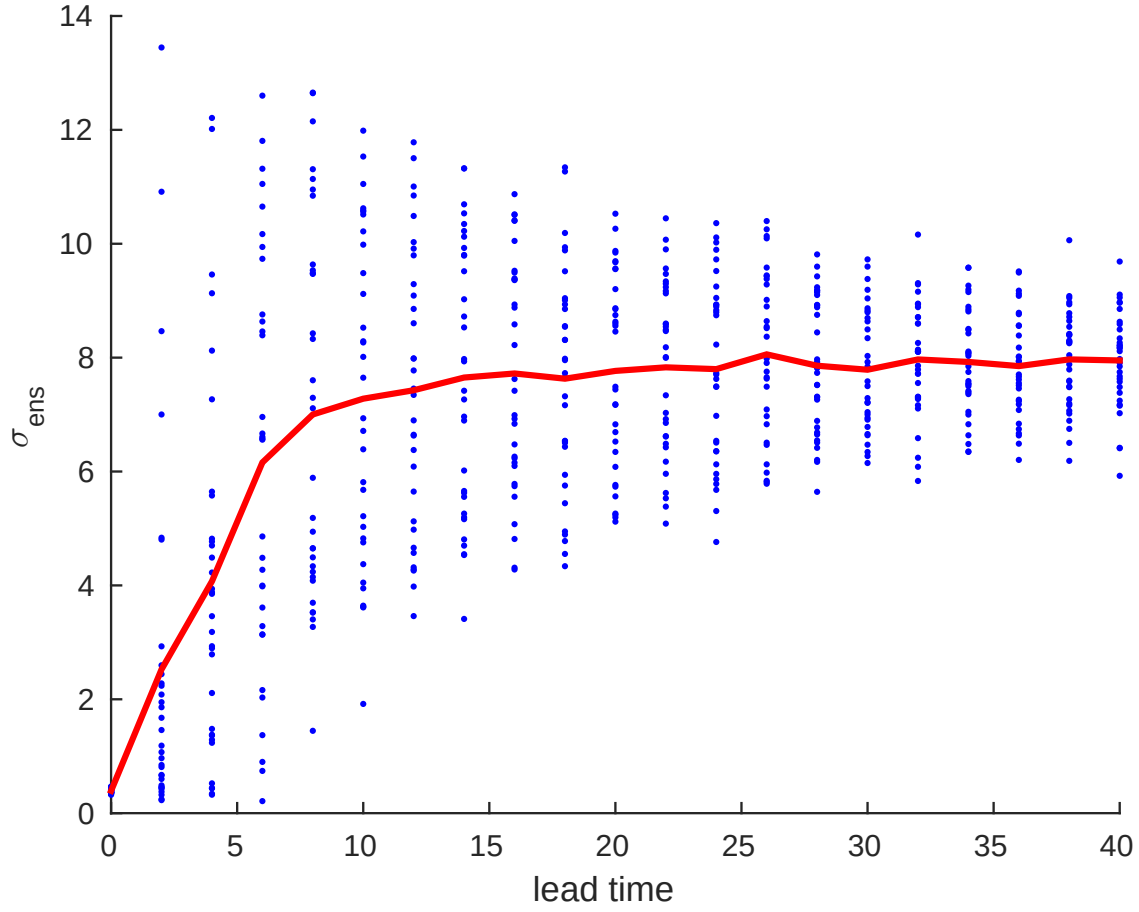

Figure 5.1: A demonstration of how, although, on average, the dispersion of ensembles increases with lead time, there is significant variation in the dispersion at each one. Each blue point represents the standard deviation at a given lead time of a 64 member inverse noise ensemble evolved forward using a perfect model of the Lorenz '63 system. The red line represents the mean standard deviation over all 32 ensembles at each lead time.

uncertainty and hence the ensemble standard deviation is much higher.

5.2 Shortcomings of simple kernel dressing

In this section, we show that, when there is significant variation in the dispersion of ensembles launched at a given lead time, the performance of some forecast densities can be worse than we might expect given the relative position of the ensemble and the outcome. The idea of kernel smoothing, used in both kernel density estimation and simple kernel dressing, is to 'smooth over' the gaps between data points to form density functions with the kernel width controlling the degree of smoothing. Generally, the closer together the ensemble members, the smaller the kernel width required. In kernel density estimation, the kernel width is usually chosen based on the properties of the individual sample and thus the dispersion is taken into account. For example, Silverman's constant [137] chooses a kernel width based on a simple function of the sample standard deviation. In simple kernel dressing, properties of the ensemble to be dressed are not directly taken into account. Instead, all ensembles launched at a given lead time are dressed with the same kernel width. A demonstration of why this can be problematic is shown in figure 5.3. Here, kernel density estimates are derived from samples drawn from logistic distributions, one with parameter $s = 1$ and the other with $s = 5$. The standard deviation of the underlying distribution of the latter is thus 5 times larger than that of the former. In the top two panels, kernel density estimates of both distributions are made using a common kernel width of $\sigma = 0.6$ whilst in the bottom two panels this is repeated but with a kernel width of $\sigma = 2.5$. Whilst each of the underlying distributions can be estimated well with a good choice of kernel width, this cannot be done when the kernel widths are constrained to be equal.

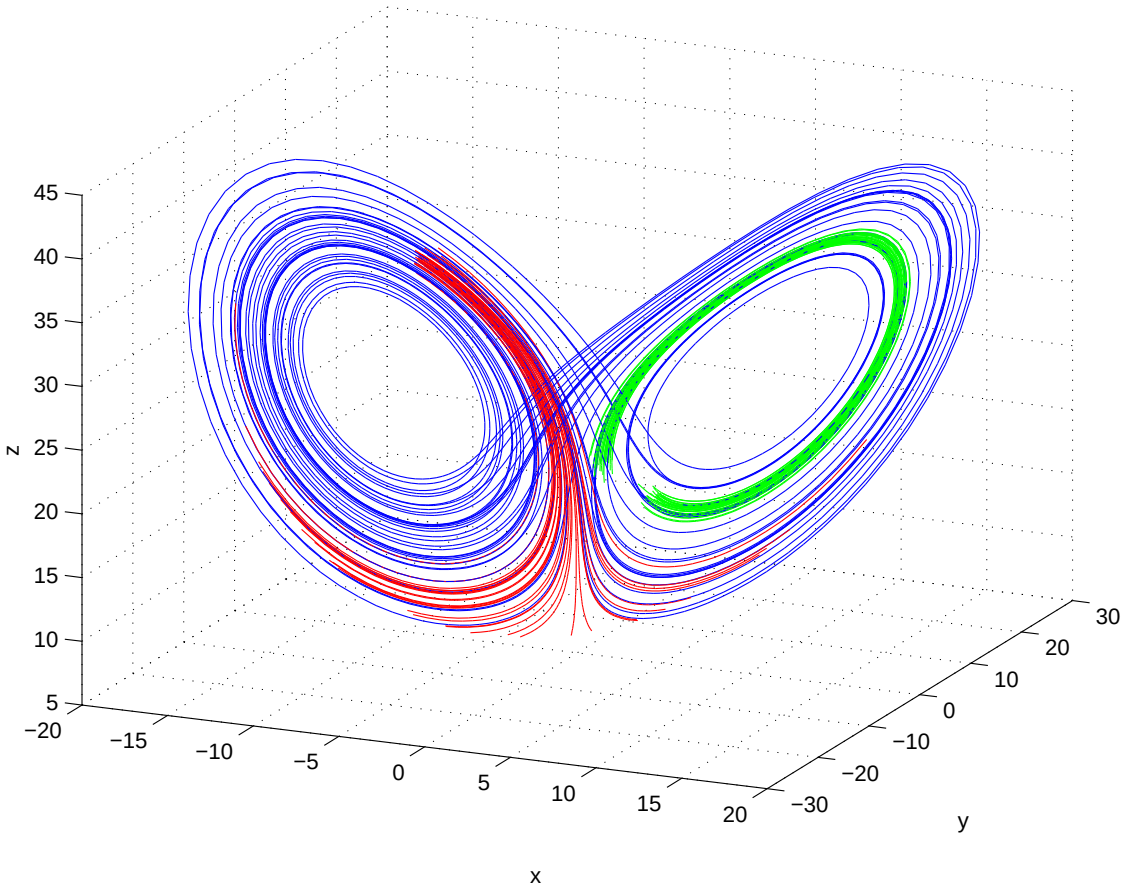

Figure 5.2: Two sets of ensemble members on the attractor of the Lorenz '63 system evolved forward 3.2 days in Lorenz time. The ensemble members coloured green evolve over a section of the attractor in which there is little uncertainty and thus stay close together. The ensemble members coloured in red, on the other hand, evolve over a section of the attractor on which there is significant uncertainty as to whether a trajectory will remain on the same wing or move to the other and thus the ensemble standard deviation is much larger.

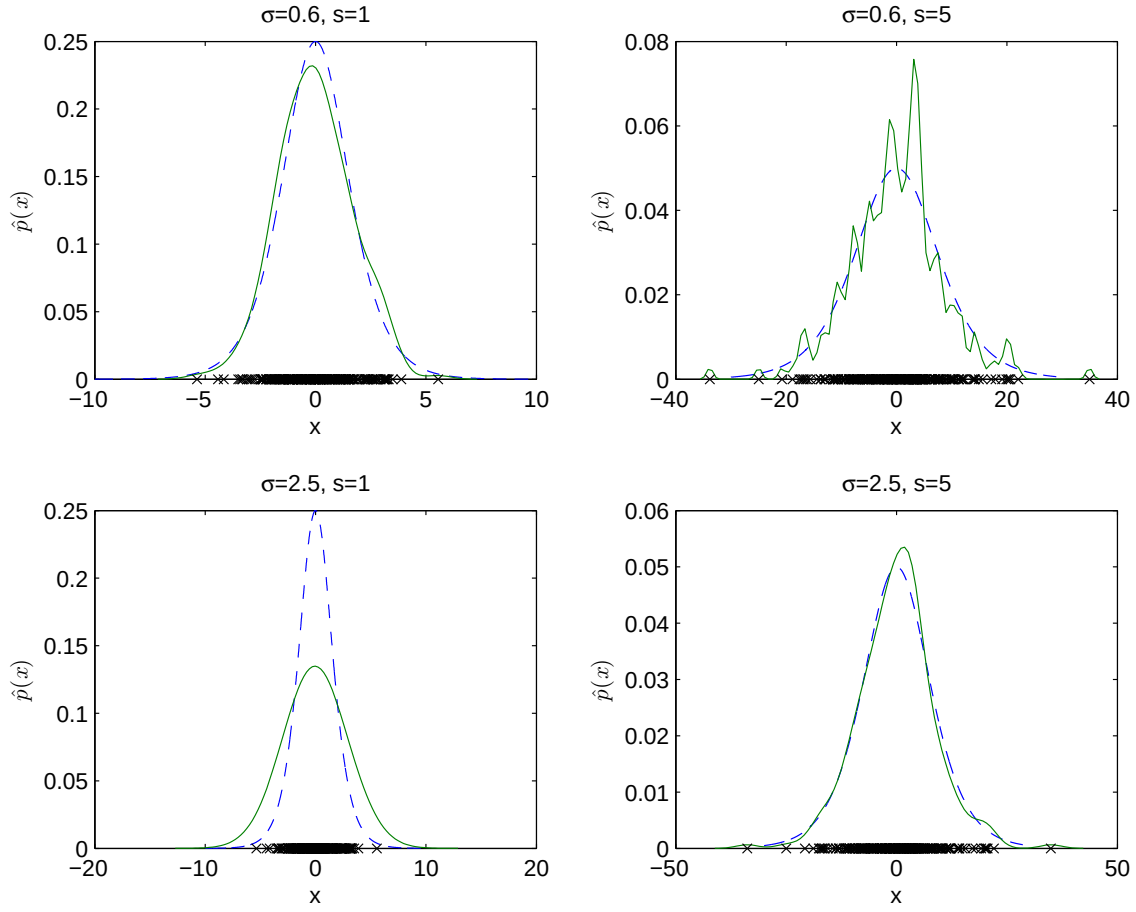

Figure 5.3: Kernel density estimates from samples of size 256 drawn from the logistic distribution with, top left: $s = 1$ and $\sigma = 0.6$, top right: $s = 5$ and $\sigma = 0.6$. Bottom left: $s = 1$ and $\sigma = 2.5$, bottom right: $s = 5$ and $\sigma = 2.5$. In each case, the dashed line represents the true distribution from which the sample was drawn and the green line represents the estimated distribution. The black crosses at the bottom represent the positions of the sample members. Each of the underlying distributions can be estimated well with a good choice of kernel width but this cannot be done when the kernel widths are constrained to be equal.

quantity	value
α	1.000
σ	0.109
minimised value in training set	-4.342
mean ignorance	-4.334

Table 5.2: The optimised values of the blending parameter α and the kernel width σ, the optimised mean ignorance score over the training set and the mean ignorance over the test set from applying simple kernel dressing to the ensembles in experiment 5.A formed using a perfect model of the Lorenz '63 system.

We demonstrate the shortcomings of simple kernel dressing using a perfect model of the Lorenz 63 system. To do this, we form 6.4 day ahead 64 member inverse noise ensembles of randomly chosen states on the Lorenz '63 attractor. From these, we form a training set and a test set both consisting of 2048 ensemble-outcome pairs. Details of the experiment, which we call experiment 5.A, are listed in table B.4. We apply simple kernel dressing, optimising the parameters over the training set, and evaluate its performance over the test set using the ignorance score. The estimated parameter values are shown in table 5.2 along with the optimised mean ignorance from the training set and the mean ignorance of the forecasts formed from the test set.

The blending parameter α is extremely close to unity so our forecasts consist almost entirely of model based information. The resulting forecast densities are highly skillful, on average placing more than 16 times more density on the outcome than the climatological distribution. This is not surprising given that the model is perfect and the lead time is reasonably short. The mean ignorance of forecast densities formed from ensembles in the test set closely resembles the optimised ignorance over the training set so we can be confident that our parameter estimates are robust. Performing very well on average, however, does not necessarily tell the whole story.

A histogram of ignorance scores of forecast densities derived from ensembles in the test set is shown in figure 5.4. Generally, as expected, the forecast densities have very high skill with the vast majority performing far better than climatology. Some, however, have extremely high ignorance implying very poor forecast skill. We shouldn't necessarily be concerned when the climatology performs better than the model based forecasts in a small number of cases. Even if our forecast density describes the system density perfectly, there will be times when the outcome lies in the tail where the density from the climatological distribution may be higher. When this happens, however, we expect our outcome to lie either outside of the range of the ensemble or very close to the edge of it. If this is not the case, our kernel dressing approach may be insufficient in making the most out of the information contained in the ensemble.

We now investigate the reasons for which the ignorance score of some of our forecast densities is so high. The forecast density with the worst ignorance score of all those formed from the ensembles in the test set is shown in figure 5.5 along with the outcome (green circle) and the positions of the ensemble members (black crosses). Although the outcome lies well within the range of the ensemble, the forecast places very little density on it. This appears not to be a result of the relative position of the outcome and the ensemble members but rather the shape of the forecast density. High density is placed in positions close to ensemble members whilst relatively low density is placed in between. This suggests that the ensemble members are undersmoothed.

The explanation for the unusual shape of this forecast density can be found by considering the dispersion of the ensemble. The standard deviation is over 6 times larger than the mean standard deviation of the ensembles in the training set. As we

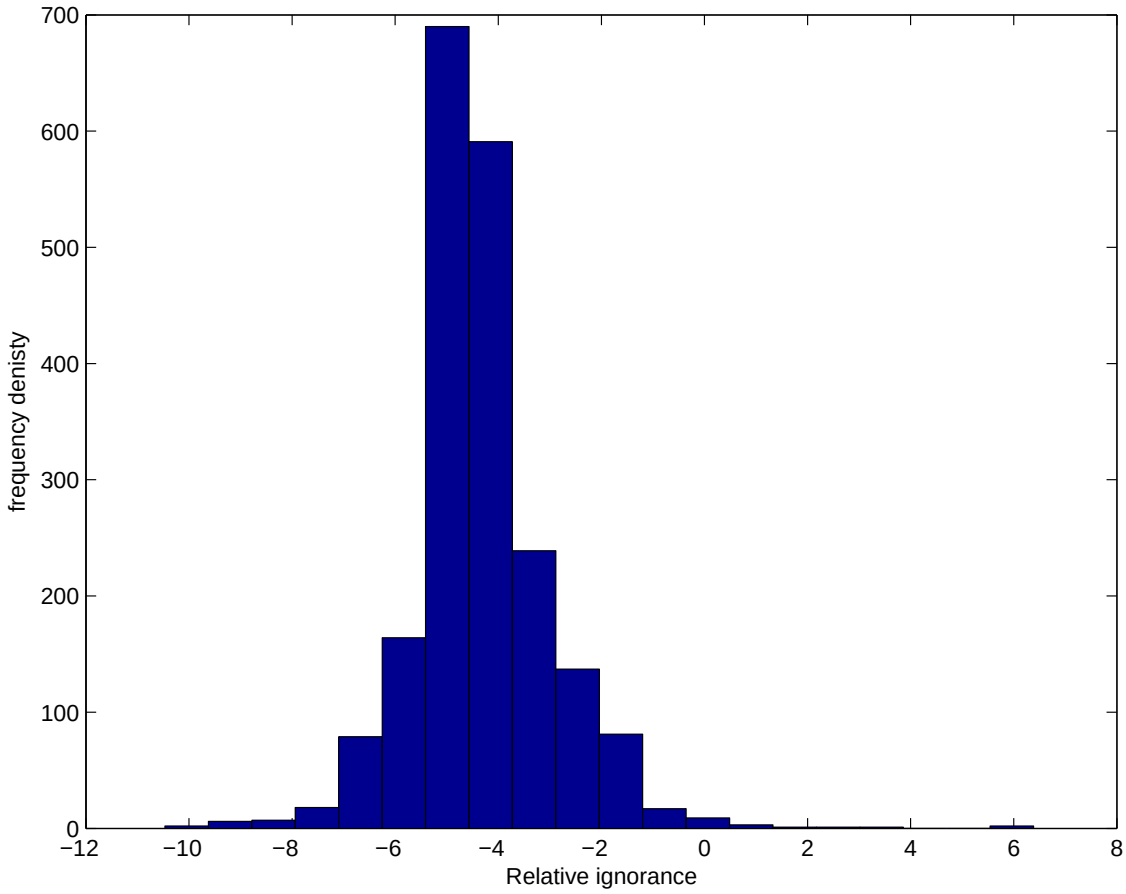

Figure 5.4: Histogram of the ignorance scores of forecast densities formed from ensembles in the test set in experiment 5.A using simple kernel dressing. Although most of the forecasts perform much better than climatology, some have very high ignorance scores.

described in section 5.2, a kernel width that applies a reasonable level of smoothing to ensembles with low dispersion cannot simultaneously provide enough smoothing for ensembles with significantly higher dispersion. The lack of forecast skill thus appears to be as a result of the way in which the forecast density is formed rather than the performance of the ensemble itself.

To illustrate this point further, the forecast density formed from the ensemble in the test set with the highest standard deviation is shown in figure 5.6. Although the shape of the forecast density appears to be unrealistic, in this example, the outcome lies very close to several ensemble members and therefore high density is placed on it resulting in a better ignorance score than in the previous example. Although the forecast density shows relatively high skill, it is not clear whether its proximity to some of the ensemble members is by chance.

The relationship between the dispersion of the ensembles and the skill of the resulting forecast densities is further demonstrated in figure 5.7 in which the standard deviation of each ensemble is plotted against the ignorance of the forecast density formed from it. Whilst ensembles with low standard deviation almost always yield forecast densities more skillful than the climatology, this is much less likely to be the case for forecast densities formed from ensembles with high standard deviation. This is because the spiky shape of the forecasts formed from widely dispersed ensembles makes the ignorance much more sensitive to the precise location of the outcome.

In the following sections, we propose three new kernel dressing methods that account for variation in the dispersion of ensembles with the aim of providing more informative forecast densities.

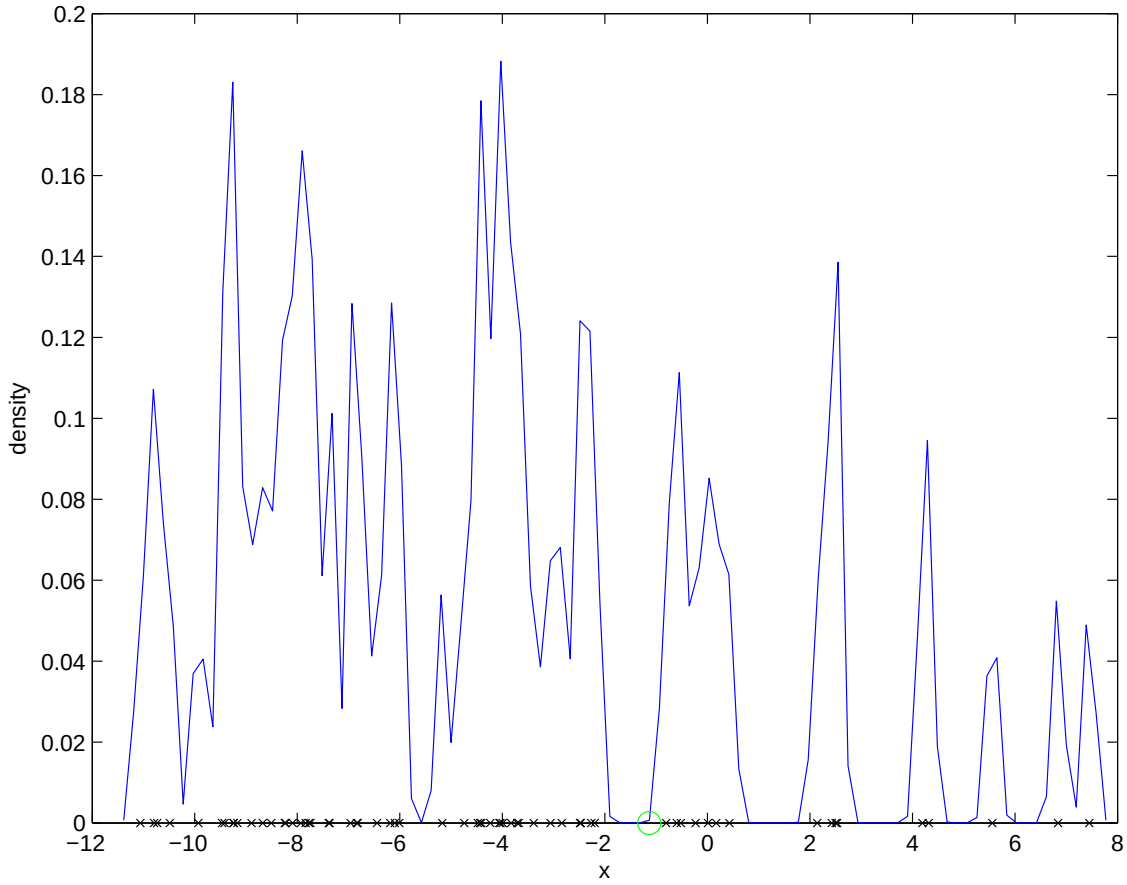

Figure 5.5: The forecast density with the worst ignorance score of all those formed from ensembles in the test set of experiment 5.A using simple kernel dressing. The ensemble and the outcome are represented by the black crosses and the green circle respectively. The shape of the forecast density suggests that the kernel width is too small and hence the level of smoothing is insufficient. The ignorance in this case is particularly high because of the relatively large distance between the outcome and the nearest ensemble member to it.

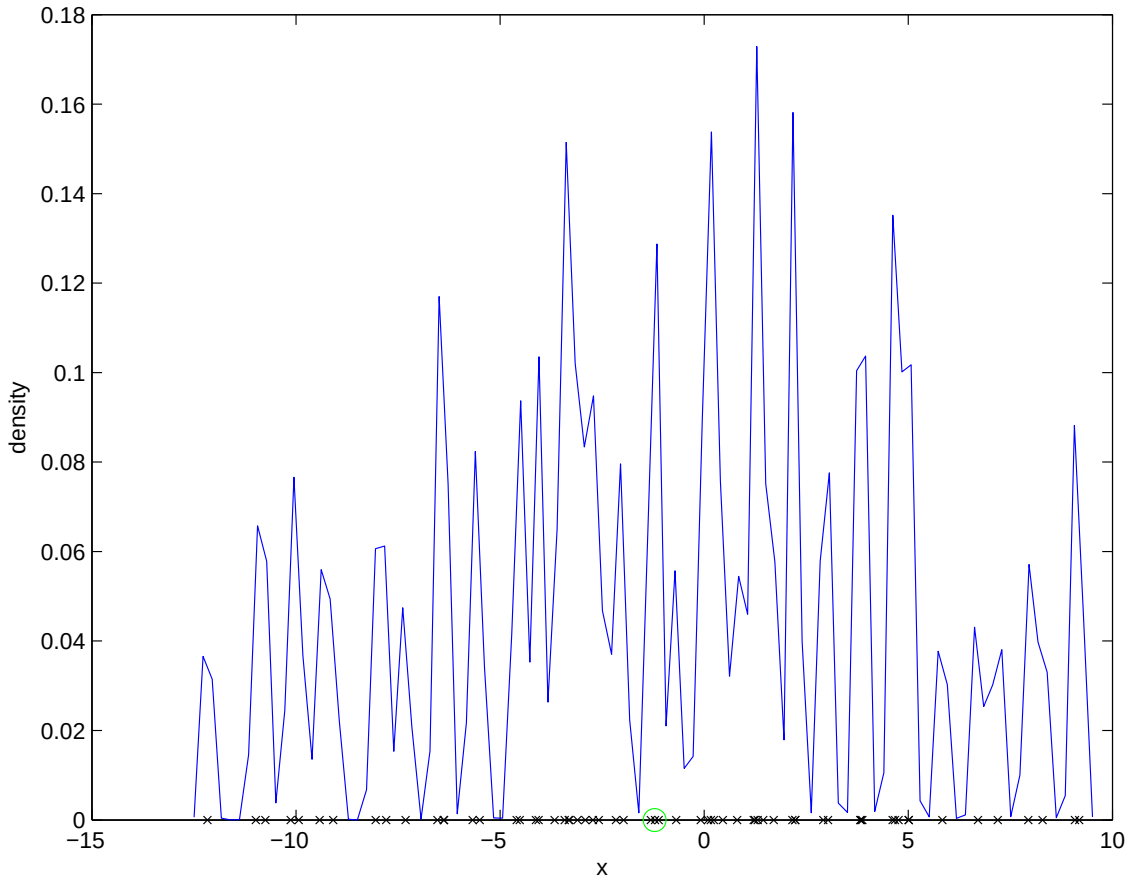

Figure 5.6: The forecast density formed using simple kernel dressing from the ensemble with the highest standard deviation of all those in the test set in experiment 5.A. The ensemble and the outcome are represented by the black crosses and the green circle respectively. As in figure 5.5, it appears that the kernel width is too small. In this case, however, the outcome happens to lie much closer to the nearest ensemble member and hence the ignorance score is much better.

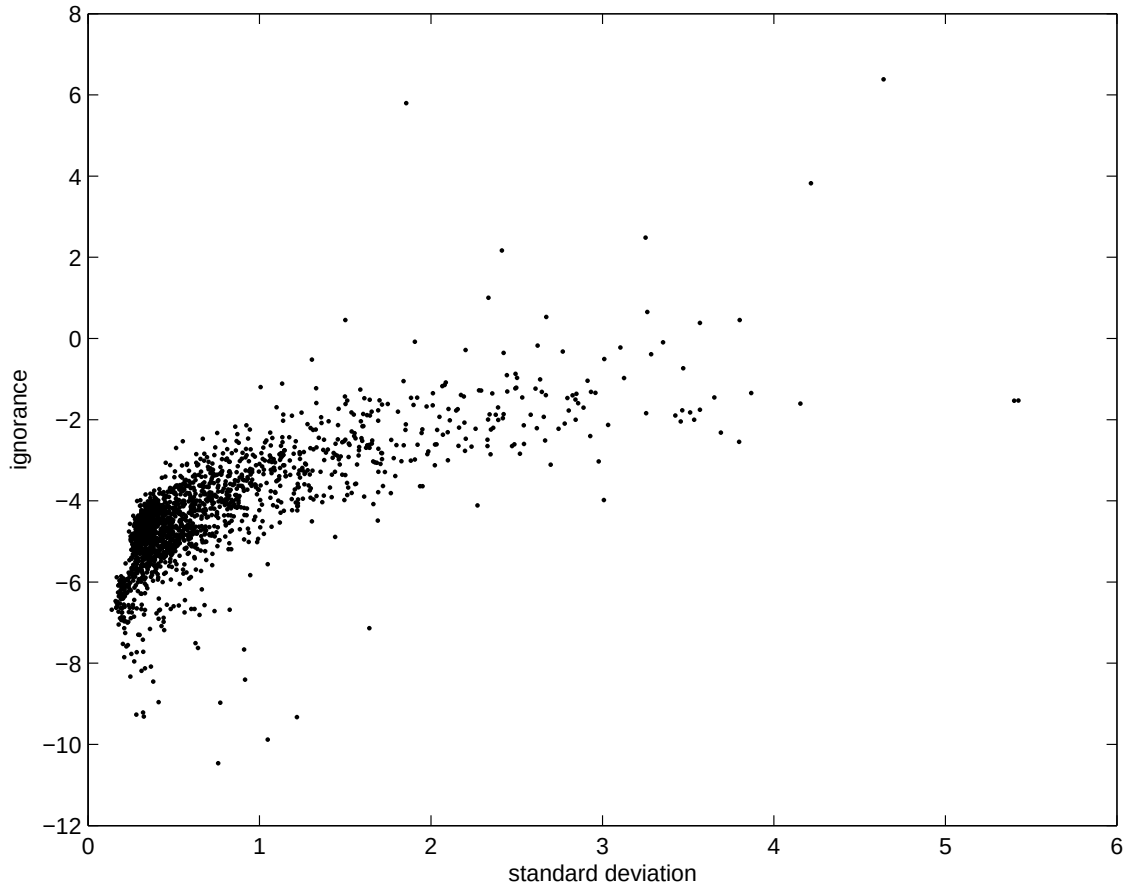

Figure 5.7: Scatter plot of the standard deviation of ensembles in the test set against the ignorance of the forecast densities formed using simple kernel dressing in experiment 5.A. Ensembles with high standard deviation are more likely to yield forecast densities that perform worse than climatology.

5.3 K groups kernel dressing

In the previous section, we showed that, in simple kernel dressing, when the dispersion of ensembles at a given lead time varies significantly, forecast densities formed from the most dispersed can be unrealistic. Motivated by this observation, we introduce a new approach called *K groups kernel dressing*. With this method, the training set is divided into K different groups and different kernel dressing parameters are found for each one. A new ensemble is then dressed with the parameters corresponding to the group containing the ensemble in the training set with the most similar standard deviation to its own.

Let $P_{(1)}, .., P_{(N_{tr})}$ represent ensemble-outcome pairs in the training set listed in ascending order of the standard deviation of the ensemble such that $P_{(1)}$ and $P_{(N_{tr})}$ consist of the ensembles with the smallest and largest standard deviations and their corresponding outcomes respectively. Let $G_1, .., G_K$ represent K mutually exclusive subsets, called groups, of the training set, where G_1 contains N_1 pairs, G_2 contains N_2 pairs and so on. Define the first group to be $G_1 = \{P_{(1)}, .., P_{(N_1)}\}$ such that it contains the N_1 ensemble-outcome pairs with the smallest ensemble standard deviation. The second group is then defined as $G_2 = \{P_{(N_1+1)}, .., P_{(N_1+N_2)}\}$ and so on up to the K_{th} group which is defined as $G_K = \{P_{(N_{tr}-N_K+1)}, .., P_{(N_{tr})}\}$. Simple kernel dressing parameters are then optimised over each group resulting in $\{\upsilon_1, .., \upsilon_K\}$, $\{\sigma_1, .., \sigma_K\}$ and $\{\alpha_1, .., \alpha_K\}$ where υ_i. σ_i and α_i are the offset, kernel width and blending parameters respectively of the i_{th} group. A new ensemble is then dressed with the dressing parameters obtained from the group that contains the ensemble with the smallest absolute difference from its own. Although the size of the groups is up to the user, it is adviseable for each one to contain a similar number of pairs.

As an example of how the parameter estimates vary between groups, consider the sets of ensemble-outcome pairs defined in experiment 5.A. Taking a value of $K = 2$ and thus dividing the training set into 2 equally sized groups of 1024 ensemble-outcome pairs each, the mean ignorance scores over G_1 and G_2 are shown as a function of their kernel widths in the lower left and lower right panels of figure 5.8. Similarly, the mean ignorance over the entire training set as a function of σ is shown in the upper panel. Clearly, σ_1, that is the optimal kernel width over G_1, is much smaller than σ_2, the optimal kernel width over G_2, whilst the optimal width over the entire training set lies between the two. In all cases, in finding the kernel width, a balance must be found between the performance of forecasts derived from the most dispersed and the least dispersed ensembles. By dividing the training set into smaller groups and hence reducing the range of ensembles over which the parameters are found, much less of a compromise is required and thus more relevant parameter values can be found.

The upper panel of figure 5.9 demonstrates how the training set is divided into groups for different values of K. Each dark blue line represents the standard deviation of an ensemble whilst the longer light blue, green and red lines define the boundaries that determine which set of parameter values should be used to dress a new ensemble for $K = 2$, $K = 4$ and $K = 8$ respectively. For example, if we use choose K to be 2 and the standard deviation of a new ensemble lies to the left of the light blue line, it is dressed using the parameters obtained from G_1. The optimal kernel widths for each group are shown in the lower panel aligned with the boundaries in the top panel and colour coded accordingly. Once again, it is clear that the optimal kernel width increases with the dispersion of the ensembles contained in each group.

In order to determine whether K groups kernel dressing performs better than simple

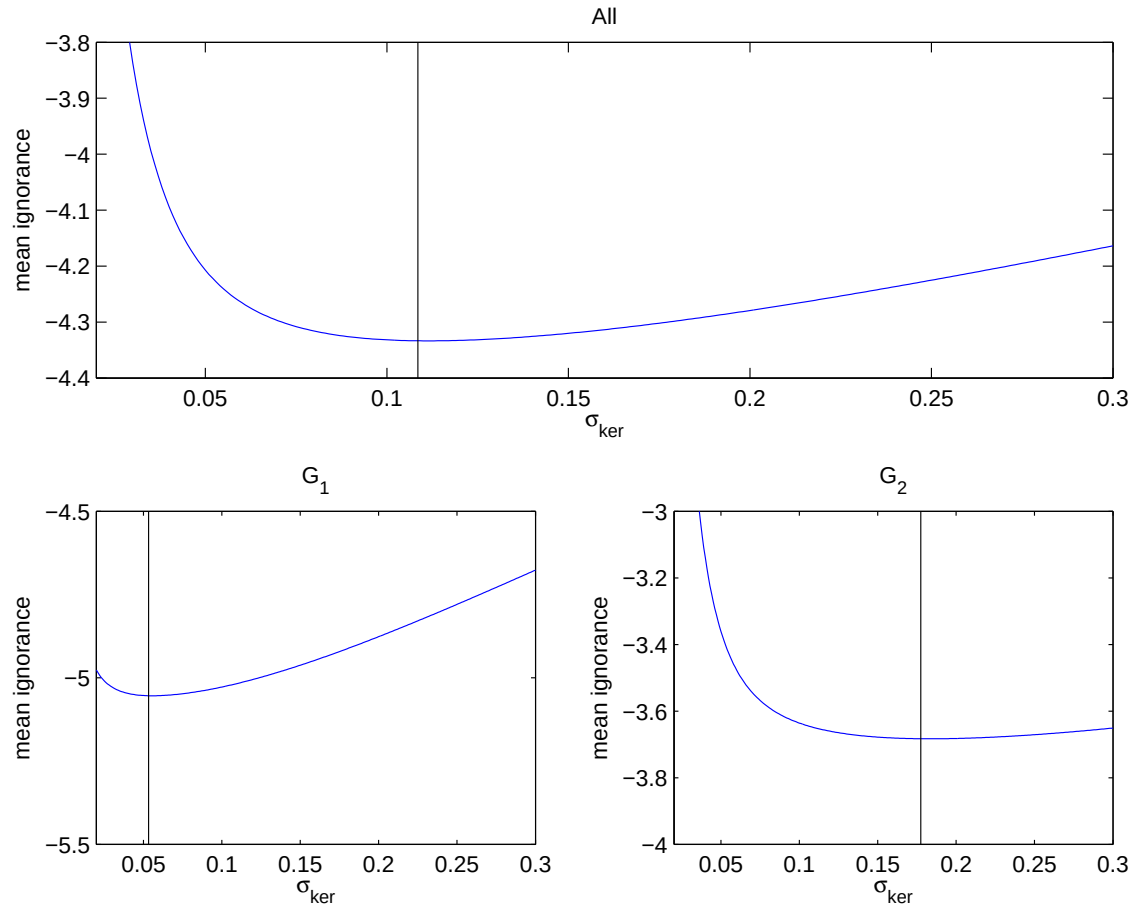

Figure 5.8: The mean ignorance as a function of the kernel width for the entire training set (top), G_1 (bottom left) and G_2 (bottom right) in K groups kernel dressing for the case when $K = 2$ for the ensemble-outcome pairs defined in experiment 5.A. In each case, the optimal kernel widths, represented with black vertical lines, are different.

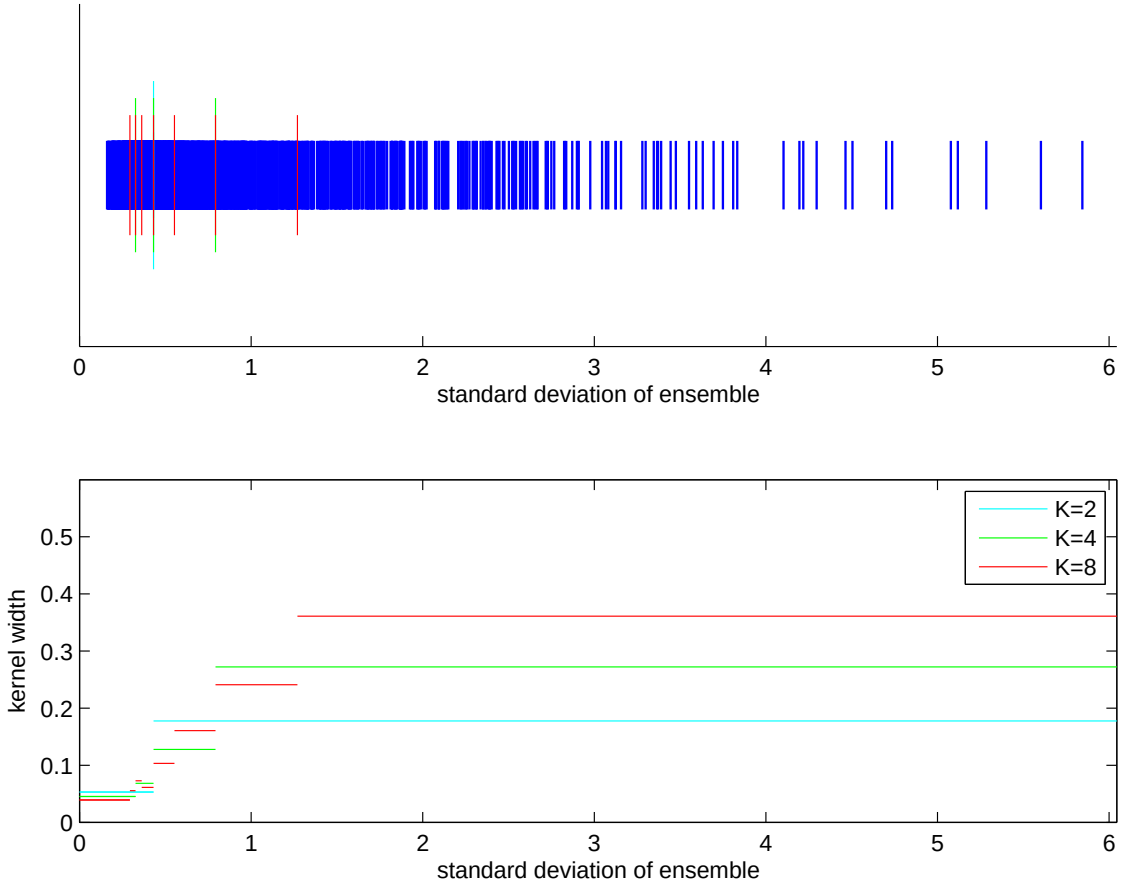

Figure 5.9: Top: The standard deviation of ensembles in the training set (short dark blue lines) of experiment 5.A along with lines indicating the boundaries that determine which set of parameters should be used to dress a new ensemble for $K = 2$ (light blue), $K = 4$ (green) and $K = 8$ (red) in the K groups method. Bottom: Kernel widths chosen to dress a new ensemble with a given standard deviation for $K = 2$ (light blue), $K = 4$ (green) and $K = 8$ (red).

K	Mean ignorance	Resampling interval
2	-0.033	(-0.470,-0.023)
4	-0.045	(-0.062,-0.032)
8	-0.048	(-0.066,-0.033)
16	-0.049	(-0.066,-0.034)
32	-0.036	(-0.055,-0.021)
64	-0.031	(-0.051,-0.015)
128	0.007	(-0.017,0.036)
256	0.150	(0.077,0.342)
512	0.466	(0.289,0.810)

Table 5.3: Mean ignorance scores relative to simple kernel dressing of forecast densities formed from the test set of experiment 5.A using K groups kernel dressing for different values of K. Also shown are resampling intervals of the mean relative ignorance. Since, for the first 6 values of K considered, zero does not fall into the resampling intervals, K groups kernel dressing yields significantly more skillful forecasts than simple kernel dressing for these values of K.

kernel dressing over the ensemble-outcome pairs defined in experiment 5.A, we compare their performance using the ignorance score. Recall that the mean ignorance score of the forecasts formed using simple kernel dressing was -4.334. The mean ignorance scores obtained using K groups kernel dressing for various values of K are shown in table 5.3. The most effective number of groups out of those considered appears to be 16 since, on average, this yields the most skillful forecasts. Note how increasing the number of groups beyond this causes the ignorance to rise. The reason for the drop in skill for larger values of K is that as the number of groups increases, the number of ensemble-outcome pairs in each falls, causing the parameter estimates to be less robust. Therefore a trade off must be made between the number of groups and the number of ensemble-outcome pairs contained in each. An optimal value of K thus exists which minimises the ignorance score of the forecasts. Whilst it would be computationally intensive to test all values of K, it is desirable for its value to be chosen such that it is as close to the optimal number as possible.

Taking a value of $K = 16$, optimal kernel widths for each group (red circles) are compared with the single width obtained using simple kernel dressing (blue crosses) in the upper panel of figure 5.10. The mean ignorance over each group for each method is shown in the lower panel using the same markers. Here, it is clear how the kernel width obtained using simple kernel dressing overestimates the kernel width for ensembles with low dispersion and underestimates for ensembles with high dispersion. The result is that, within these groups, the mean ignorance of the forecast densities formed using K groups kernel dressing is much better than of those formed using simple kernel dressing. For ensembles with less extreme dispersion, the difference between forecast densities formed using each method is minimal. Therefore the improvement in forecast skill achieved by using K groups rather than simple kernel dressing appears almost entirely to result from ensembles with either very high or very low dispersion.

5.3.1 Choosing the value of K

As discussed in the previous section, the most effective choice of K in K groups kernel dressing is a trade-off between the number of groups and the number of ensemble-outcome pairs contained in each. In our example, we determined that the optimal number of groups out of those considered was 16 since this provided the most skillful forecasts. In practice, however, we are not usually able to test the parameters out of sample and thus we need to determine the number of groups purely from the training set. We cannot, however, simply choose the value of K that yields the best skill over the training set. This is because, although the ignorance over the training set will usually be a decreasing function of the number of groups, forecast densities formed from ensembles in the test set will not continually improve in the same way. This problem is sometimes known as over training [7] but is similar

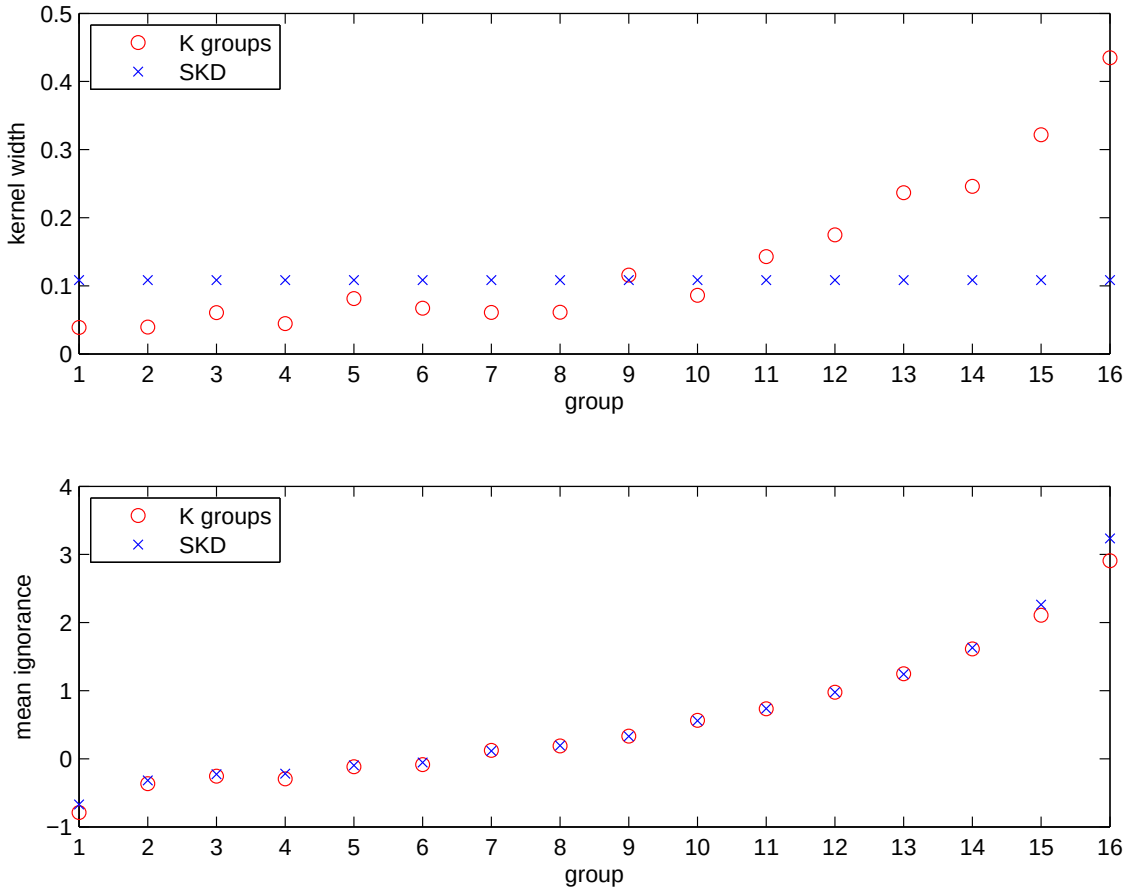

Figure 5.10: Top: Fitted kernel widths for ensembles in the test set of experiment 5.A that fall into each group using K groups kernel dressing with $K = 16$ (red circles) and simple kernel dressing (blue crosses). Bottom: Mean ignorance of forecast densities formed from ensembles that fall into each group using K groups method (red circles) and simple kernel dressing (blue crosses). The increased skill achieved from using K groups rather than simple kernel dressing comes mainly from forecast densities derived from the most dispersed and least dispersed ensembles.

to the more well known problem of over fitting [59] in which too many parameters are fitted to a model for it to generalise out of sample. A common approach to problems like this is to use cross validation techniques which attempt to assess how well the parameters would generalise to an independent set.

One particular approach is leave one out cross validation in which the parameter values are optimised over all but one of the ensemble-outcome pairs in the training set and tested on the remaining one, repeating so that the parameters are tested on each pair exactly once. This is most effective when the number of pairs in each group is large so that the effect of leaving out one pair is only small. For large training sets, however, this approach is likely to be extremely time consuming and likely too computationally intensive to be feasible in many instances.

An alternative approach is K-fold cross validation [65]. This is done by separating the set of ensemble-outcome pairs into several subsets, optimising over each separately and testing over the others. For example, in 2 fold cross validation,we would divide each group into 2 sets, optimise the parameters over one and test them on the other and vice versa. This, however, cannot give an accurate indication of the optimal number of groups since the number of ensemble-outcome pairs used to optimise the parameters is halved. This approach, therefore, does not reflect the robustness of parameters optimised over the whole group. This problem is demonstrated in figure 5.11 where the optimal mean ignorance over the training set (blue), the optimal mean ignorance over the test set (red) and the ignorance using 2 fold cross validation (green) are shown as a function of the number of groups K. Although the number of groups that minimises the ignorance over the test set is 16, this is not reflected in the cross validated ignorance which suggests an optimal number of 32 groups.

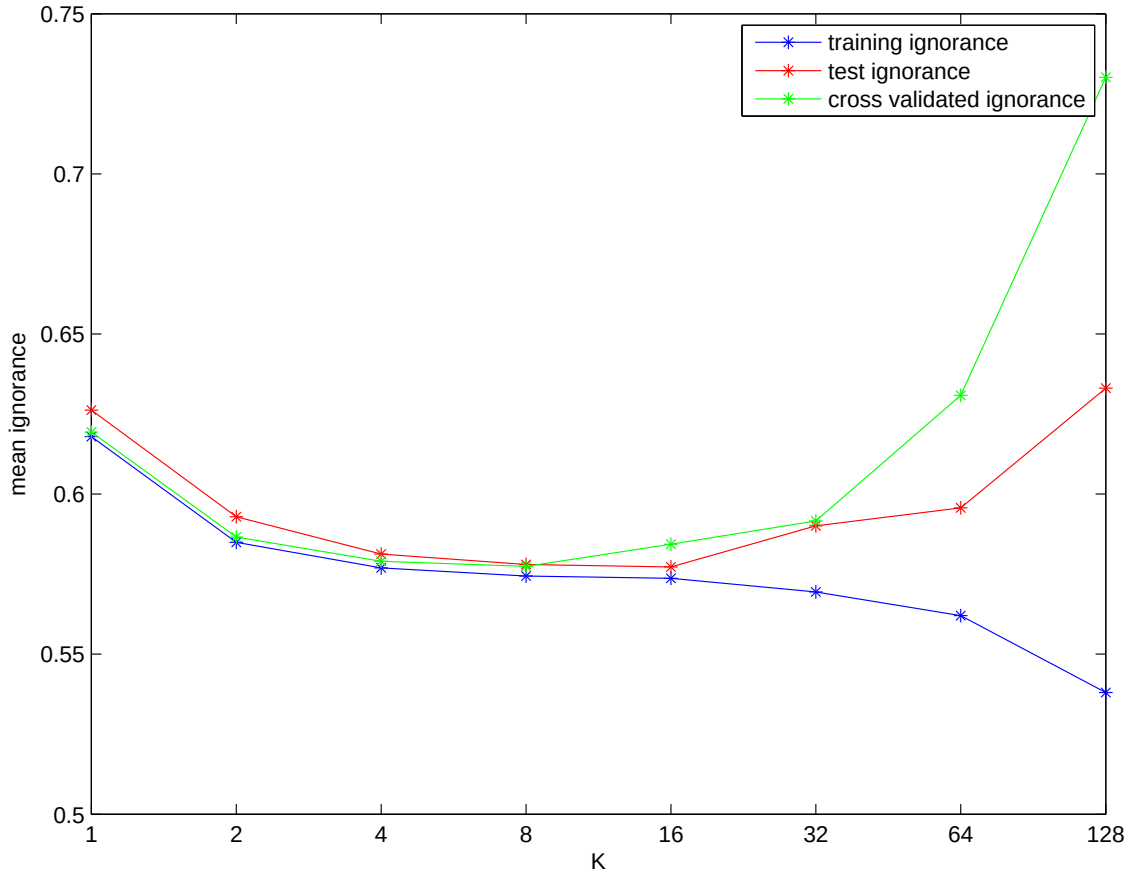

Figure 5.11: The optimised mean ignorance over the training set (blue), the mean ignorance obtained using 2 fold cross validation (green) and the mean ignorance over the test set (red) as a function of K for the ensemble-outcome pairs defined in experiment 5.A. Here, it is clear how 2 fold cross validation fails to identify the optimal number of groups.

5.4 Fixed Window Kernel Dressing

In K groups kernel dressing, the training set is divided into K distinct subsets and kernel dressing parameters are found for each one. A new ensemble is then dressed using the parameters obtained from the group that contains the ensemble with the closest standard deviation to its own. If the standard deviation of the ensemble to be dressed lies close to a group boundary, however, the parameters used to form the forecast density are less likely to be as relevant as for those ensembles whose standard deviation lies close to the middle of a group range.

We now introduce another new method called *fixed window kernel dressing* in which, as with K groups kernel dressing, the parameters to dress an ensemble are optimised over a subset of the training set. The ensemble-outcome pairs contained in the subset are chosen such that the median standard deviation is as close to that of the ensemble to be dressed as possible. As with the K groups method, let $\boldsymbol{P} = \{P_{(1)}, .., P_{(N_{tr})}\}$ represent the ensemble-outcome pairs ordered by ensemble standard deviation from the smallest to the largest. The subset P_{win} of $\boldsymbol{P}$ over which the parameters are optimised is chosen to be the $\frac{M}{2}$ pairs in $\boldsymbol{P}$ with the closest standard deviation that is strictly larger than the standard deviation of the ensemble to be dressed and the $\frac{M}{2}$ with the closest standard deviation that is strictly smaller. The parameters used to dress an ensemble are thus determined by the position of its standard deviation relative to the ensembles in the training set. If there are not $\frac{M}{2}$ pairs in the training set with strictly larger or strictly smaller standard deviation, extra pairs are taken with the opposite property. For example, if only one ensemble in the training set has larger standard deviation than the ensemble to be dressed, an additional $\frac{M}{2} - 1$ pairs with strictly smaller standard deviation are included in the window instead.

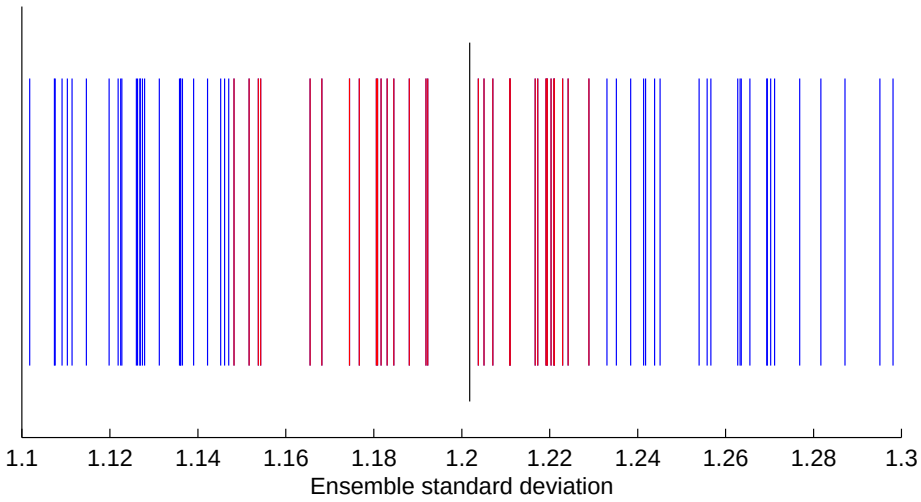

Figure 5.12: An example of how the subset of the training set is selected in fixed window kernel dressing for the case $M = 32$. The long black line represents the standard deviation of the ensemble to be dressed while the smaller lines represent the standard deviation of the ensembles in the training set. Those that are coloured red represent the ensemble-outcome pairs over which the parameter values are optimised.

Once an appropriate subset P_{win} has been found, simple kernel dressing parameters α, υ and σ are found and used to dress the ensemble.

The way in which a window of the training set is chosen is demonstrated in figure 5.12 for an ensemble in our Lorenz 63 PMS example in experiment 5.A using a window size of $M = 32$. Here, the long black line represents the standard deviation of the ensemble to be dressed whilst the shorter lines represent the standard deviation of the ensembles in the training set. Those coloured in red, immediately to the left and to the right of the black line represent the ensemble-outcome pairs over which the parameters are optimised. Those coloured in blue do not fall into the window and thus are not used in the optimisation process.

For a fixed training set of N_{tr} ensemble-outcome pairs, there are $N_{tr} + 1$ possible positions that the standard deviation of a new ensemble can fall relative to the

ensembles in the training set. The kernel widths that would be applied to ensembles falling into each of these positions are shown in figure 5.13 for different values of M. The blue, green and red lines show the fitted kernel width for window sizes of $M = 16$, $M = 128$ and $M = 1024$ respectively. For $M = 16$, the kernel width is volatile with respect to the exact position of the ensemble standard deviation. This suggests that the window size is too small thus failing to produce robust kernel widths with which to dress the ensembles. For $M = 128$, the kernel width is much less volatile although there is still significant movement in the chosen kernel width as the position changes. A window size of $M = 1024$ yields a smoothly increasing kernel width with movement from one position to the next having little effect on the optimised parameters. Of course, if M were increased to 2048, the parameters would be optimised over the entire set and the approach would reduce down to simple kernel dressing. Thus, as is the case with K groups kernel dressing, a balance has to be struck between the coverage of the window and the number of ensemble-outcome pairs included within it.

The results of applying fixed window kernel dressing to the ensemble-outcome pairs in experiment 5.A with different values of M are shown in table 5.4 in which the mean ignorance score over the test set is shown for each along with 95 percent resampling intervals of the mean. As with K groups kernel dressing, for some window sizes, significant improvement is made over simple kernel dressing. According to the ignorance score, a window size of $M = 256$ performs the best out of the tested window sizes. Of course, whilst we only test a small number of window sizes, many other sized windows could be chosen, some of which are likely to perform better than those tested.

Fixed window kernel dressing has a significant advantage over K groups kernel

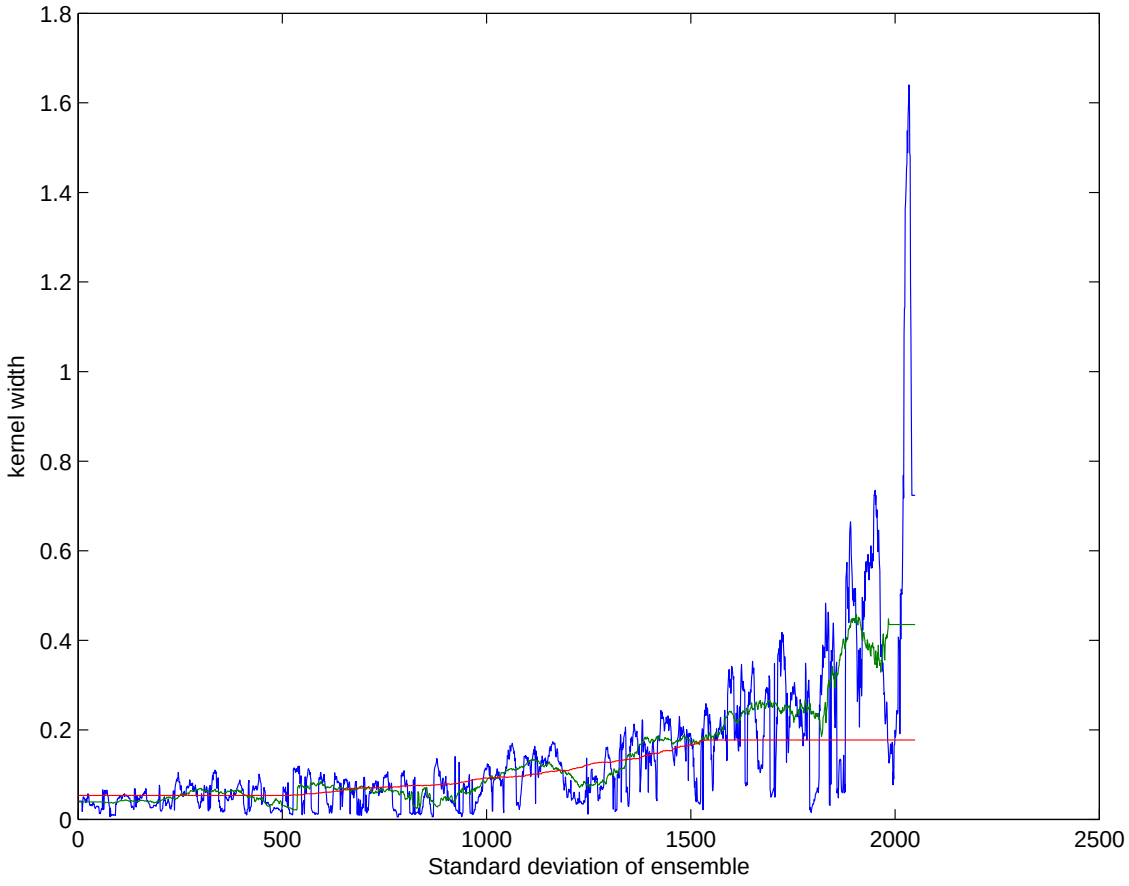

Figure 5.13: The kernel width used to dress an ensemble with a given standard deviation for $M = 16$ (blue), $M = 128$ (green) and $M = 1024$ (red) using fixed window kernel dressing with the ensemble outcome pairs defined in experiment 5.A.

M	Relative ignorance	Resampling interval
16	0.010	(-0.015,0.041)
32	-0.022	(-0.046,-0.004)
64	-0.040	(-0.058,-0.025)
128	-0.047	(-0.066,-0.032)
256	-0.049	(-0.066,-0.035)
512	-0.047	(-0.063,-0.035)
1024	-0.047	(-0.051,-0.030)
2048	0	(0,0)

Table 5.4: Mean ignorance scores relative to simple kernel dressing of the forecast densities formed from the test set of experiment 5.A using fixed window kernel dressing for different values of M. Also shown are 95 percent resampling intervals of the mean relative ignorance in each case. Since zero does not fall into these intervals, a significant improvement is made over simple kernel dressing for all values of M considered except $M = 16$. The most effective value of M out of those considered appears to be 256.

dressing. In the latter, the training set is divided into distinct groups and hence each ensemble-outcome pair can only occur in one. This means that when the ensemble to be dressed lies close to the boundaries of a group in terms of its standard deviation, the optimised parameters are likely to be less relevant than if it lies in the middle. In fixed window kernel dressing, the standard deviation of the ensemble to be dressed usually lies close to the median of the set over which the parameters are optimised and hence the parameters can be expected to be more appropriate.

Like in K groups kernel dressing, choosing the optimal window size is likely to be problematic. If K-fold cross validation were used, the parameters would need to be optimised over subsets of each group and would thus fail to give an accurate assessment of the robustness of parameters with each window size. Leave one out cross validation, on the other hand, is computationally intensive.

5.5 Dynamic Kernel Dressing

We have shown that when there is significant variation in the dispersion of ensembles at a given lead time, forecast densities formed using K groups and fixed window kernel dressing can perform significantly better, on average, than those formed using simple kernel dressing. As we have explained, however, both of these methods have significant limitations rendering them impractical in many cases.

We now consider another method which we refer to as *dynamic kernel dressing*. Under this approach, instead of dividing the training set into subsets, an extra parameter is fitted such that the kernel width is chosen to be a linear function of the ensemble standard deviation. Allowing the kernel width to be a linear function of the ensemble variance was proposed in [22] and thus, although the exact nature of dynamic kernel dressing is new to this thesis, the concept is not.

In dynamic kernel dressing, the kernel width σ_d used to dress an ensemble is chosen using the relation

$$\sigma_d = a + bs, \tag{5.1}$$

where s is the ensemble standard deviation and a and b are parameters to be found. Similarly to simple kernel dressing, the parameters a and b, the offset parameter v and the blending parameter α are optimised simultaneously with respect to the mean ignorance score over the training set. Positive values of b allow larger kernel widths to be applied to more widely dispersed ensembles whilst, when $b = 0$, this approach reduces down to simple kernel dressing with $\sigma_d = a$. The approach in which the kernel width is chosen to be a function of the standard deviation of a sample is also commonly taken in kernel density estimation. For example, Silverman's constant [137] is a particular case of this parameterisation with $a = 0, b = 1.06n^{-5}$ and

α	1.000
a	0.0000
b	0.1995
Optimal ignorance over training set	-4.383
Mean ignorance over test set	-4.386

Table 5.5: The optimised parameter values, the minimised mean ignorance over the training set and the mean ignorance over the test set obtained from applying dynamic kernel dressing to the ensemble-outcome pairs in experiment 5.A.

$\upsilon = 0$. Once again, we demonstrate this approach with the ensemble-outcome pairs in experiment 5.A. The parameter values, the minimised mean ignorance over the training set and the mean ignorance over the test set are shown in table 5.5. Note that, since the model is perfect, we have set the offset parameter υ to zero.

The large value of the gradient term b implies that the kernel width is heavily dependent on the ensemble standard deviation and thus that the inclusion of this parameter is beneficial. The mean ignorance score achieved using this method is better than that obtained using simple kernel dressing (recall that simple kernel dressing yielded a mean ignorance score of -4.334). In figure 5.5, we showed how simple kernel dressing fails to provide a realistic forecast density from the most highly dispersed ensemble in the test set. The forecast density formed from the same ensemble using dynamic kernel dressing is shown in green in figure 5.14 along with, for comparison, the density obtained using simple kernel dressing which is shown in blue. The former appears to be much more realistic than the latter. This is because the kernel width in dynamic kernel dressing takes account of the large ensemble standard deviation and therefore fits a much larger kernel width allowing more smoothing than simple kernel dressing.

We now consider an example from the opposite extreme in which the ensemble standard deviation is very low compared to the average in the training set. The

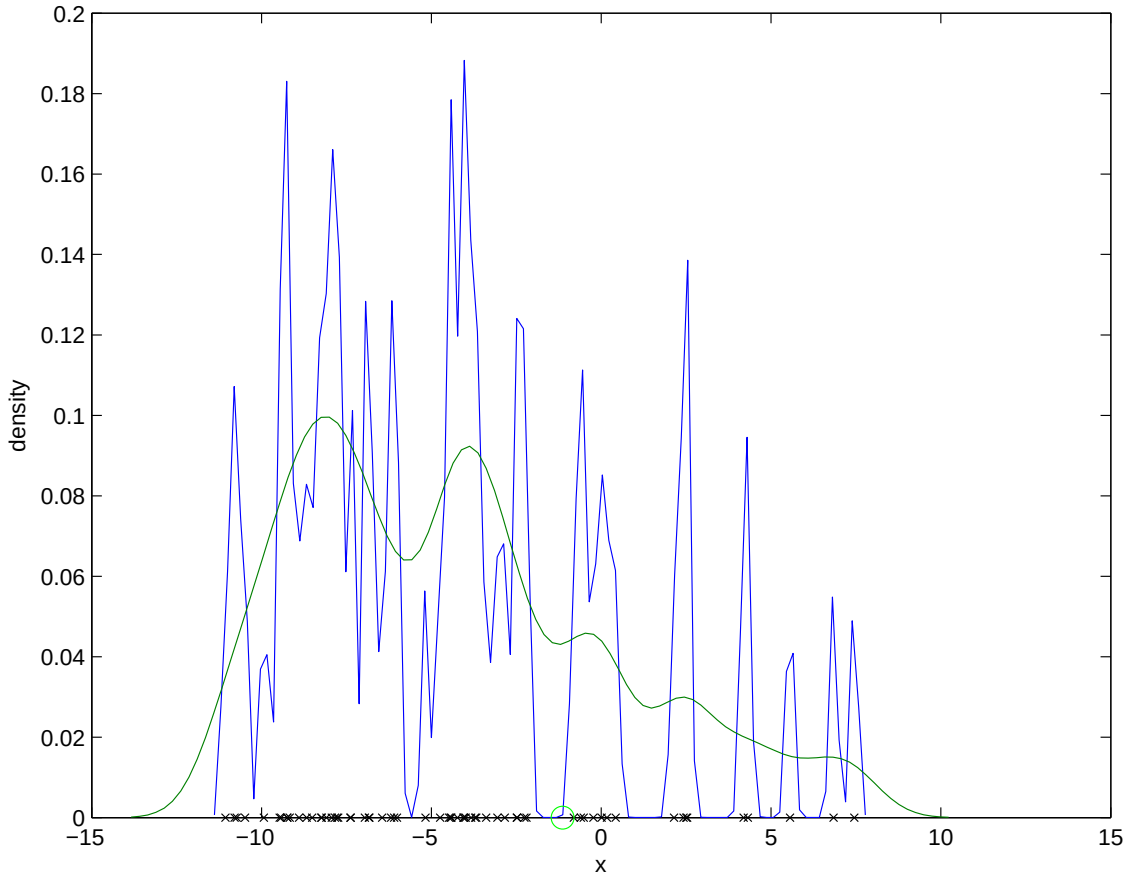

Figure 5.14: Forecast densities formed using dynamic (green) and simple (blue) kernel dressing from the ensemble in the test set of experiment 5.A which produces the highest ignorance score when simple kernel dressing is applied (as shown in figure 5.5). The black crosses along the x axis show the positions of the ensemble members and the green circle the position of the outcome. By taking account of the ensemble standard deviation, dynamic kernel dressing is able to apply a larger kernel width and hence yield a more informative forecast density.

forecast densities formed from the ensemble in the test set with the lowest standard deviation using each method are shown in figure 5.15. In this case, the forecast density obtained from simple kernel dressing is wide and covers a large area outside of the range of the ensemble. This suggests that the kernel width is too large and hence that the data are over smoothed. The forecast density obtained from dynamic kernel dressing, on the other hand, takes account of the fact that the standard deviation of the ensemble is small and therefore applies a much smaller kernel width and thus the forecast density retains the shape of the data much better.

Whilst the above two examples appear to show that dynamic kernel dressing yields more consistent forecast densities than simple kernel dressing, we can only draw conclusions about their performance by comparing the forecast skill over many forecast densities. We have already shown that dynamic kernel dressing outperforms simple kernel dressing in this experiment. We now show, however, that the majority of the improvement in skill comes from the ensembles with either very high or very low standard deviation. A scatter plot of the ignorance scores obtained from each method for all of the ensembles in the test set is shown in figure 5.16. Each of the points is coloured according to the ensemble standard deviation with warmer colours denoting the ensembles with the largest standard deviation. Those points that fall above the blue line indicate ensembles in which dynamic kernel dressing yields a better ignorance score than simple kernel dressing whilst the opposite is true for those that fall below it.

The forecasts with the highest ignorance scores in both cases tend to be those derived from the most dispersed ensembles. This is not surprising since the most dispersed ensembles tend to result in wider distributions and so we expect to place less density on the outcomes. For such ensembles, most of the points lie above the blue line and

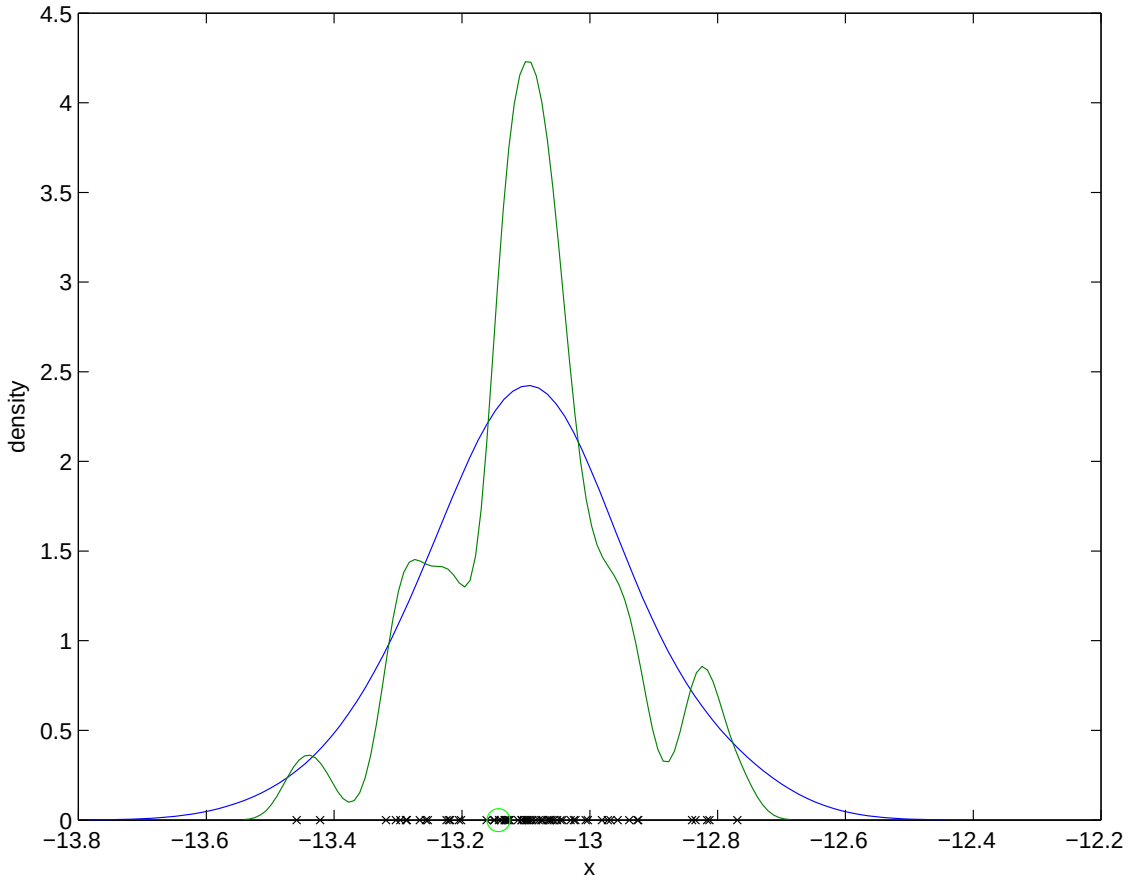

Figure 5.15: Forecast densities formed using dynamic (green) and simple (blue) kernel dressing from the ensemble in the test set of experiment 5.A with the lowest standard deviation. The black crosses along the x axis show the positions of the ensemble members and the green circle the position of the outcome. The forecast density formed using simple kernel dressing places a large amount of density outside of the range of the ensemble members suggesting that the kernel width is too large. The kernel width obtained from dynamic kernel dressing, on the other hand, is smaller and hence the resulting forecast density appears to reflect the distribution of the ensemble members better.

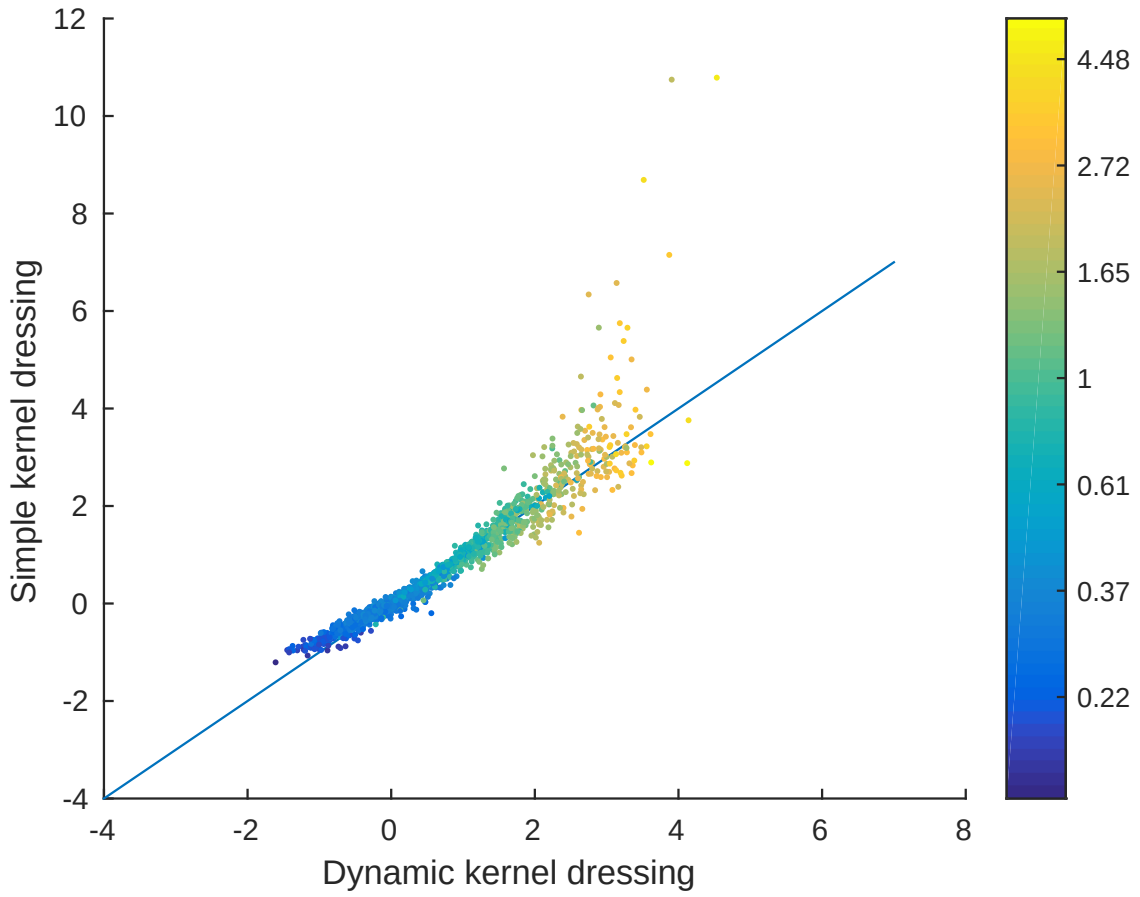

Figure 5.16: Scatter plot of the ignorance of forecast densities formed from the test set of experiment 5.A using simple and dynamic kernel dressing. Each point is coloured according to its ensemble standard deviation with warmer colours indicating those with high standard deviation. The blue line indicates the points at which the ignorance of both methods would be equal. It is clear that most of the improvement in skill yielded from dynamic kernel dressing comes from ensembles with relatively low or relatively high standard deviation since these points are much more likely to lie above the blue line.

hence it is clear that the forecast densities obtained from dynamic kernel dressing tend to perform better than those obtained from simple kernel dressing. In the same way, ensembles with low dispersion tend to result in low ignorance. For these ensembles, again, many of the points lie above the blue line and hence dynamic kernel dressing tends to yield the best forecast densities. Towards the middle of the range, there is little difference between the forecast densities formed using the two approaches. This is because the kernel widths fitted by the two methods are more similar for such cases than for when the ensembles are either very widely or very narrowly dispersed. It appears then that the improvement in the mean ignorance of forecast densities formed using dynamic kernel dressing comes almost entirely from the ensembles with the lowest and the highest standard deviation whilst little difference is made to those in between.

5.5.1 Robustness of dynamic kernel dressing

In all types of kernel dressing, the parameter values are optimised over some training set of ensemble-outcome pairs. As with any form of parameter estimation, however, small training sets can lead to non-robust estimates. We demonstrated the effect of this in section 5.3 in the context of K groups kernel dressing. Here, decreasing the number of ensemble-outcome pairs contained in each group lead to continually improving optimised skill in the training set but eventually lead to a decrease in the skill of the forecasts in the test set.

Dynamic kernel dressing requires the optimisation of an additional parameter over simple kernel dressing. Whenever an additional parameter is required to be optimised, the risk of over fitting [38], in which more parameters are fitted than can be justified given the sample size, is increased. For large training sets, this is un-

likely to be a problem since dynamic kernel dressing only requires the simultaneous optimisation of four parameters. In some applications in practice, however, only a small training set may be available. We have shown that dynamic kernel dressing can be expected to perform at least as well as simple kernel dressing when the training set is large. We now perform an experiment in order to compare the effect of different training set sizes on the performance of each method. Once again, we use the ensemble-outcome pairs formed in experiment 5.A. Since more information is contained in larger ensembles, we may expect a small training set to be less of a hindrance than when the ensemble size is small. We therefore perform the experiment twice, once with a small ensemble size of $M = 8$ (in this case, we randomly draw 8 members from each ensemble in experiment 8.A) and again with the larger ensemble size of $M = 64$. For each training set size, we randomly draw N_{tr} ensemble-outcome pairs from the archive of 2048 pairs, optimise the parameter values and apply them to the test set of 2048 pairs. This is repeated 64 times for each training set size to assess the effects of sampling error.

The results for the case in which the ensemble size is 8 are shown in figure 5.17. The solid lines represent the mean over all mean ignorance scores of the forecast densities formed using dynamic (green) and simple (blue) kernel dressing for each training set size. The error bars represent 90 percent bootstrap resampling intervals. The dashed lines show the mean of the optimised mean ignorance scores over the training set for each size. Even when the training set is very small, dynamic kernel dressing provides significantly better forecast densities than simple kernel dressing. This difference is even more significant when the size of the training set is increased. When the training set is small, the optimised ignorance is much lower than the ignorance of the forecast densities in the test set. This is because for small training

sets, parameter values are likely to be found that depend heavily on the nature of the specific ensemble-outcome pairs within them. Larger training sets provide a better representation of the range of ensemble-outcome pairs contained in the test set and thus the ignorance scores are more similar.

The results of the same experiment using an ensemble size of 64 are shown in figure 5.18. This time, for the smallest sized training sets, although the ignorance using dynamic kernel dressing is lower, the bootstrap intervals overlap so there is no significant difference between the methods. For training set sizes of 32 pairs or larger, however, dynamic kernel dressing performs significantly better than simple kernel dressing.

This begs the question of why, for the smallest training set size, dynamic kernel dressing yields a significant improvement over simple kernel dressing when the ensemble is small but not when it is larger. The explanation for this is likely to be that the size of the kernel width has a lower impact on forecast skill when the ensemble is large. This is because the shape of the distribution is affected more strongly by the ensemble members than by the kernel width. Dynamic kernel is thus likely to be most valuable when the ensemble size is small.

5.6 Perfect Forecast Scenario

In probabilistic forecasting, even when the model dynamics are identical to those of the system, unless the initial condition is known precisely, in which case all future states can be predicted with probability one and an ensemble forecast is not required, the ensemble members will not be drawn from the same distribution as the outcome. This is because ensemble formation schemes generally cannot be expected to sample the observational uncertainty at the forecast launch time perfectly. We therefore

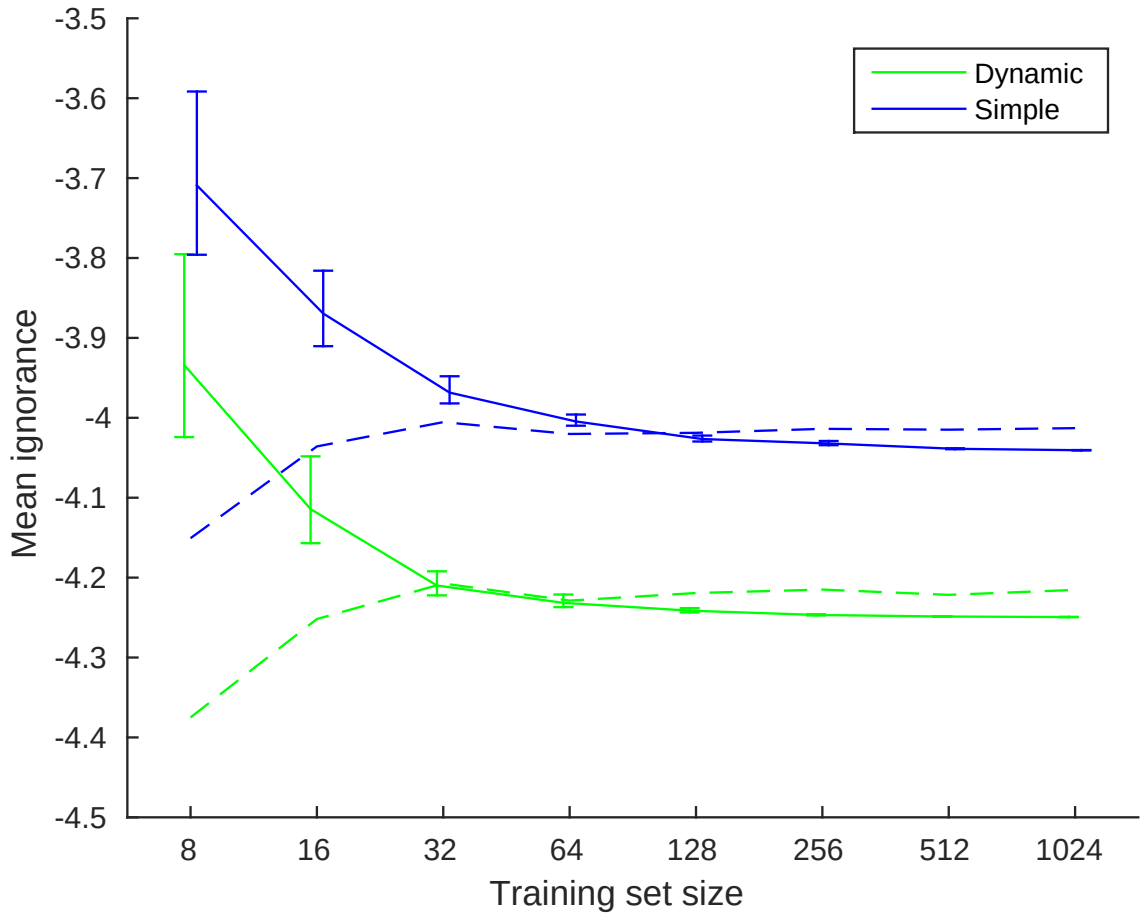

Figure 5.17: The mean ignorance of forecast densities formed from 8 member ensembles using simple (blue solid lines) and dynamic (green solid lines) kernel dressing averaged over 64 repeats for each training set size. The error bars represent 90 percent bootstrap resampling intervals. Using the same colour scheme, the dashed lines show the mean optimised mean ignorance for each training set size. For ensembles of this size, dynamic kernel dressing yields a significant improvement in ignorance even when the training set is very small.

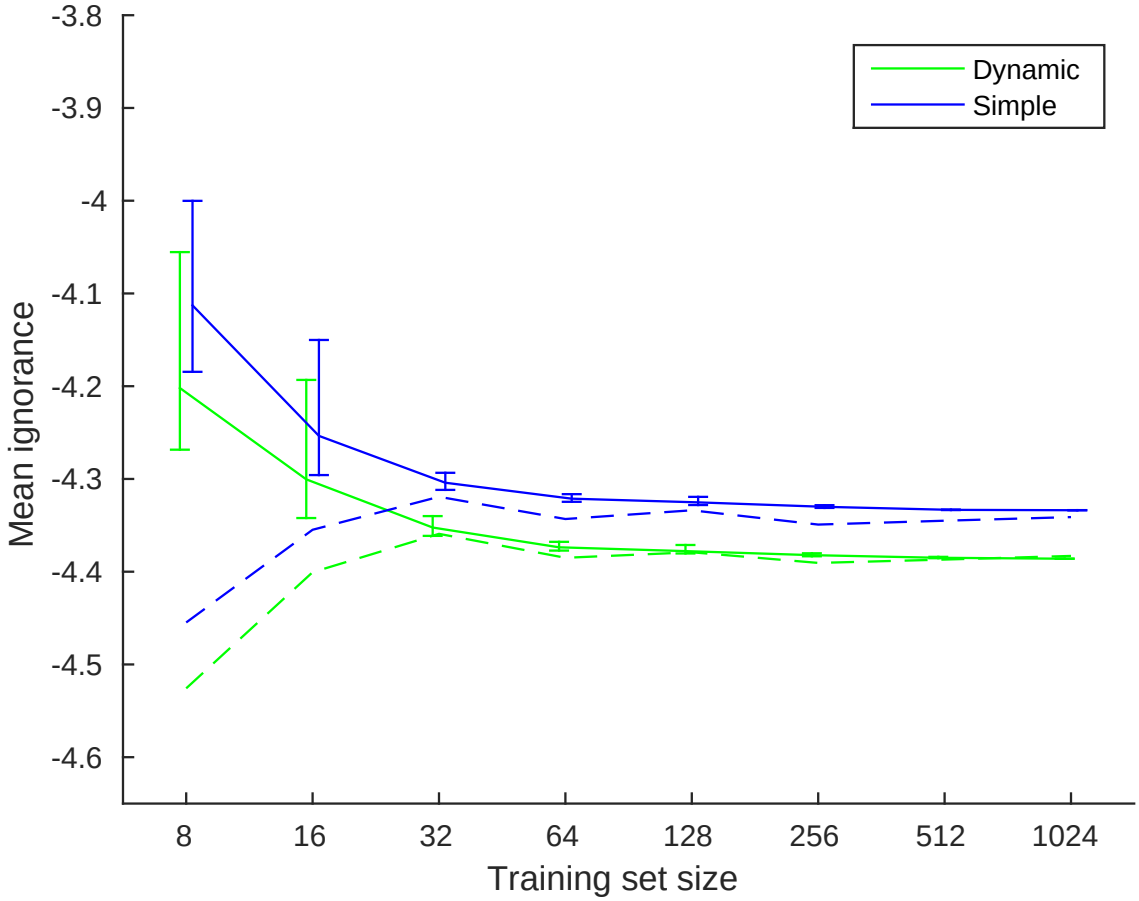

Figure 5.18: The mean ignorance of forecast densities formed from 64 member ensembles using simple kernel dressing (blue solid lines) and dynamic kernel dressing (green solid lines) averaged over 64 repeats for each training set size. The error bars represent 90 percent bootstrap intervals. Using the same colour scheme, the dashed lines show the mean optimised ignorance for each training set size. Dynamic kernel dressing yields a significant improvement in forecast skill for training sets consisting of 32 ensemble-outcome pairs and more. For the smallest training set sizes, there is still some improvement but this difference is not significant for this ensemble size.

make a distinction between the scenario in which the underlying system dynamics are known perfectly from that in which the ensemble is drawn from precisely the same distribution as the outcome. In this section, we focus on the latter case, which we refer to as the perfect forecast scenario (PFS). To apply our experiments in the PFS we must know the system density and thus like the PMS, it can only be constructed.

The PFS, although artificial, is useful. Unlike in the PMS and IMS, we can find the 'optimal' kernel width that minimises some measure of the difference between the forecast density and the system density. As a result, we can compare the performance of each method in terms of how close it gets to recovering this value. In addition to this, we can compare the resulting forecast densities from each method with the system densities. To contrast the model densities constructed using each kernel dressing method with the system densities, a means of comparison is required. Although many such comparison techniques exist [87, 120, 108], a particularly relevant method is the Kullback Leibler divergence (KL divergence) [90] due to its relationship with the ignorance score.

5.6.1 Ignorance and Kullback-Leibler Divergence

The Kullback-Leibler Divergence is a measure of the difference between two distributions. For continuous distributions, it is defined by

$$D_{kl}(p||q) = \int p(x) \log_2(\frac{p(x)}{q(x)}) \tag{5.2}$$

and can be interpreted as the expected information lost in bits when a true distribution p is approximated using some distribution q. For our applications in the PFS, the distribution $p(x)$ is the system density from which both the outcome and

the ensemble are drawn and $q(x)$ is the forecast density constructed using kernel dressing. The relationship between the KL divergence and the ignorance score is outlined below. Applying simple algebra [16] to the formula for the KL divergence yields

$$\begin{aligned} \mathrm{D}_{kl}(p||q) &= \int p(x)\log_2(\frac{p(x)}{q(x)}) \\ &= \int (p(x)\log_2(p(x)) - p(x)\log_2(q(x))) \\ &= \int p(x)\log_2(p(x)) - \int p(x)\log_2(q(x)), \end{aligned}$$

where $\int p(x)\log_2(p(x))$ is the negative of the Shannon entropy [134] of the system density $p(x)$ and $-\int p(x)\log_2(q(x))$ is the expected ignorance using $q(x)$ as a forecast of $p(x)$. The Shannon entropy depends only on $p(x)$ and is thus constant with respect to $q(x)$. Therefore the kernel width that minimises the KL divergence $D_{kl}(p||q)$ will also minimise the expected ignorance. When $q(x)$ and $p(x)$ are identical, the expected ignorance is equal to the Shannon entropy and thus the KL divergence is 0.

5.6.2 The performance of dynamic and simple kernel dressing in the perfect forecast scenario

In this experiment, we make use of the PFS to compare the performance of simple and dynamic kernel dressing in terms of how close each method gets to recovering both the optimal kernel width σ_{KL} and the system density. To create a scenario in which there is a large degree of dispersion in a set of system densities, we make use of the ensembles formed in experiment 5.A. We use them to form the system densities

from which we draw both the ensembles and the outcomes. To do this, as with our climatology, we apply kernel density estimation to each ensemble, finding kernel widths by minimising the mean ignorance using leave one out cross validation. The result is a set of system densities that vary significantly in both shape and dispersion. From each of these densities, we draw both an ensemble and a outcome. In this experiment, both the training set and the test set consist of 2048 ensemble-outcome pairs each drawn from different system densities. We compare the performance of simple and dynamic kernel dressing with a variety of different ensemble sizes using both the ignorance score and the KL divergence between the forecast and system densities.

We present the results of this experiment in two stages. In the first stage, we compare how close each kernel dressing method gets to recovering σ_{KL} for the case in which each ensemble consists of 256 members. In the second stage, we compare the performance of each method using a number of different ensemble sizes.

In figure 5.19, for each ensemble, the kernel width fitted using simple (blue crosses) and dynamic kernel dressing (green crosses) and the optimal kernel width σ_{KL} (red points) are plotted against the ensemble standard deviation. Here, it is clear that there is strong correlation between the standard deviation of an ensemble and its optimal kernel width. This is not surprising since, as we discussed in section 5.2, more smoothing is required for more dispersed ensembles. Of course, since simple kernel dressing takes no account of the dispersion of an ensemble, the kernel width is constant. The kernel widths obtained from dynamic kernel dressing, on the other hand, much better reflect the relationship between the standard deviation of the ensembles and their optimal kernel widths.

In figure 5.20, the forecast densities formed using simple kernel dressing (blue) and

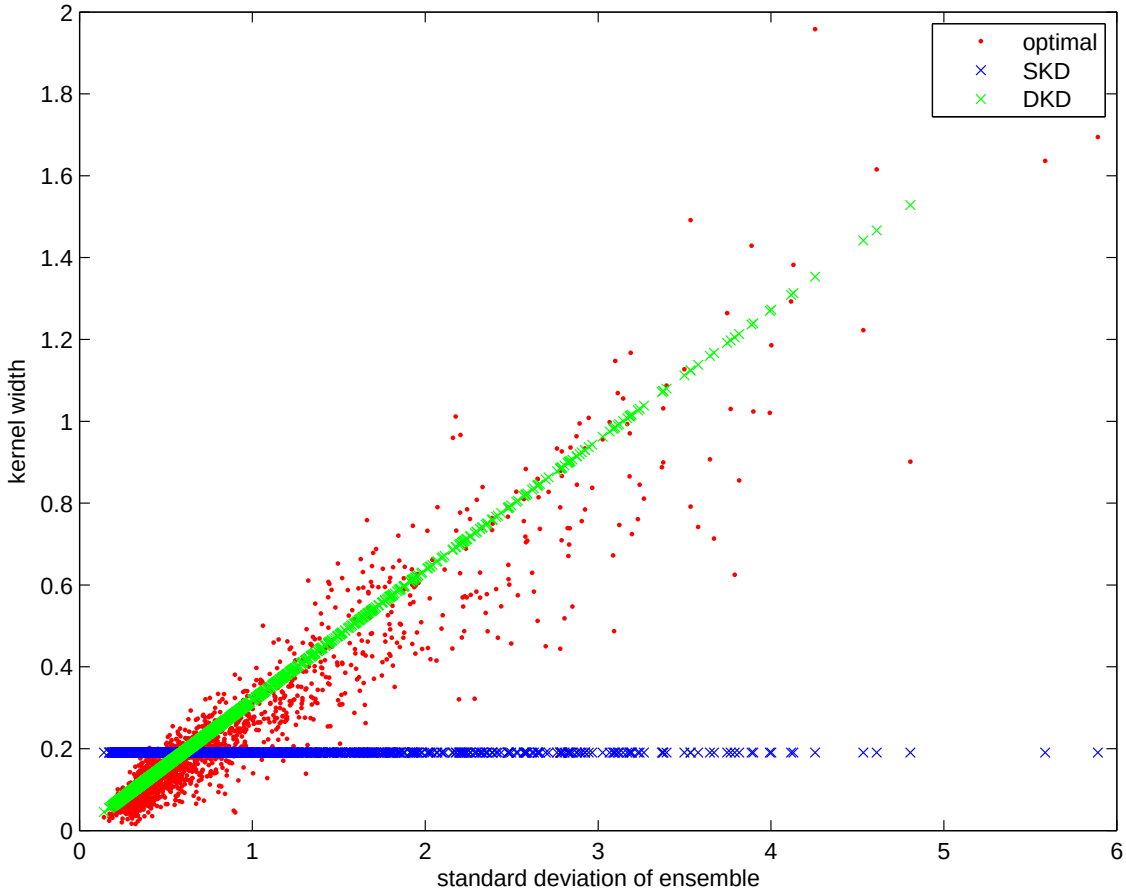

Figure 5.19: The kernel width that optimises the KL divergence (red dots), the kernel width fitted using dynamic kernel dressing (green crosses) and the kernel width fitted using simple kernel dressing (blue crosses) for each ensemble against its standard deviation in the perfect forecast scenario. By taking the ensemble standard deviation into account, dynamic kernel dressing gets much closer to recovering the optimal kernel widths.

dynamic kernel dressing (green) are shown for the ensembles in the test set with the smallest standard deviation (lower panel) and the largest standard deviation (upper panel). For comparison, the system density (black dashed lines) and the forecast density obtained using the optimal kernel widths (red) are shown. In both cases, the forecast density formed using dynamic kernel dressing is similar to that formed using σ_{KL} and hence appears to be close to optimality in terms of the KL divergence. Those densities formed using simple kernel dressing, however, do not get close to recovering the system density. This is because, for the most highly dispersed ensemble, the kernel width is far too small, resulting in an undersmoothed forecast density, whilst for the ensemble with the lowest dispersion, the ensemble is over smoothed and hence the resulting forecast density is much wider than the system density.

We now compare the performance of dynamic and simple kernel dressing for a variety of different ensemble sizes. The results are shown in figure 5.21. In the top left panel, the mean ignorance of the forecasts formed using simple (blue) and dynamic (red) kernel dressing are shown along with that obtained using the optimal kernel width (green). In the top right panel, the mean relative ignorance between the forecasts formed using dynamic and simple kernel dressing (green) and between the optimal forecasts and those formed using dynamic kernel dressing (red) are shown. The same two plots are shown in the bottom left and right panels but with the ignorance replaced with the KL divergence. In terms of both the mean ignorance and the mean KL divergence, there is a significant difference between the performance of dynamic and simple kernel dressing. The difference is largest for smaller ensemble sizes providing further evidence to suggest that this is when dynamic kernel dressing is at its most valuable. Notably, there are no significant differences between the

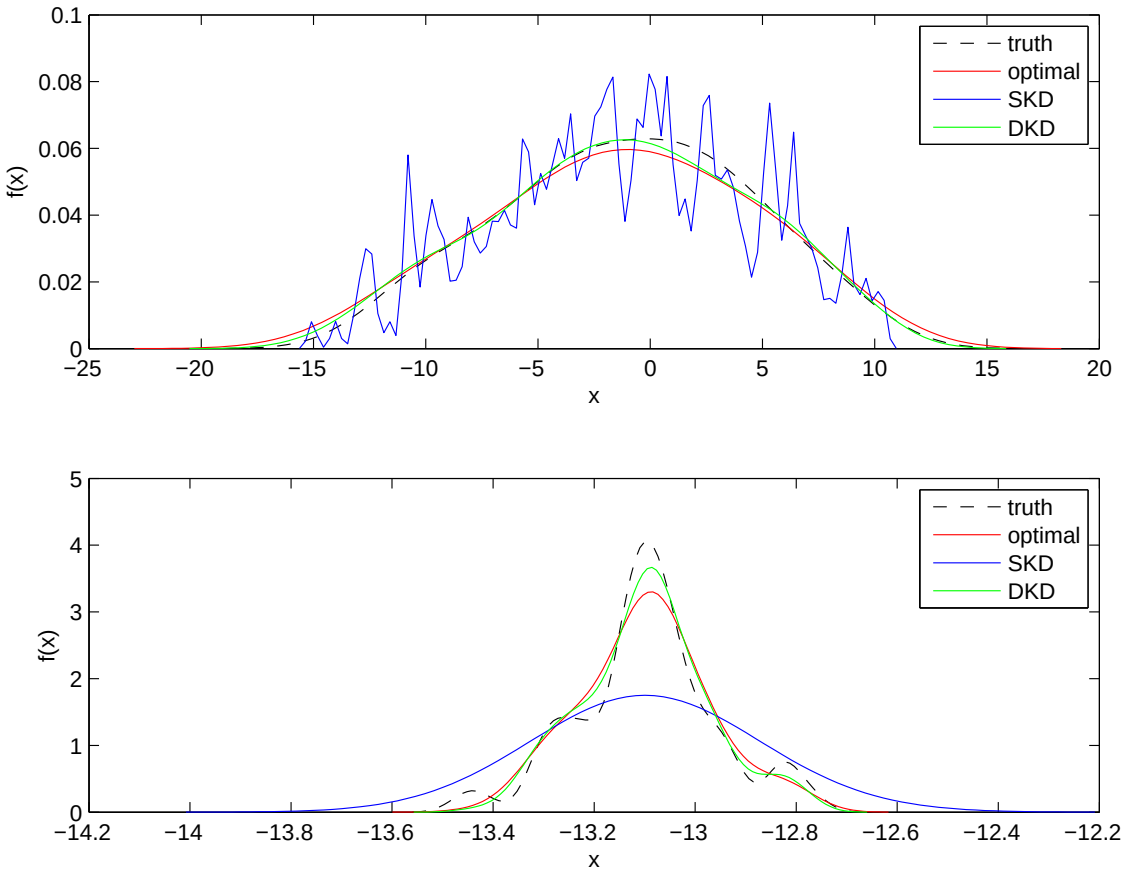

Figure 5.20: Forecast densities formed from the ensembles in the test set with the largest (upper panel) and smallest (lower panel) standard deviation using simple (blue) and dynamic (green) kernel dressing. Also shown is the forecast density obtained by applying the kernel width that minimises the KL divergence (red). The black dashed line represents the system density from which both the ensemble and outcome are drawn. In both cases, dynamic kernel dressing gets much closer to recovering the system density.

performance of dynamic kernel dressing and the forecasts formed using the optimal kernel widths in terms of mean ignorance although a significant difference is found for all but the largest ensemble sizes when comparing the the mean KL divergence. This suggests that, in this case, there is little difference between the forecasts formed using dynamic kernel dressing and the forecasts formed using the optimal kernel widths. Simple kernel dressing performs poorly, on the other hand, yielding much less skillful forecasts than both dynamic kernel dressing and those formed using the optimal widths.

Here, we can see that dynamic kernel dressing can yield forecasts of the same quality as simple kernel dressing using much smaller ensemble sizes. For example, the forecast densities formed using dynamic kernel dressing with ensembles consisting of 8 members perform better than those formed using simple kernel dressing with both 16 and 32 member ensembles. In large scale weather and climate models, the computational cost of running an extra ensemble member can be extremely high. By changing the kernel dressing method, we could find as much improvement in skill as would be achieved by doubling the ensemble size and retaining simple kernel dressing as our method of building forecast densities.

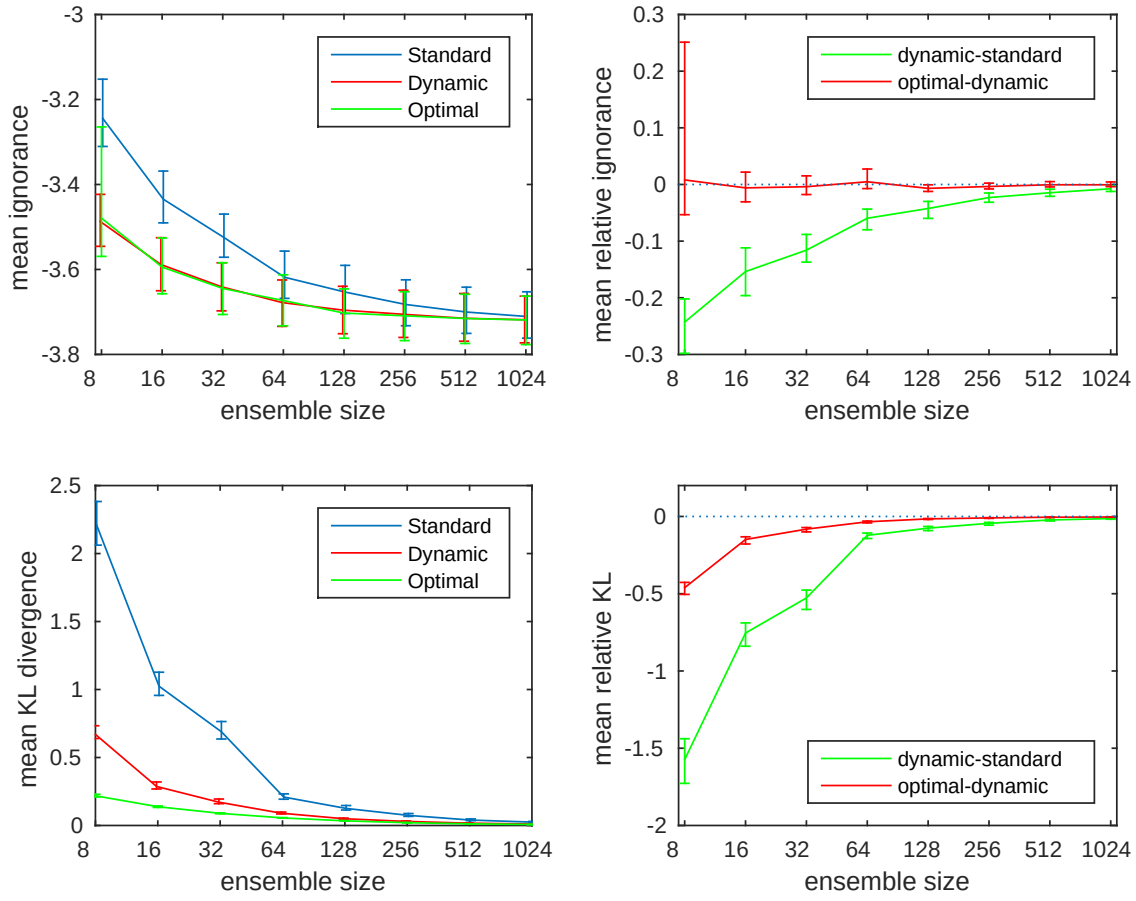

Figure 5.21: Top left:the mean ignorance score of forecast densities formed using simple (blue) and dynamic (red) kernel dressing along with that of the forecasts formed using the optimal kernel width (green) as a function of ensemble size. Top right: The mean relative ignorance between forecasts formed using dynamic and simple kernel dressing (green) and of forecasts formed using the optimal kernel widths and dynamic kernel dressing (red). Bottom left:the mean KL divergence of forecast densities formed using simple (blue) and dynamic (red) kernel dressing along with that of the forecasts formed using the optimal kernel width (green) as a function of ensemble size. Top right: The mean difference in KL divergence between forecasts formed using dynamic and simple kernel dressing (green) and of forecasts formed using the optimal kernel widths and dynamic kernel dressing (red).

Chapter 6

Beyond Bayesian Updating of Forecasts

In operational weather forecasting, new ensembles are commonly launched around every 6 or 12 hours with each simulation extending to include lead times a week or more from their launch time [152, 117]. This means that new ensembles are constantly emerging for each given forecast target time. Generally, newly launched ensembles are expected to contain more information than those launched at earlier times and, as a result, to yield more informative forecast densities. In practice, earlier launched ensembles are often disregarded as new ones becomes available. This seems a curious discontinuity in reliance, and would be fundamentally inconsistent with a Bayesian approach were the forecast densities considered to be probabilities.

Assuming that each forecast density is formed using a single ensemble, in which each member is launched at the same point in time, it seems a sensible approach to favour the most recently launched ensemble for this purpose (assuming that all

other factors, such as ensemble size are kept equal). After all, when an event is closer, there is less time for inaccuracies from initial condition and model error to take hold and so the ensemble can usually be expected to perform better than any of those launched previously. It doesn't necessarily follow, however, that previously launched ensembles cease to be of any value to a forecaster as soon as a new one is launched. For example, suppose we wish to make a forecast of the temperature in exactly 3 days time and so we immediately launch an ensemble. Then, suppose we launch another ensemble exactly 5 minutes later. It would not make sense to discard all of the information in the first ensemble in favour of the second simply because it was launched slightly earlier. Clearly, there are scenarios in which the launch times of 2 ensembles are so close together that, in practice, their performance is indistinguishable. Given an arbitrarily small difference between the launch time of two ensembles, discarding the first ensemble would arguably be throwing away around half of the information available (assuming the ensembles were of equal size). An improvement in forecast skill could almost certainly be achieved by combining the two ensembles into one larger ensemble and using it to construct a forecast density.

In practice, things are less clear cut since forecasts are rarely launched so close together in time. If one ensemble is much more informative than the other, a forecast density formed by combining the ensembles in this way may be *less* informative than one constructed using only the most recently launched ensemble. This is because the importance of the most informative ensemble is reduced to accommodate the less informative ensemble. The longer the time between ensemble launches, the less similar the 2 ensembles will be and thus a threshold value must exist above which this approach is counterproductive. We show that this threshold exists in

section 6.2.2.

We do not suggest the naive approach described above as an efficient way of constructing forecast densities. Nevertheless, we use it to make a very important point: in some circumstances, the most recently launched ensemble can be combined with previously launched ensembles to yield more skillful forecast densities than those constructed solely from the most recently launched. With this encouraging thought, we spend the rest of this chapter in pursuit of more sophisticated techniques aimed at combining forecasts from multiple lead times.

In section 6.2.3, we introduce a more sophisticated approach to the combination of ensembles from multiple lead times called the time-weighted method. The time-weighted method accounts for the different quality of ensembles launched at different times and ensures that the resulting forecast densities are never expected to perform worse, on average, than those formed using only the most recently launched ensemble.

In section 6.3, we take a different approach. Instead of combining the ensembles themselves, we combine forecast densities. We first use a Bayesian updating approach in which the previous forecast density is updated each time a new one becomes available and argue that, whilst Bayesian updating provides an effective approach when a set of forecast densities can be interpreted as probabilities, in practice, the method can fail due to the existence of flaws in the forecasting system, in particular model error.

We then propose a new method called sequential blending which extends the approach of blending described in chapter 2 to the combination of forecast densities formed at different times. This approach is 'safe' when there is no useful informa-

Method	Origin σ
Standard Method	Kernel dressing and blending method introduced in [22]
Pure Bayes Method	Derived [109] from Bayes Theorem [12].
Naive method	Introduced in this thesis
time-weighted method	Introduced in this thesis
Sequential Blending	Introduced in this thesis

Table 6.1: Summary of the origins of methods used in this chapter.

tion contained in previously launched ensembles because the method reduces down to the standard approach of issuing the forecast derived purely from the most recently launched ensemble. We show that sequential blending can be effective both in the idealised perfect model scenario and the more realistic imperfect model scenario.

In table 6.1, we summarise the origins of the methods used in this chapter.

6.1 Combining forecasts

The idea of combining forecasts launched at different points in time is not new. A method called lagged forecasting [69, 31, 70], in which a set of point forecasts launched at different times are combined, much like an ensemble forecast, was introduced by Hoffman and Kalnay in 1983 [69]. Since simulations launched earlier in time are expected to have more forecast error, it was proposed in [42] that each simulation should be weighted differently according to forecast performance. More recently, time-lagged ensembles have been used as inputs for multilinear regression to form probabilistic forecasts of weather variables [101] and to model the dispersion of volcano plumes [154].

The idea of combining forecasts from different *models*, rather than different lead times, has been investigated more thoroughly. In probabilistic forecasting, it is well documented that a weighted average of forecasts obtained from different predictive

models can sometimes yield lower mean squared error, than each of the models individually [11]. The idea that weighted averages of multiple models can outperform the individual forecasts has lead to a large body of research investigating ways to choose the weights placed on each model. A common and well known approach to this is Bayesian model averaging [84, 94] in which the weights are placed on each of the models according to the posterior probability that the data were generated from each one given some observed or forecasted information. Another common approach is to weight models by to their performance according to information criteria [155] which score models by favouring those with high likelihood but penalise each additional parameter to attempt to avoid over fitting [6, 133]. Other studies have found using multiple regression to choose the weights can be successful [122, 35] whilst even placing equal weighting [35, 32] on each candidate can yield effective multimodel forecasts.

In addition to the studies aforementioned, the weighting of forecast models is something we have discussed already in section 2.5.8 of this thesis. When we perform kernel dressing, forecast densities derived from ensembles are blended with the climatological distribution and the resulting densities can often be found to outperform both [22]. When blending is performed, the weighting parameters are found by optimising the forecast skill over a training set. Although, so far in this thesis, we have only considered blending a model density with the climatology, it is just as easily extended to the blending of forecast densities obtained from separate models.

In fact, the idea of combining forecasts from different lead times is not dissimilar to that of combining forecasts from different models. In multi-model forecasting, although one of the models will outperform the others when used individually, combining the 'best' model with one or more others can yield an improvement in perfor-

mance[1] [11]. Similarly, the most recently launched forecast is expected to perform better than any of those previously launched. This, however, is not to say that better forecasts cannot be formed by combining the most recently launched with one or more forecasts launched earlier.

6.2 Forecast densities from multiple lead time ensembles

In this section, we introduce the concept of multiple lead time ensembles and suggest ways in which they can be used to build forecast densities. We define a multiple lead time ensemble to be an ensemble that consists of members launched at more than one lead time. Multiple lead time ensembles are formally defined below.

6.2.1 Multiple lead time ensembles

Let $\boldsymbol{x}_i = x_{i,1}, ..., x_{i,n_i}$ denote an ensemble launched at a time t_i and evolved to a target time τ using the model. Similarly let $\boldsymbol{x}_j = x_{j,1}, ..., x_{j,n_j}$ denote another ensemble launched at a time $t_j > t_i$ also evolved forward to τ. Let $h = t_j - t_i$ represent the difference between the launch time of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The ensembles $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be combined to form a new ensemble $\boldsymbol{x}_c = x_{c,1}, ..., x_{c,n_i+n_j}$. We call $\boldsymbol{x}_c$ a *multiple lead time ensemble.*

[1] Although this is often shown using the mean squared error, which is a questionable metric for nonlinear systems, the fact that blending with the climatology can improve the ignorance score suggests that this can be true even when using more suitable measures of performance.

6.2.2 Naive method of building forecast densities from multiple lead time ensembles

A simple approach to building forecast densities from multiple lead time ensembles is to treat each ensemble member equally and apply simple kernel dressing in precisely the same way as is done with single lead time ensembles. Formally, a forecast density is formed using

$$p_n(x) = \alpha(\frac{1}{n\sigma}\sum_{i=1}^{n} K\Big(\frac{x - x_i - \upsilon}{\sigma}\Big)) + (1-\alpha)p_{clim}(x), \tag{6.1}$$

where x_i represents the i_{th} member of $\boldsymbol{x}_c$ and α, υ and σ are parameters to be found by optimising the mean ignorance score over a training set. We call this approach the *naive method* and that of applying simple kernel dressing to the most recently launched ensemble as the *standard method*. We argue that, for sufficiently small differences in the lead times of members of the multiple lead time ensemble, applying the naive approach will almost certainly yield improved skill over the standard method.

To demonstrate the time scales over which our claim is accurate, we make forecasts with a perfect model of the Lorenz '63 system using the naive method with multiple lead time ensembles consisting of 32 members launched 96 hours ahead and 32 members launched $96 + h$ hours ahead. This is repeated for different values of h so that we can evaluate how the time between launches affects forecast skill. For comparison, we also form forecasts using the standard method in which simple kernel dressing is applied only to those ensemble members launched 96 hours ahead. Since the forecasts are formed using a perfect model, the offset parameter υ is set to zero. The rest of the parameters are optimised over a training set consisting of

2048 ensemble-outcome pairs and the performance is measured over a test set of 8192 pairs. Details of the experiment, which we label experiment 6.A, are listed in table B.5.

In figure 6.1, the mean ignorance score of forecast densities formed using the naive method, expressed relative to that of the standard method, are represented with green stars for each value of h. The error bars represent 95 percent resampling intervals of the mean relative ignorance. The mean relative ignorance achieved by setting $h = 0$ (i.e. doubling the ensemble size at the shortest lead time) is represented with the dotted line.

For those values of h less than or equal to 54 hours, forecast densities formed using the naive method perform significantly better, on average, than those formed using the standard method. For values of h greater than 78 hours and more, however, the naive method performs significantly worse and is thus counterproductive. We have thus demonstrated that some threshold value of h exists above which the naive approach ceases to be effective. When h exceeds this threshold, one of two things must be true. Either no extra forecast skill can be gained from using members from the earlier launched ensemble or the methodology does not use the information efficiently enough to improve forecast skill.

The naive method, of course, is somewhat simplistic. All members from the multiple lead time ensemble are treated equally regardless of launch time and thus no adjustment is made to account for any differences in performance. Since, in simple kernel dressing, each ensemble member is weighted equally, increasing the size of the ensemble reduces the overall impact of each individual member. This means that moving from the standard method, in which kernel dressing is applied to single lead time ensembles, to the naive method, in which forecast densities are formed from

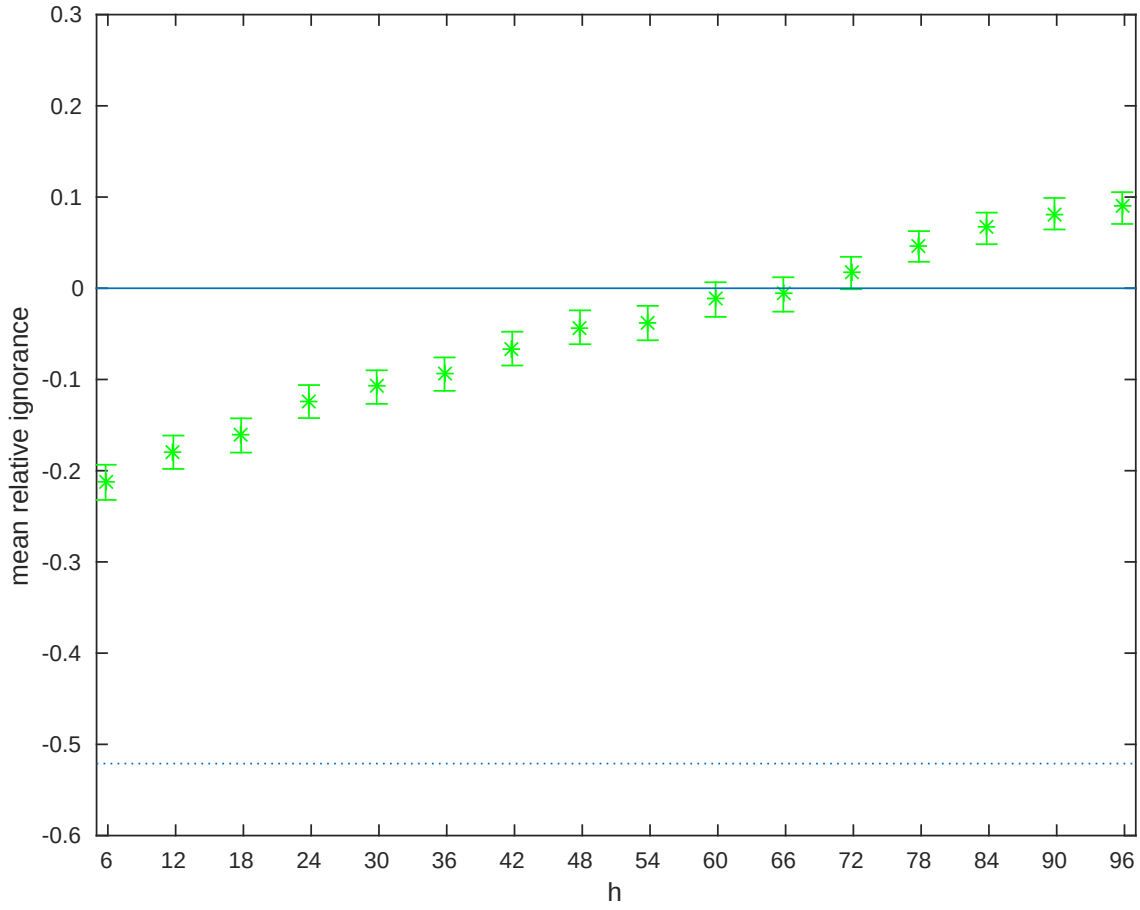

Figure 6.1: The mean ignorance score, expressed relative to that of the standard method, of 8192 forecast densities formed using the naive method (green stars) on multiple lead time ensembles where 32 ensemble members were launched at a lead time of 96 hours and another 32 at $96 + h$ hours for different values of h. The error bars represent 95 percent resampling intervals of the mean relative ignorance. The dotted line represents the mean relative ignorance achieved by setting $h = 0$ (i.e. doubling the ensemble size 96 hours ahead.) Since zero does not lie within the resampling intervals, when h takes a value of 54 hours or less, there is significant evidence that the naive method of combining forecast densities yields improved skill over the standard method. For larger values of h, there is no significant evidence of an improvement in skill whilst, for values of h of 78 hours and longer, there is significant evidence that the forecast densities perform worse on average than the standard method.

larger multiple lead time ensembles, has the effect of reducing the weighting on the most recently launched ensemble members so that equal weighting is placed on earlier launched members which may be significantly less informative. This seems like a suboptimal way of using multiple lead time ensembles and for this reason, we do not advocate the use of the naive method. The results presented above, however, do illustrate an important point: the most recently launched forecast information can sometimes be combined with previously launched forecast information to yield improved forecast skill.

In section 6.2.3, we propose a more sophisticated method of forming forecast densities from multiple lead time ensembles in which the relative performance of ensemble members launched at each lead time is taken into consideration.

6.2.3 A time-weighted approach to building forecast densities from multiple lead time ensembles

In the naive method outlined above, we argued and demonstrated that, whilst for small enough values of h, more skillful forecast densities than those obtained from the standard method can usually be found, the opposite is true when h exceeds some threshold value. This is because the naive method treats each ensemble member equally regardless of its launch time. This means that no account is taken of the fact that those launched nearest to the forecasted event can usually be expected to be the most informative. The performance of the naive method is therefore a trade off between the added benefit of increasing the number of ensemble members and the effect of lowering the weighting on the better performing members so that less well performing members can be included. Whilst the former dominates for small values of h, the latter dominates when h is large.

We now introduce a new approach, which we call the *time-weighted method*, that takes a more considered approach to the construction of forecast densities from multiple lead time ensembles. Forecast densities formed using the time-weighted method take the form

$$p_w(x) = \alpha(\sum_{i=1}^{K} w_i p(x|\boldsymbol{x}_i, \upsilon_i, \sigma_i)) + (1-\alpha)p_{clim}(x) \tag{6.2}$$

where $\boldsymbol{x}_i = x_{i,1}, .., x_{i,n_i}$ is a vector of ensemble members launched at the i_{th} lead time, $p(x|\boldsymbol{x}_i, \sigma_i) = \frac{1}{n\sigma_i}\sum_{j=1}^{n_i} K(\frac{x-x_{i,j}-\upsilon_i}{\sigma_i})$ and $\sum_{i=1}^{K} w_i = 1$. Like earlier in the thesis, we use a Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}u^2}$. The parameters α, $\{w_1, .., w_K\}$, $\{\upsilon_1, .., \upsilon_K\}$ and $\{\sigma_1, .., \sigma_K\}$ are found simultaneously by minimising the empirical skill over the training set (which must be sufficiently large). Given the likely existence of multiple minima, care should be taken in the choice of minimisation algorithm. For the results shown in this thesis, we have used simulated annealing due to its efficiency in finding global minima.

The time-weighted method allows different weightings and kernel widths to be placed on ensemble members launched at different times. This, if required, allows more weighting to be placed on the most recently launched ensemble members whilst potentially still making use of the entire multiple lead time ensemble. Moreover, if ensemble members from one or more launch times have no value to add to the forecast, we expect a weighting close to zero to be placed upon them. If the only lead time of value to the forecast is the most recent, the optimal weighting on this lead time is $w_K = 1$ and the method reduces down to the standard method. Therefore, forecast densities formed using the time-weighted method are expected to perform at least as well, on average, as those formed using the standard method as long as

the parameter estimates are robust[2]. It should be noted, however, that this does not imply that individual forecast densities formed using this method will always perform better.

We apply the time-weighted method to the ensembles formed in experiment 6.A defined in section 6.2.2. As with the naive method, since the model is perfect, we set the offset parameters $\{\upsilon_1, .., \upsilon_K\}$ to zero. The results are shown in figure 6.2. Here, the red stars represent the mean ignorance relative to the standard method of forecast densities formed using the time-weighted method whilst, as in figure 6.1, the green stars represent the mean ignorance achieved using the naive method. Again, in each case the error bars represent 95 percent resampling intervals of the mean and the dotted line represents the case where $h = 0$ and thus where the ensemble size at the shortest lead time is doubled.

In this example, the time-weighted method performs better than the naive method and at least as well as the standard method for all values of h. This is easily explained by considering how the forecasts are formed. For larger values of h, the earlier launched ensemble may be of some value but not enough for it to be given equal weighting. Training the parameters allows us to estimate the most effective weights in terms of maximising forecast skill.

6.3 A Bayesian approach to the combination of forecast densities

In weather forecasting, forecasting centres launch new ensembles at regular intervals of typically 6 or 12 hours [152, 117]. This means that for any given target time, a

[2]Cross-validation techniques can be used to determine whether this is the case

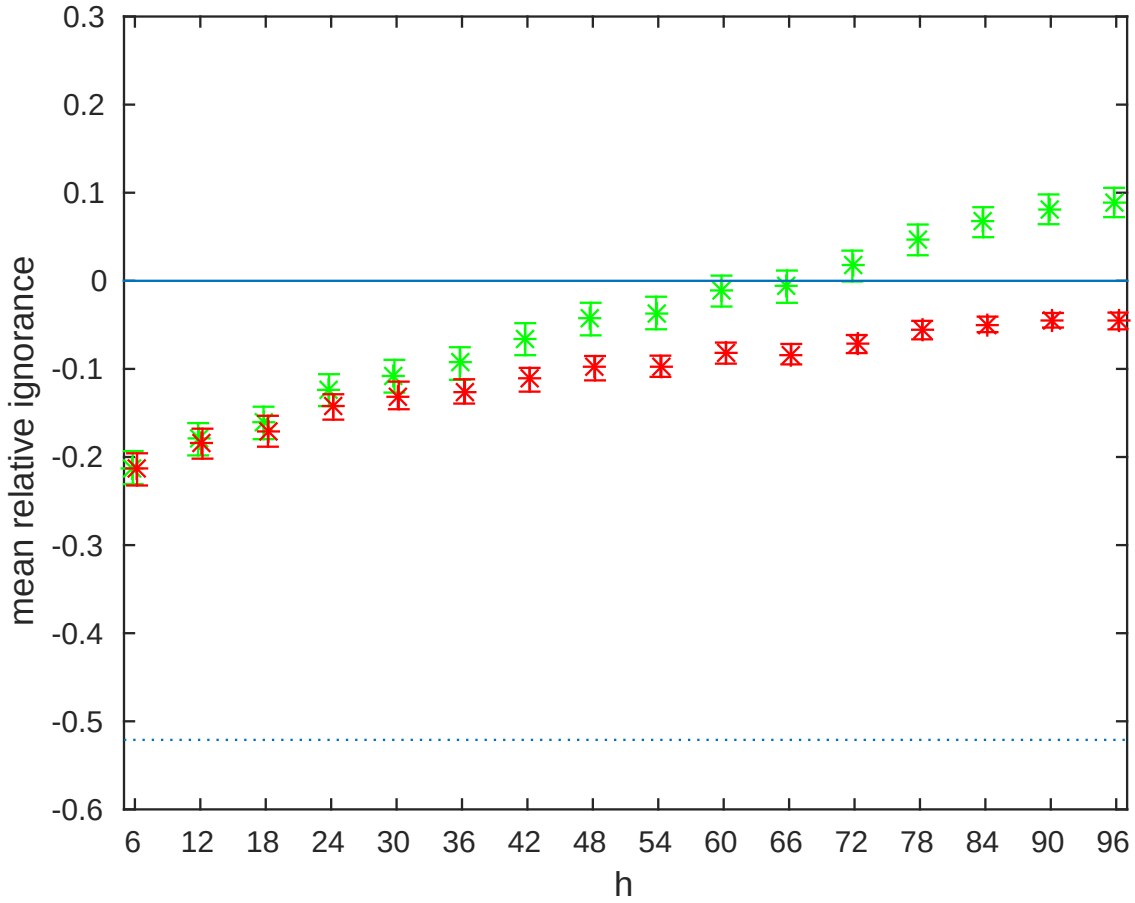

Figure 6.2: The same as figure 6.1 with the results of using the time-weighted method added (red stars). Applying this approach results in skillful forecast densities from larger values of h. Unlike those formed using the naive method, forecast densities formed using the time-weighted method are never expected to perform worse than those formed using the standard method.

steady stream of new forecast information is available. If forecast densities can be treated as probabilities, a Bayesian approach, in which the forecasts are updated each time a new ensemble becomes available, is expected, on average, to make optimal use of the information [162] in terms of maximising the likelihood (and thus minimising the ignorance). We show that, whilst this approach can be effective in the perfect model scenario, when the model is imperfect, the resulting posterior densities can perform worse, on average, than the original forecast densities used in the updating procedure. This calls into question the use of a Bayesian approach in this setting and therefore the rationale of treating forecast densities formed from imperfect models as probabilities. To help describe the Bayesian approach, we first define Bayes theorem and give some background on the concepts of Bayesian inference.

6.3.1 Bayes' Theorem

In probability theory, Bayes' theorem is an important result concerning the probability of a hypothesis given some knowledge of another related event. Let the events $A_1, ..., A_k$ form a partition of the sample space S. The probability of the event A_i given another observed event B is given by the formula

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{p(B)}, \tag{6.3}$$

where $p(B|A_i)$ is the probability of B given A_i and $p(A_i)$ and $p(B)$ are the unconditional probabilities of A_i and B respectively. The probability $p(B)$ is found using the law of total probability such that $p(B) = \sum_{i=1}^{k} p(B|A_i)p(A_i)$. Bayes' theorem was first suggested by the Reverend Thomas Bayes in his essay of 1763 [12] whilst the modern formulation described above was first described by Pierre-Simon

Laplace in 1812 [92]. In the 20th century, the wide variety of potential applications of Bayes theorem were realised and the foundations of Bayesian inference developed [76]. Today, a large proportion of statistics research is Bayesian in nature.

6.3.2 Bayesian Inference

Bayes theorem naturally leads to Bayesian Inference [160] in which a set of prior probabilities are updated with some new information $\mathbf{X}$. This approach is commonly taken when data sets are modelled using parametric distributions. Unlike in the older frequentist approach, in which estimates of parameters are given single values, the parameters in Bayesian inference are represented with probability distributions. The parameters $\boldsymbol{\theta}$ are assigned a distribution reflecting some prior belief which is then updated when new data become available using the relation

$$p(\boldsymbol{\theta}|\mathbf{X}, \alpha_h) = \frac{p(\mathbf{X}|\boldsymbol{\theta}, \alpha_h)p(\boldsymbol{\theta})}{p(\mathbf{X}, \alpha_h)}, \tag{6.4}$$

where $\boldsymbol{X}$ is a vector of data points, $p(\boldsymbol{\theta})$ is the distribution of the parameters before any data have been observed which is usually termed the *prior* and $p(\mathbf{X}|\boldsymbol{\theta})$ is the probability of the data conditional on the parameters, often termed the likelihood. The marginal likelihood $p(\mathbf{X})$ is the distribution of $\mathbf{X}$ marginalised over the parameters, that is $p(\mathbf{X}, \alpha_h) = \int p(\mathbf{X}|\boldsymbol{\theta}, \alpha_h)p(\boldsymbol{\theta})$. α_h is known as a hyperparameter and represents the parameters of the prior distribution.

6.3.3 Bayesian Updating with Probability Density Functions

Bayesian updating provides a principled approach to updating a prior PDF with some relevant new information. Commonly, the new information takes the form of

one or more data points. Bayesian updating can, however, be performed in cases in which the new information takes the form of a probability density function. Define a variable of interest x, a prior PDF $p_{pr}(x)$ and a probability density function $f_{new}(x)$, independent of $p_{pr}(x)$, with which to update the prior. Applying Bayes Theorem directly yields the relation

$$p(x|f_{new}) = kp(f_{new}|x)p_{pr}(x), \tag{6.5}$$

where $p(x|f_{new}(x))$ is the posterior PDF, $p(f_{new}(x)|x)$ is the likelihood function and k is a normalising constant. In practice, $p(f_{new}(x)|x)$ is difficult to find. It is shown by Peter A. Morris [109], in the context of combining expert judgements, however, that, given two important assumptions, which we will define shortly, the following simplification

$$p(x|f_{new}(x)) = cf_c(x)p(x) \tag{6.6}$$

can be made where $f_c(x) = C(x).f_{new}(x)$. The function $C(x)$ is called the *calibration function* which, based on a *calibration set* of past PDFs $f_1(x), .., f_{N_c}(x)$ and corresponding outcomes $y_1, .., y_{N_c}$ comparable[3] to $f_{new}(x)$, is used to calibrate the PDF $f_{new}(x)$. To define the calibration function, we first need to define the *performance function*. Define the *forecast indicator* of the i_{th} forecast to be

$$\phi_i = F_i(y_i), \tag{6.7}$$

where $F_i(x) = \int_{t=-\infty}^{x} f_i(t)dt$ is the CDF of the forecast density evaluated at x. The *performance function* Φ is defined as the frequency distribution of $\boldsymbol{\phi} = \phi_1, ..., \phi_{N_c}$ where N_c is the number of past forecasts from which the performance function is

[3]In the context of forecasting, comparable would mean the same forecasting system and lead time.

estimated. The calibration function is then

$$C(x) = \Phi(F_{new}(x)). \tag{6.8}$$

If the forecasts and outcomes in the calibration set are perfect probabilistic forecasts of the outcomes, the set $\boldsymbol{\phi}$ is randomly drawn from a uniform distribution $U(0, 1)$. The two assumptions [109] of this approach are listed below:

1. No information can be gained about the probability of an outcome falling in the tail of $f_{new}(x)$ by knowing the outcome itself.

2. The dispersion of $f_{new}(x)$ alone does not provide any information regarding the value of the outcome x_0.

We draw upon this approach to Bayesian updating in the next section.

6.3.4 Pure Bayes method

We now formally describe a Bayesian approach to the combination of sequentially launched forecast densities of the same target time. We call this approach the *Pure Bayes method*. Let $p_1(x), .., p_K(x)$ be a sequence of independent forecast densities of the same target time but launched at decreasing lead times such that $p_1(x)$ is formed from the ensemble launched first and $p_K(x)$ the ensemble launched last. Define $p_i^c = C(x).p_i(x)$ to be the i_{th} calibrated forecast density where $C(x)$ is the calibration function, that is the frequency distribution $\boldsymbol{\phi}$ of forecast indicators over the forecast-outcome pairs in the training set corresponding to the same lead time. Since $C(x)$ is a frequency distribution, it must be estimated using some approach to

density estimation. In this thesis, $C(x)$ is estimated using kernel density estimation bounded over the range $(0, 1)$.

Before the first forecast density is formed, no model based information is available and hence the unconditional climatology $p_{clim}(x)$ can be considered to be the best available information. In the Pure Bayes method, the climatology is thus used as the prior distribution. As the first forecast density $p_1^c(x)$ becomes available, it is used to update the climatology to form $p_1^{bayes}(x)$ using the relation

$$p_1^{bayes}(x) = \frac{p_1^c(x)p_{clim}(x)}{\int_{-\infty}^{\infty} p_1^c(x)p_{clim}(y)dy} \tag{6.9}$$

As further new forecasts become available, they are used to update the posterior density from the previous lead time. Hence for $i > 1$, the i_{th} posterior distribution is formed using the formula

$$p_i^{bayes}(x) = \frac{p_i^c(x)p_{i-1}^{bayes}(x)}{\int_{-\infty}^{\infty} p_i^c(x)p_{i-1}^{bayes}(y)dy} \tag{6.10}$$

The result is thus a sequence of Bayesian updated forecast densities dependent on all previous forecast information.

We now define a set of experiments in which we test the performance of this approach in both the perfect and imperfect model scenarios.

6.3.5 Experimental design

The experiments below are conducted in the following way: m member ensembles of a system variable at a fixed time in the future are made at each of K different lead times. As the first ensemble becomes available, it is used to form a forecast density which is then calibrated over the forecast-outcome archive to form $p_1^c(x)$.

The first Bayesian updated forecast $p_i^{bayes}(x)$ is then formed using equation 6.9. Each subsequent new ensemble is then used to form a new forecast density which is then used to update the posterior from the previous step using equation 6.10. To test the effect of applying the Pure Bayes method with different numbers of lead times, we apply the approach K times, starting the updating process at a different starting point each time. We use the ignorance score to compare the performance of the forecast densities formed using the Pure Bayes method with that of the original forecast densities.

6.3.6 Imperfect Model Scenario

First, we test the Pure Bayes method in the IMS using the structurally imperfect model of the Lorenz '63 system defined in section A.2.1 with imperfection parameter $c = 16$. At each of 10 different lead times, forecast densities of the x variable are formed from 64 member inverse noise ensembles using dynamic kernel dressing. The first forecast is launched 8 days ahead of the forecasted event. New forecasts are then launched every day until the target time is one day away after which additional forecasts are launched at 12 and 6 hours ahead. Both the training set and the test set consist of 1024 ensemble-outcome pairs. Details of the experiment, which we label experiment 6.B, are shown in table B.6.

The results of applying the Pure Bayes method are shown in figure 6.3. The mean ignorance scores of the original forecasts, formed using dynamic kernel dressing and blending with climatology, are represented with the black stars for each lead time. The mean ignorance scores of forecast densities formed using the Pure Bayes method are represented with blue lines where the beginning of the line indicates the lead time at which Bayesian updating was first applied. The colour of the points on the

lines indicate whether the mean ignorance of the Pure Bayes method is significantly (at the 5 percent level) worse than (red), better than (green) or not significantly different (yellow) from the mean ignorance of the original forecasts. From these results, it is evident that, in general, in this scenario, the forecast densities formed using the Pure Bayes method perform significantly worse than the original forecast densities. Generally, the earlier the first updating step, the worse the performance of the subsequent forecasts. The exception, however, is at the first updating step, at which the climatology is used as the prior, which tends to yield significantly improved forecast densities. Therefore, it appears that, although in this scenario, the Pure Bayes method tends to be counterproductive, there may be some benefit in issuing posterior forecast densities using the climatology as the prior.

This leads to the question of why the Pure Bayes method does not, on average, yield improved forecast densities. Bayes theorem is a fundamental relationship between a set of probabilities. If the original forecast densities represent probabilities of the system or can be calibrated as such, the Pure Bayes method is expected to make optimal use of the information in terms of maximising the likelihood. The fact that the Pure Bayes method performs worse, on average, than the standard method suggests that the original forecast densities, whilst informative, can not be considered to represent probabilities of the system. We can speculate as to why this may be (though a definitive answer is beyond the scope of this thesis). Since, in the imperfect model scenario, the model dynamics do not perfectly represent the system dynamics, it is unlikely that a forecast density formed in this fashion will provide a good enough representation of the uncertainties in the system for the Pure Bayes method to be effective. It is also unlikely that calibrating the forecast densities will be sufficient to allow them to be considered probabilities of the system. In fact, any

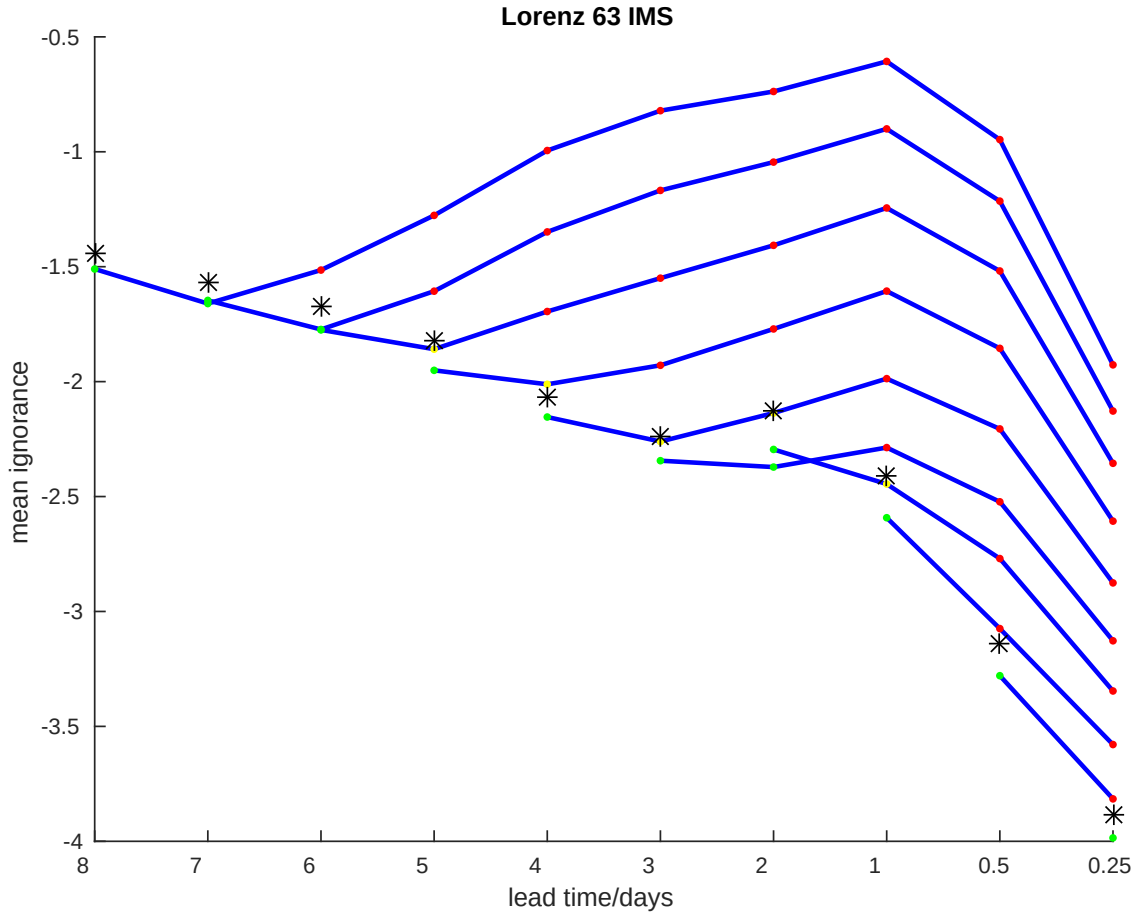

Figure 6.3: The results of applying the Pure Bayes method in the Lorenz '63 imperfect model scenario of experiment 6.B. The black stars show the mean ignorance of the original forecasts whilst the blue lines show the mean ignorance using the Pure Bayes method where the starting point of each line indicates where the updating process was first applied. The colour of the points on the blue lines indicate whether the Pure Bayes method performs significantly worse than (red), better than (green) or not significantly different to (yellow) the original forecasts. Although there appears to be some benefit in using the climatology as a prior, in general, the Pure Bayes method appears to be counterproductive in this scenario.

forecasting system will contain uncertainties regardless of how well the model and system dynamics coincide and thus the assumption of optimality is lost.

6.3.7 Perfect Model Scenario

We now attempt to 'fix' the model such that the forecast densities are more representative of the true uncertainty in the future development of the system. To do this, we replace our imperfect model with a perfect model. If the ensemble members are exchangeable with the outcome, that is drawn from the same distribution, finding a forecast density is simply a case of trying to recover this distribution. In fact, even in the PMS, due to other uncertainties in the forecasting system such as an imperfect ensemble formation scheme, the ensemble and the outcome can *not* generally be considered to be exchangeable. Nonetheless, in the PMS, estimating the distribution of the ensemble is likely to prove effective in forming forecast densities of the outcome. We therefore form forecast densities using kernel density estimation, selecting the kernel width for each ensemble using leave one out cross validation (rather than dynamic kernel dressing and blending as in the IMS).

The results of the experiment in the PMS are shown in figure 6.4. Here, the Pure Bayes method performs much better than in the IMS in improving the forecasts. In most cases, the Pure Bayes method yields significant improvements to the skill of the forecasts. The clearest improvements appear to be at shorter lead times. By the time the forecasted event is 2 days away, the mean ignorance of the forecast densities formed using the Pure Bayes method is better than that of the original forecasts regardless of when the first updating step was performed. At longer lead times, the improvement is less marked and in some cases there is no significant improvement. Contrary to the IMS, using the climatology as a prior appears, on average, to be

counterproductive yielding forecasts with significantly lower skill.

In contrast to the imperfect model scenario, in the perfect model scenario, we find that the Pure Bayes method is effective, yielding posterior forecast densities that outperform the original forecast densities. In this case, the forecast densities appear to be close enough to representing probabilities of the system such that the Pure Bayes method is effective.

In conclusion, it is clear that the Pure Bayes method can not automatically be considered to be an effective method of combining forecast densities. The implication of this is that forecast densities can not always be considered to represent probabilities of the system. In the IMS, we demonstrated that, in reality, the Pure Bayes method can be counterproductive yielding forecast densities that perform worse than the original densities. Since the assumption of optimality fails, the only way to determine the effectiveness of the Pure Bayes method for a particular forecasting system is to to test its performance over past forecast-outcome pairs.

6.4 Sequential blending

In section 6.2.1, we defined multiple lead time ensembles and introduced two approaches to using them to construct forecast densities, the naive and time-weighted methods. We showed that the naive method, though simple and computationally cheap, is only of limited use in practice due to the fact that no account is taken of the relative performance of ensemble members launched at different lead times. We then defined the time-weighted method in which different weightings and kernel widths are placed on ensemble members launched at different times. We showed that this approach can be much more effective than the naive method because it can place smaller weights on less informative ensemble members. Although, in theory,

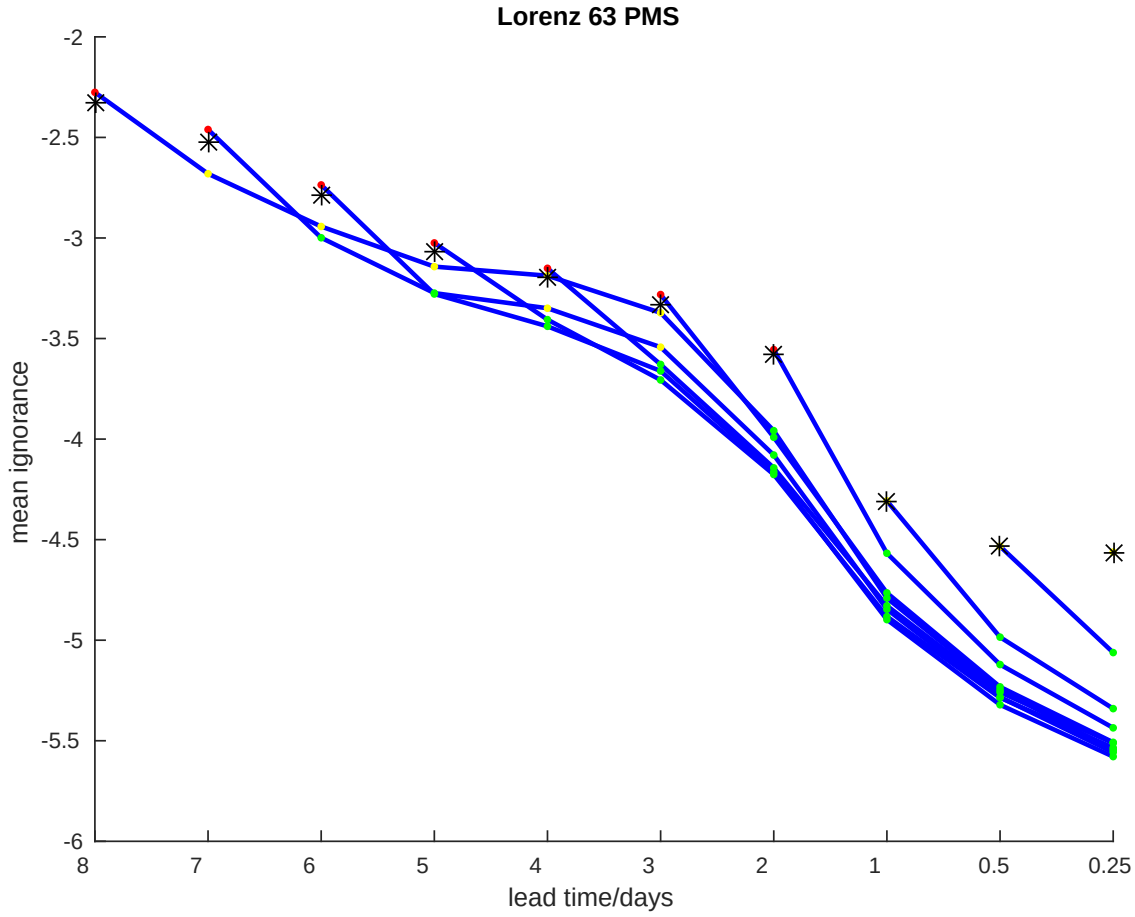

Figure 6.4: The results of applying the Pure Bayes method in the Lorenz '63 perfect model scenario of experiment 6.B. The black stars show the mean ignorance of the original forecasts whilst the blue lines show the mean ignorance using the Pure Bayes method where the starting point of each line indicates where the updating process was first applied. The colour of the points on the blue lines indicate whether the Pure Bayes method performs significantly worse than (red), better than (green) or not significantly different to (yellow) the original forecasts. In this scenario, in many cases, the Pure Bayes method yields significantly improved forecasts whilst using the climatology as a prior, on the other hand, appears to be counterproductive.

this approach can be applied to multiple lead time ensembles consisting of ensemble members from any number of lead times, in practice, the number that can be included is likely to be limited due to the large number of parameters that are required to be optimised simultaneously. For a multiple lead time ensemble consisting of K different lead times, the optimisation of $3K + 1$ parameters is required. This means that including ensemble members from an extra lead time requires the estimation of 3 extra parameters. As a result, the size of the training set required to find robust parameter estimates is likely to grow quickly with the number of lead times considered. Moreover, there is a significant risk of converging to local minima and, although sophisticated optimisation algorithms exist that increase the chance of convergence to the global minimum, this can never be guaranteed. In practice, it is likely that the multiple lead time ensemble can only consist of ensemble members from a relatively small number of different lead times. This constraint means that, although there may be ensemble members that could, in theory, increase forecast skill, they may need to be excluded.

In section 6.3, we defined a Bayesian approach to combining forecast densities from multiple lead times. The sequential nature of this approach means that it has the desirable property that posterior densities can be a combination of forecast information from any number of lead times. Whilst this method can be shown to improve forecast skill in the PMS, in the more realistic IMS, the method is likely to fail yielding posterior forecast densities far worse than the original forecasts used to form them. This means that, unless the forecast model gets very close to reproducing the system dynamics, this approach is unlikely to be of practical use in the real world.

We now define a different approach to the combination of forecast densities launched at different lead times called *sequential blending*. Like the Pure Bayes method,

this method can combine forecast information from any number of lead times. We show, however, that sequential blending can be effective even in the IMS when the Pure Bayes method fails. Moreover, sequential blending requires the simultaneous optimisation of significantly fewer parameters than the time-weighted method.

In sequential blending, the blending approach described in chapter 2, in which forecasts are formed using a weighted average of a model based forecast and the climatology, is extended to the combination of forecasts formed at different lead times. Under this approach, forecasts are continually updated by blending the current forecast with new forecasts as they become available. As before, denote the forecast obtained from the i_{th} available ensemble for a particular target time by $p_i(x)$. When the first ensemble is launched, no previous forecast information exists and so the first sequentially blended forecast is given by

$$p_1^s(x) = p_1(x) \tag{6.11}$$

Subsequent sequentially blended forecasts are then found using the iterative relationship

$$p_i^s(x) = r_i p_i(x) + (1 - r_i) p_{i-1}^s(x), \tag{6.12}$$

where each of the weights $0 < r_i < 1$ are found sequentially by minimising the empirical skill over the training set. Unlike the Pure Bayes method, as long as robust parameter estimates are found, this method is not expected to perform worse, on average, than the original forecast densities because a low weighting can be placed on densities from lead times that add little or no value to the overall forecast skill.

The results of applying sequential blending to the Lorenz '63 perfect model scenario in experiment 6.B are shown in figure 6.5. Here, the blue line represents the mean

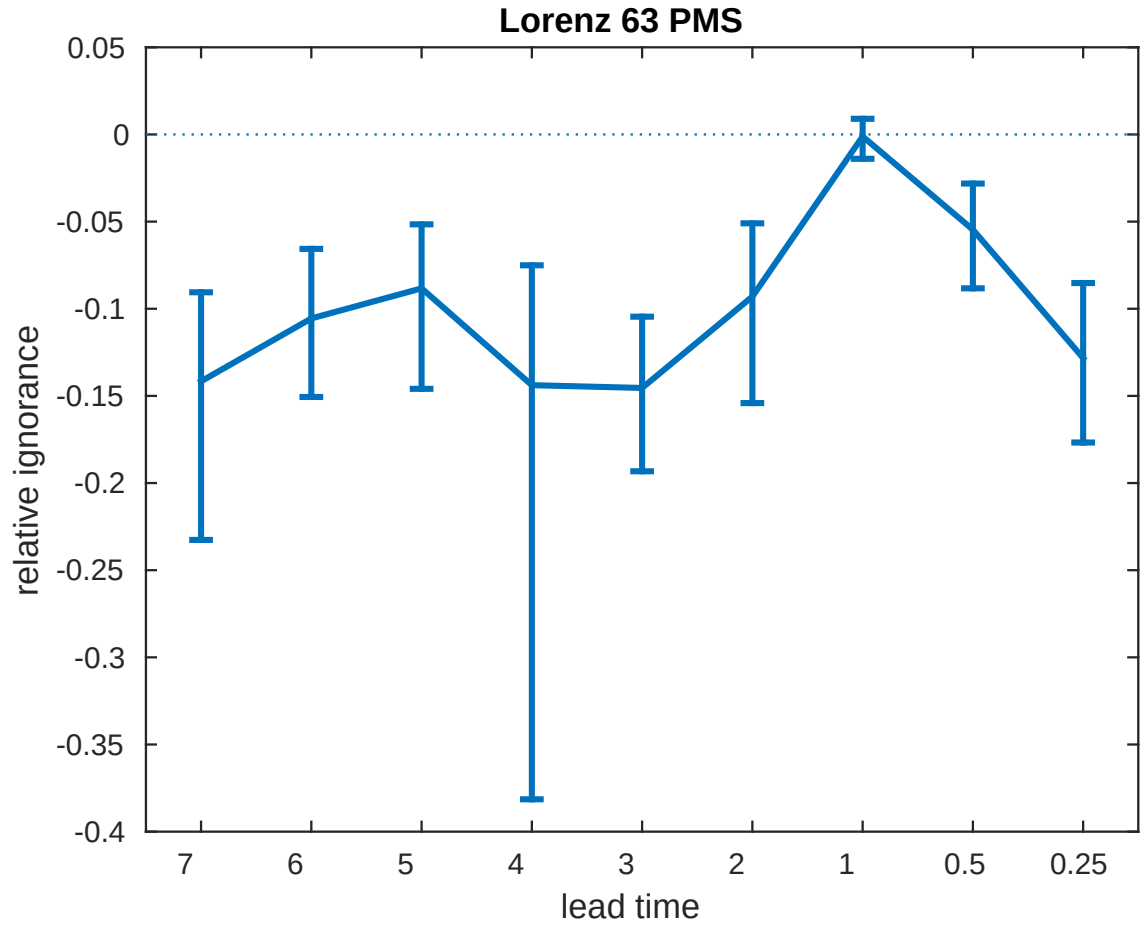

Figure 6.5: Mean ignorance scores of sequential blending expressed relative to that of the standard method with 95 percent bootstrap resampling intervals of the mean in the Lorenz '63 PMS. Since sequential blending yields a lower mean ignorance score and the resampling intervals do not contain zero, there is significant evidence that this approach performs better than the standard method in 8 out of 9 lead times considered.

i	1	2	3	4	5	6	7	8	9
w_i	0.783	0.775	0.936	0.815	0.696	0.865	1.000	1.000	0.844

Table 6.2: Weighting parameters of sequential blending for the Lorenz '63 PMS in experiment 6.B.

ignorance of sequential blending expressed relative to that of the standard method as a function of lead time. The error bars represent 95 percent resampling intervals of the mean. At 8 of the 9 lead times considered, the sequentially blended forecasts perform significantly better than the original forecasts. This demonstrates that sequential blending can successfully improve forecast skill. The weighting parameters are shown in table 6.2.

Whilst this approach is successful in improving the skill of the forecast densities in the PMS, we do not necessarily expect it to outperform the Pure Bayes method. This is because, when the forecast densities can be considered to be probabilities, the Pure Bayes method makes optimal use of the information. Although our model densities almost certainly can *not* be considered to be probabilities even when the model is perfect, the model densities will often be good enough for this approach to be effective. In the IMS, we showed that the Pure Bayes method can perform very poorly. We now show that the sequential blending method can still yield improved forecast skill in this scenario. The results of the experiment using the same imperfect model as in section 6.3 are shown in figure 6.6.

Here, there is significant improvement in forecast skill at all but the shortest lead times considered. This demonstrates that, unlike the Pure Bayes method, sequential blending can still yield improved skill even when the model is imperfect.

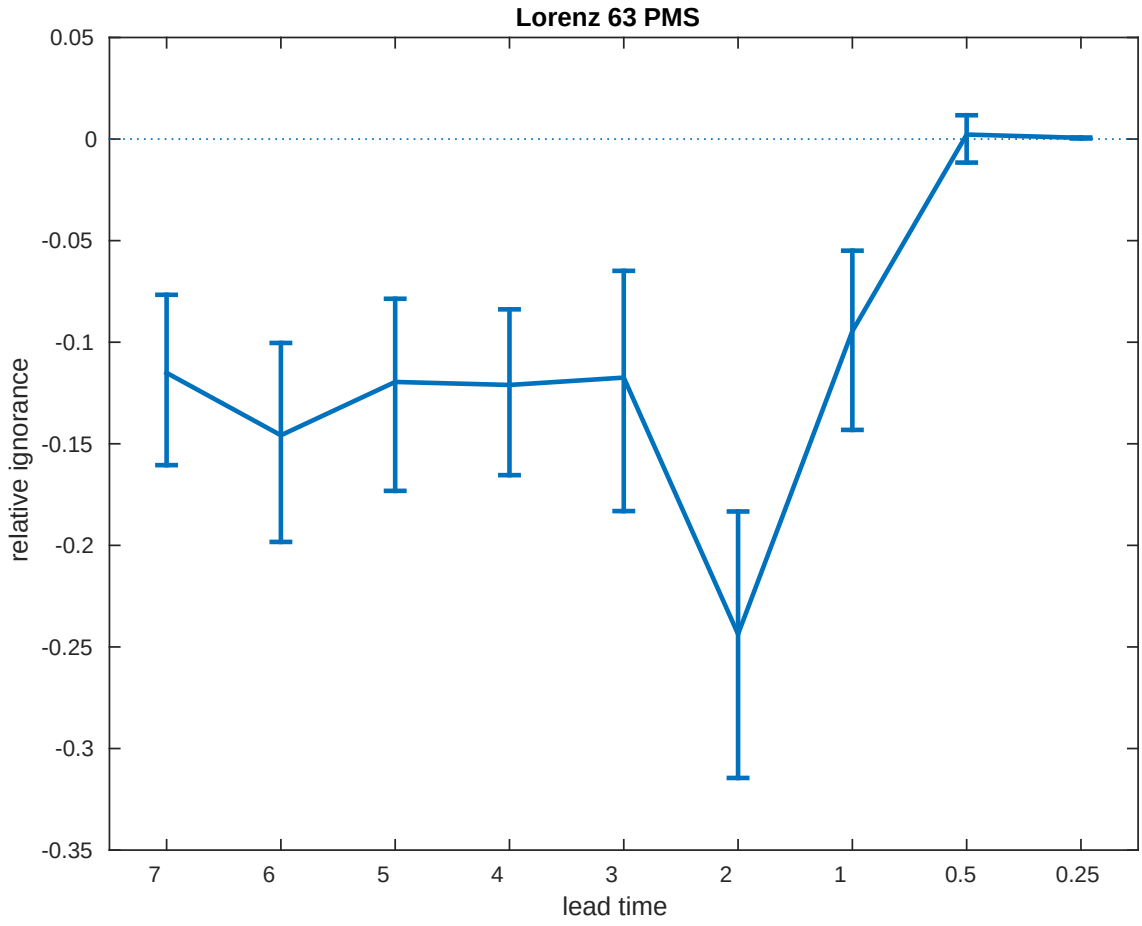

Figure 6.6: Mean ignorance scores of sequential blending expressed relative to that of the standard method with 95 percent bootstrap resampling intervals of the mean in the Lorenz '63 IMS. Since sequential blending yields a lower mean ignorance score and the resampling intervals do not contain zero, there is significant evidence that this approach performs better than the standard method in 7 out of 9 lead times considered.

Chapter 7

Evaluating Data Assimilation with Shadowing Ratios

In this chapter, we discuss the comparison of data assimilation techniques and, in section 7.1, argue that shadowing, specifically the comparison of mean shadowing lengths, provides a useful approach to this kind of comparison. Shadowing ratios, introduced in section 4.1, offer an alternative view of the nature of any difference in the mean shadowing length. We use both approaches in three experiments aiming to make comparisons of data assimilation techniques. In section 7.3, we show that, for a perfect model of the Duffing Map, replacing the Jacobian matrix in PDA with some approximation makes little difference to the performance of the algorithm. In section 7.4, we demonstrate how PDA can fail to find shadowing pseudo-orbits when the climatological standard deviation of each of the variables varies significantly. We demonstrate, using the simple correction suggested in [81], that the resulting forecast trajectories perform much better when this correction is applied. Finally, in section 7.5, we compare the performance of PDA and 4DVAR in an increasingly

nonlinear environment and show that the performance of 4DVAR is much more negatively affected by nonlinearities than PDA.

The new contributions of this chapter are:

1. Comparison of the performance of pseudo-orbit data assimilation (PDA) with and without exact gradient information using shadowing lengths and shadowing ratios.

2. Comparison of the performance of the PDA algorithm with and without using the rescaling approach to PDA suggested in [81] using shadowing lengths and shadowing ratios.

3. Comparison of PDA and 4DVAR in a varyingly nonlinear environment.

7.1 Using Shadowing to Compare Data Assimilation Techniques

In chapter 4, we discussed shadowing and introduced a new method called shadowing ratios as approaches to the comparison of forecasting systems. One particular aspect of the forecasting system that shadowing ratios can be used to assess is the data assimilation scheme. In this section, we perform a number of experiments making comparisons of this type.

Data assimilation has been shown to be invaluable in improving the performance of both deterministic and probabilistic forecasts of nonlinear systems with many different schemes having been proposed over the years [147, 47, 82, 60]. Consequently, there is much discussion surrounding which types of data assimilation schemes give

the best results in each circumstance. Whilst many studies have been performed with the aim of identifying the most effective methods, little research has been undertaken into the best way of comparing such methods.

Generally, there is not much variation in the approach taken to comparing data assimilation schemes. For real life systems, in which the observations provide the only information about the underlying system, comparisons are often made between the point forecasts initialised using the assimilated values and future observed values. This is commonly done by measuring the mean squared error (MSE) between the forecasts and the outcomes at one or more lead times [25, 145, 161]. Another common approach is to compare the distance or the MSE between the analysis and the observations [37] over the assimilation window. Alternatively, techniques are sometimes evaluated by finding the MSE between the analysis and the system states [93, 48]. This, of course, can only be done when the system values are known and hence is only possible for artificial data. These approaches, however, have the potential to be misleading because, as we discussed in chapter 4, MSE can favour imperfect models over perfect models. Shadowing does not have this problem since, all else being equal, a trajectory formed using a perfect model is always expected to shadow longer, on average, than one formed using an imperfect model. Comparing shadowing lengths is thus a simple and intuitive method of comparing data assimilation techniques. Shadowing ratios can be used alongside shadowing lengths to give further information regarding any difference in performance between two different techniques.

7.2 Experimental Design

We now describe the design of the experiments in this chapter. To gain robust results, each assimilation technique is applied to sets of observations deriving from N randomly chosen positions on the system attractor. For each one, a set of observations forming both the assimilation window and the future states are created by adding uncorrelated white noise at a fixed sampling rate. Each assimilation technique is then applied to the observations over the assimilation window and the resulting analyses are used to initialise the model. The shadowing length of each model trajectory is then recorded and shadowing ratios calculated for each technique.

7.3 PDA with estimated gradient information

PDA uses gradient information in the form of the Jacobian matrix to obtain a shadowing pseudo-orbit that approaches a model trajectory with each iteration. For simple low dimensional models, the gradient information is usually easily obtained directly from the model equations. However, for more complex models such as those found in meteorology and climatology this is usually not possible.

It is shown in [82] that the exact Jacobian matrix is not necessarily required to find a shadowing pseudo-orbit in PDA. This is because, whilst the exact gradient information allows us to minimise the cost function $L(x)$ using the path of *steepest* descent, many other paths exist that also provide a route to a shadowing pseudo-orbit. In fact, in some cases, even the identity matrix can be sufficient in achieving this [82]. Although we don't necessarily need the exact gradient information, we might expect that making such an approximation would come at some cost to the quality of the

resulting analysis. In this experiment, we investigate the effect of replacing the exact Jacobian matrix with some approximation. We use the approaches of shadowing and shadowing ratios to compare the performance in each case.

To approximate the Jacobian matrix, we use a simple forward differencing technique. From the first principles of calculus, a derivative $f'(x)$ of a function $f(x)$ is defined by the limit

$$f'(x) = \lim \epsilon \to 0 \frac{f(x+\epsilon) - f(x)}{\epsilon}. \tag{7.1}$$

Often this limit cannot be found analytically and so, instead, an approximation can be made by simply using a small finite value of ϵ. This approach can be extended to partial derivatives and thus can be used to approximate the Jacobian matrix.

To compare the performance of PDA using the exact Jacobian and an approximate one, we perform an experiment in which PDA is applied with a perfect model of the Duffing map defined in appendix A.1.3. The experiment is applied to $N = 1024$ repeats and the noise model is Gaussian at the 5 percent level, trimmed at 3 standard deviations. We apply each technique to a number of different length assimilation windows. Details of the experiment, which we label experiment 7.A, are listed in table B.7.

The results of the experiment are shown in figure 7.1. The top panel shows the mean shadowing length of model trajectories formed using PDA with an exact (blue) and an approximate (red) Jacobian matrix as a function of the assimilation window. The error bars represent 95 percent resampling intervals of the mean in each case. Since the shadowing lengths are so similar, it is difficult to distinguish the lines on the plot. In the middle panel, the mean differences between the shadowing lengths for model trajectories formed using the exact and approximate Jacobian are shown with

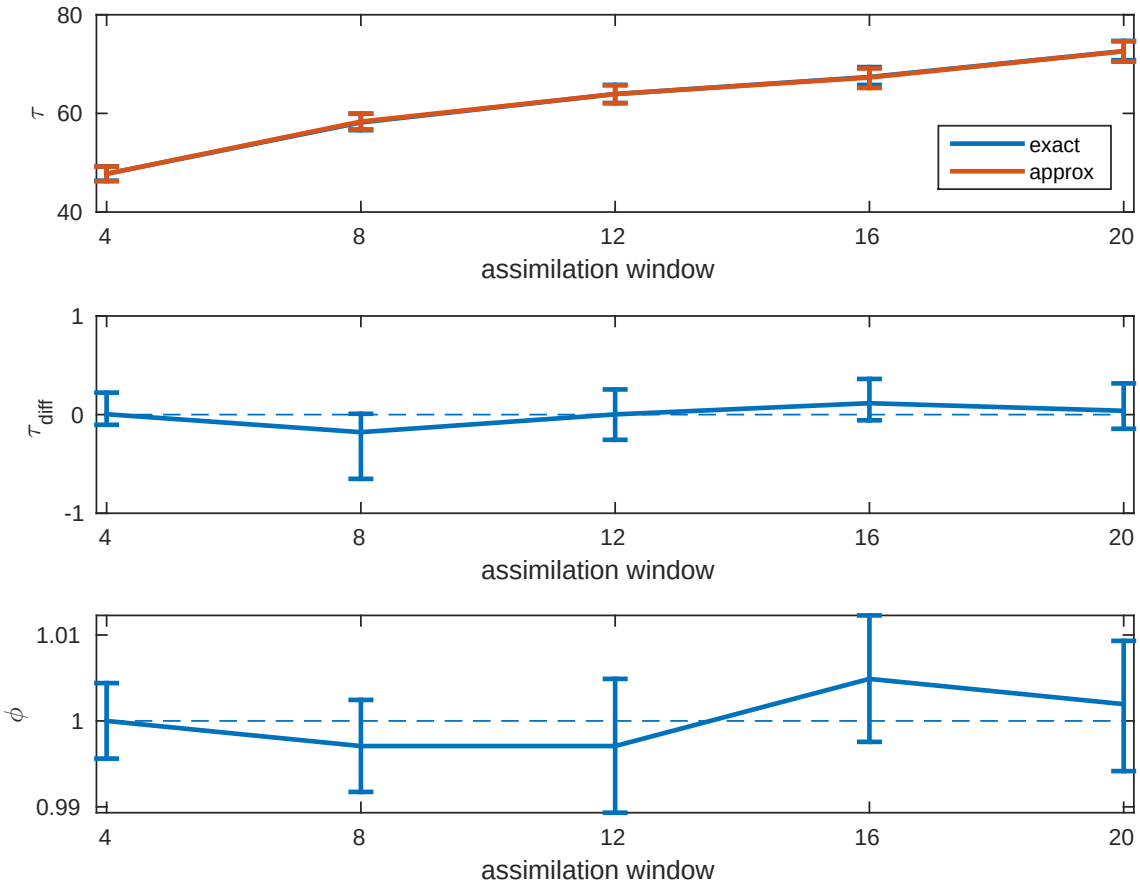

Figure 7.1: Top panel: Mean shadowing length of model trajectories formed using PDA with an exact (blue) and an approximate (red) Jacobian as a function of the assimilation window. The error bars represent 95% resampling intervals of the mean in each case. Since the means are so close, it is difficult to distinguish the lines. Middle panel: The mean pairwise difference between the shadowing lengths of model trajectories formed using PDA with an exact and an approximate Jacobian with 95% resampling intervals of the mean difference. Lower panel: Shadowing ratios of model trajectories formed using an exact and approximate Jacobian. The error bars represent 95% resampling intervals of the shadowing ratio. Since the mean difference between the shadowing lengths is not significantly different from zero and the shadowing ratio is not significantly different from 1, both measures suggest a lack of a difference between the performance of each approach.

95% resampling intervals of the mean difference. Since all of the intervals contain zero, there is little evidence of a difference between the two approaches. In the lower panel, shadowing ratios between model trajectories formed using an exact and an approximate Jacobian are shown. The error bars represent resampling intervals of the shadowing ratio. Since 1 always falls within the interval, there is no significant difference in performance in terms of shadowing ratios.

From these results, it is clear that, in this case, there is no significant difference to the performance of PDA between estimating the Jacobian with a forward differencing technique and using the exact Jacobian. Although in this particular case, there appears to be little difference in performance from using an approximate rather than the exact Jacobian, further analysis is needed to check the effect of this approximation in, for example, the imperfect model scenario and in cases in which observations are taken less frequently. This is beyond the scope of this thesis, however.

7.4 PDA for systems with large variation in the standard deviation of variables

When PDA is applied to a set of observations, the standard deviation of each variable should be comparable. When this is not the case, shadowing pseudo-orbits are not always found. We now demonstrate how the simple adjustment to PDA suggested by Judd in [81][1] can improve the chance of finding a shadowing pseudo-orbit in such cases.

In some dynamical systems, the climatological variability of each of the variables is similar given the units in which they are measured. In others, there is much

[1]Here, the author states that "Ideally, one should either use non-dimensional coordinates, so that all variables are of order one, or a physically and dynamically relevant metric, such as energy"

greater variation. Consider two of the variables commonly featuring in numerical weather prediction models, the atmospheric pressure at sea level and the maximum temperature 2 metres above ground. The former is often measured in millibars and observations tends to lie in between 890mb and 1060mb. The latter is often measured in Celsius and observations rarely fall outside of the range from $-60 \circ C$ to $60 \circ C$. When the variables are measured in these units, the climatological standard deviations are similar. Atmospheric pressure, however, is sometimes expressed in Pascals, which are equivalent to one hundredth of a millibar. On this scale, the standard deviation is much larger. Consider a set of observations of atmospheric pressure $\boldsymbol{x} = x_1, ..., x_n$ measured in Pascals with standard deviation σ. The observations can be converted into millibars by multiplying them by 100. Therefore the standard deviation of the observations when measured in millibars is $\text{std}(100\boldsymbol{x}) = 100\sigma$, a hundred times larger than when they are measured in Pascals. In fact, it is only the measurement scale that affects the relative standard deviation of the variables in a system. It is thus always possible to convert the scales on which each variable is expressed so that the climatological standard deviations are similar.

Consider the 5 dimensional PST system defined in appendix A.2.4. In this system, the climatological standard deviations of the variables vary significantly. These are shown for each variable in table 7.1. Notably, the standard deviation of the Z variable is much smaller than that of the other variables.

We now demonstrate how the difference in climatological variability can cause PDA to fail to find a shadowing pseudo-orbit if it is not adjusted accordingly. We apply PDA to a long assimilation window consisting of 256 observations sampled every 0.1 units of time using a perfect model with the initial conditions defined in table B.8. The results are shown in figure 7.2 in which the black lines represent the system

Variable	Standard Deviation
x	0.1650
y	0.1650
X	0.6006
Y	0.2978
Z	0.0220

Table 7.1: Climatological standard deviations of the the PST system defined in appendix A.2.4

trajectory for each variable, the red dots represent the observations and the green lines the analysis pseudo-orbit obtained from applying PDA directly to the observations. Whilst the analysis remains close to the observations for the x, y, X and Y variables, it does not stay close to observations of the Z variable. Given that the model is perfect and a long assimilation window is used, we would expect to be able to obtain a pseudo-orbit that stays close to the observations of all variables. We now show that shadowing pseudo-orbits can indeed be found by adapting PDA to account for the different levels of dispersion in each variable.

7.4.1 Rescaling the variables

We now show how shadowing pseudo-orbits can be found by applying the simple adjustment suggested in [81]. Since the choice of the scale of each of the variables is entirely arbitrary, the observations in a system can easily be rescaled so that each one has a similar standard deviation. PDA can then be applied to the rescaled observations and the resulting pseudo-orbit transformed back to its original scale. The procedure is performed as follows.

Define a dynamical system described by a set of ordinary differential equations in

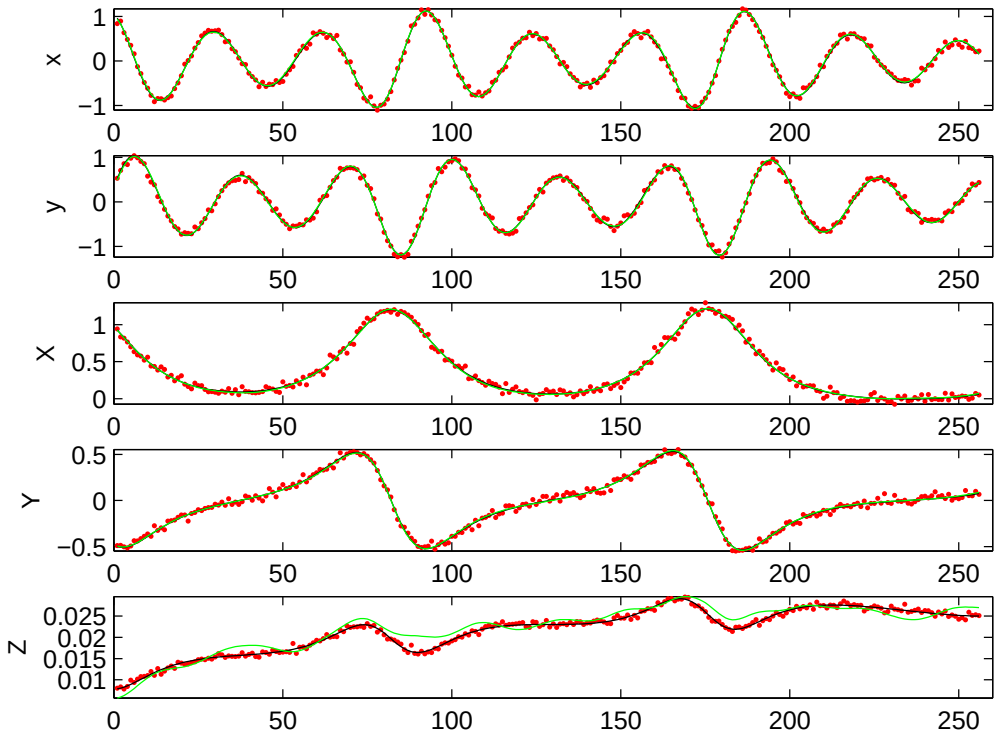

Figure 7.2: The results of assimilating using PDA without rescaling the variables first. For each variable, the black lines are the true trajectory, the red dots the observations and the green lines are the analysis. Although PDA has found a pseudo-orbit that stays close to the observations of the x,y,X and Y variables, this is not the case for the Z variable.

the form

$$\begin{aligned} \frac{dX_1}{dt} &= f_1(x_1, ..., x_n) \\ &\vdots \\ \frac{dX_n}{dt} &= f_n(x_1, ..., x_n). \end{aligned}$$

Let σ_i define the climatological standard deviation of the i_{th} variable in the system. The observations of each variable can be rescaled by multiplying by a scalar w_i. To account for the rescaling of the observations, the model equations must also be rescaled using the simple transformation

$$\begin{aligned} \frac{dZ_1}{dt} &= w_1 \frac{dX_1}{dt} = f_1(\frac{x_1}{w_1}, ..., \frac{x_n}{w_n}) \\ \vdots & \qquad \vdots \\ \frac{dZ_n}{dt} &= w_n \frac{dX_n}{dt} = f_n(\frac{x_1}{w_1}, ..., \frac{x_n}{w_n}), \end{aligned}$$

where w_i is the i_{th} weight which is given by $w_i = \frac{max(\sigma_1,..,\sigma_n)}{\sigma_i}$. PDA is applied to the rescaled observations using the transformed system of equations. The analysis for the true system can then be found by transforming it back to the original scale using the relation $X_i = \frac{Z_i}{w_i}$.

The analysis obtained from the same observations as those in figure 7.2, but adapting PDA using the process described above, is shown in figure 7.3. It is now clear that the analysis shadows the observations of all variables.

In order to test the performance of PDA with and without rescaling, we perform data assimilation 128 times with an assimilation window of 16 observations using the same initial conditions but with different sets of observations in each case and use

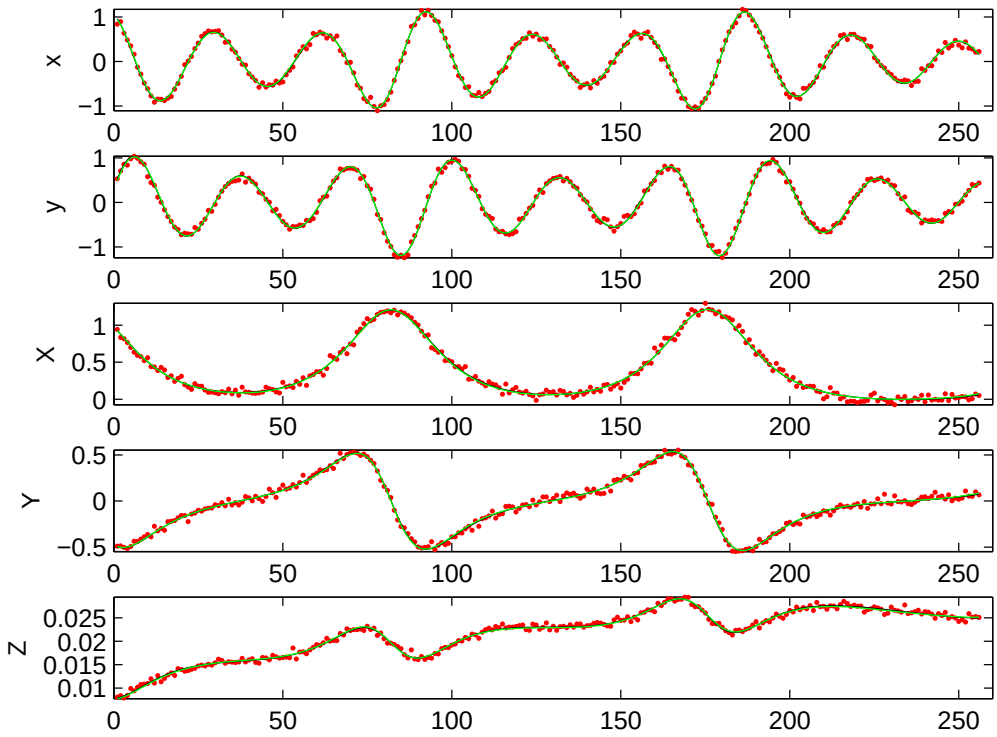

Figure 7.3: The same as figure 7.2 but with the modified version of PDA in which the variables are rescaled. This time, the pseudo-orbit stays close to the observations of all of the variables.

Mean shadowing length with rescaling	24.27
Mean shadowing length without rescaling	20.66
Mean Difference in Shadowing Lengths	3.62 (0.14,6.48)
Shadowing Ratio	1.61 (1.19,2.16)

Table 7.2: Results of applying PDA with and without rescaling the observations first. Values in brackets represent 95 percent resampling intervals.

lengths and shadowing ratios to compare the performance of the resulting forecast trajectories. For simplicity, the noise is trimmed at 3 standard deviations. Details of the experiment, which we label experiment 7.B, are listed in table B.9. Table 7.2 shows the mean shadowing length with and without rescaling, the mean pairwise difference between the two with a 95% resampling interval of the mean difference and the shadowing ratio between the two with a 95% resampling interval of the shadowing ratio. Since zero does not fall within the resampling interval of the mean difference in shadowing length, there is significant evidence that rescaling the observations before applying PDA yields model trajectories that shadow longer than when rescaling is not applied. The shadowing ratio of 1.61 implies that forecast trajectories initialised using the analysis from PDA after rescaling shadow longer 1.61 times more often than those formed without rescaling. Since the resampling interval of the shadowing ratio does not contain one, this result is significant. There is thus significant evidence that, for this model, rescaling before applying PDA provides longer shadowing lengths on average than applying PDA without rescaling.

7.5 The effect of nonlinearity on PDA and 4DVAR

In many practical applications of forecasting, the system dynamics are highly nonlinear. In this example, we test the performance of PDA and 4DVAR in an increasingly nonlinear environment. In this example, we show, using shadowing and shadowing ratios, that the performance of PDA is less affected by higher levels of nonlinearity

(defined below) than 4DVAR providing evidence that PDA may be more suitable in highly nonlinear systems. Below, we describe a practical method of measuring the nonlinearity in a system.

7.5.1 Measure of nonlinearity

To measure the nonlinearity in a system, we apply the test introduced in [54]. This test quantifies the nonlinearity in a system by measuring the average period of time for which a linear approximation of a system is defined to be satisfactory under some criterion. Consider an initial condition of a dynamical system obtained from an analysis $\boldsymbol{A}(0)$. The analysis initial condition will be displaced from the true initial condition $\boldsymbol{x}(0)$ by a fixed perturbation $\boldsymbol{\delta}(0)$ and thus can be expressed in the form $\boldsymbol{A}(0) = \boldsymbol{x}(0) + \boldsymbol{\delta}(0)$. The trajectory of the perturbed initial condition can then be described by a Taylor expansion of $\boldsymbol{x}$

$$\boldsymbol{A}(t) = \boldsymbol{x}(t) + M(\boldsymbol{x}, t)\boldsymbol{\delta}(0)\boldsymbol{\delta}^T + H(\boldsymbol{x}, t)\boldsymbol{\delta}(0) + ..., \tag{7.2}$$

where $M(\boldsymbol{x}, t)$ is the linear propagator, $H(\boldsymbol{x}, t)$ is the Hessian matrix and so on with increasing order partial derivatives. For sufficiently small values of both the initial perturbation $|\boldsymbol{\delta}(0)|$ and the time t, the linear relation

$$\boldsymbol{A}(t) = \boldsymbol{x}(t) + M(\boldsymbol{x}, t)\boldsymbol{\delta}(0) \tag{7.3}$$

gives an approximation of equation 7.2 in any smooth dynamical system. The linear propagator $M(\boldsymbol{x}, \tau)$ can be said to govern the linear evolution of a perturbation from $\boldsymbol{x}$ over the time interval from 0 to τ. The initial perturbation $\boldsymbol{\delta}_0$ can then be said to evolve linearly according to $\boldsymbol{\delta}(\tau) = M(\boldsymbol{x}, t)\boldsymbol{\delta}(0)$. The error arising from

making the linear approximation is thus $\boldsymbol{A}(t) - (\boldsymbol{x}(t) + \boldsymbol{\delta}(0))$. The quality of the linear approximation can therefore be measured as a function of time. The extent to which the linear approximation reflects the dynamics of the system is called the *linear regime*. The length of the linear regime is defined to be the period of time for which the approximation of the linear regime is defined to be acceptable. To calculate the length of the linear regime, the test exploits the fact that in a linear environment, initial perturbations of equal magnitude but opposite signs will also evolve over time equally with opposite signs. The importance of the nonlinearities in the system can thus be measured by the extent to which this is not the case.

Consider the initial condition obtained from the analysis $\boldsymbol{A}(0)$. Denote the twin perturbations from $\boldsymbol{A}(0)$ by $\boldsymbol{\delta}^+(0)$ and $\boldsymbol{\delta}^-(0)$ where $\boldsymbol{\delta}^+(0) = -\boldsymbol{\delta}^-(0)$. If error growth is exactly linear the equality $\boldsymbol{\delta}^+(t) = -\boldsymbol{\delta}^-(t)$ holds for all t. Scaling by the average magnitude of the evolved perturbations leads to a suitable statistic for evaluating the nonlinearity in a system. The *relative nonlinearity* of evolution Θ is given by

$$\Theta(\hat{\delta}, ||\boldsymbol{\delta}||, t) = \frac{||\boldsymbol{\delta}^+(t) + \boldsymbol{\delta}^-(t)||}{||\boldsymbol{\delta}^+(t)||||\boldsymbol{\delta}^i(t)||}. \tag{7.4}$$

The value of Θ corresponds to the proportion of the error resulting from the linear approximation to the system. A truly linear system would have a value of $\Theta = 0$, as long as the analysis initial condition $\boldsymbol{A}(0)$ corresponds to the exact initial condition. A threshold value of Θ can thus be chosen above which the error resulting from the linear approximation is defined to be too high. In this thesis, we follow [54] in defining the length of the linear regime to be the period of time until the relative nonlinearity exceeds 0.2.

Of course, the length of the linear regime can depend on the position on the system

attractor from which the initial condition is drawn. Figure 7.4 shows points on the attractor of the Rössler map coloured according to the length of the acceptable linear regime of a trajectory starting at that point. The darker the colour, the shorter the length of the linear regime and hence the higher the level of nonlinearity.

Since there is significant variation in the colours, it is clear that the nonlinearity is state dependent. Note how the nonlinearity tends to be highest for initial conditions lying in positions where there is substantial uncertainty as to whether the system trajectory will continue to move around the base of the attractor or on to the part of the attractor that moves in the z direction (upwards on the figure).

7.5.2 PDA versus 4DVAR in an increasingly nonlinear scenario

In this experiment, we exploit a property of model 1 of the Lorenz '96 system, defined in section A.2.3, to compare the performance of PDA and 4DVAR in a varyingly nonlinear environment. We use the mean length of the linear regime over trajectories initialised randomly on the attractor to measure the nonlinearity in the system for each parameter value. Increasing the parameter F in the Lorenz '96 system applies more forcing to the system variables resulting in a larger degree of nonlinearity. Comparing the performance of the algorithms for different values of F thus allows us to evaluate the performance for different levels of nonlinearity in the system. We apply PDA and 4DVAR with a perfect model of the 5 dimensional Lorenz '96 system with a fixed assimilation window of 32 time steps and 25 percent Gaussian noise trimmed at 3 standard deviations. For each value of F, we measure the performance over $N = 512$ repeats of the experiment. Details of the experiment, which we label experiment 7.C, are listed in table B.10.

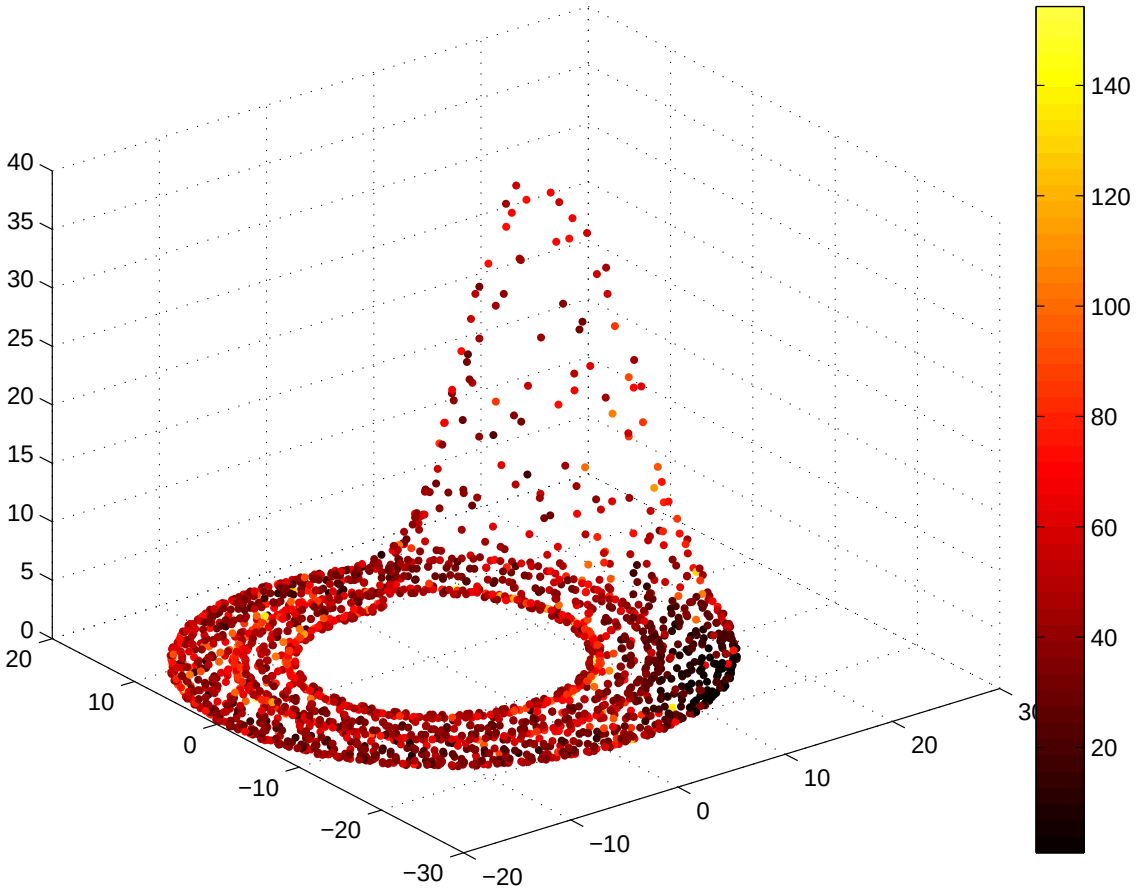

Figure 7.4: Points on the attractor of the Rössler map coloured according to the length of the linear regime of a trajectory initialised at that point. Darker colours imply a shorter linear regime (as indicated on the colour bar) and hence higher nonlinearity. Note how the nonlinearity tends to be highest for trajectories initialised close to where the attractor splits, either continuing around the base of the attractor or up in the z direction. For trajectories initialised on this part of the attractor, the linear regime tends to be less than 100 time units.

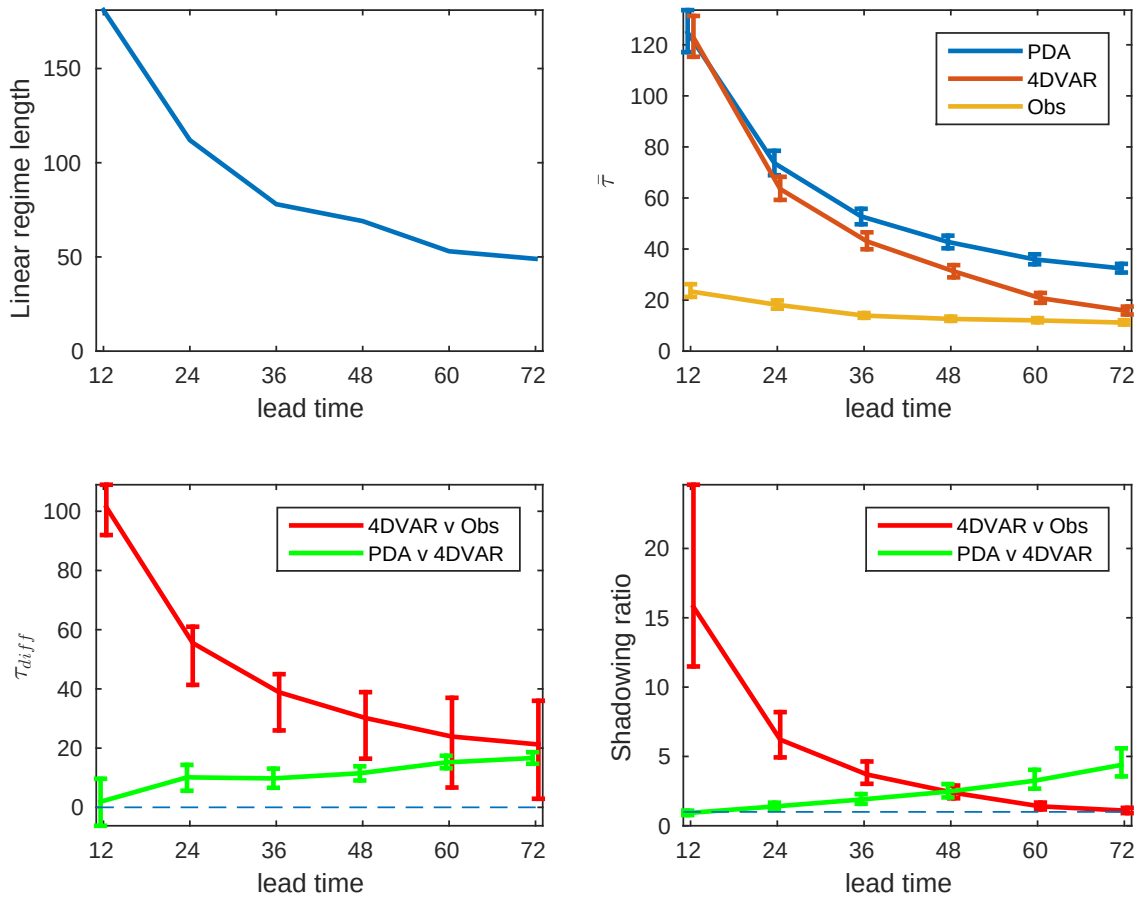

Figure 7.5: Top left panel:the mean length of the linear regime as a function of F. Top right: the mean shadowing lengths achieved using PDA (blue), 4DVAR (red) and when no data assimilation is used (orange) as a function of F. The error bars represent 95 percent resampling intervals of the mean. Bottom left: the mean difference in shadowing lengths between PDA and 4DVAR (green) and between 4DVAR and when no data assimilation is used (red) as a function of F. The error bars represent 95 percent resampling intervals of the mean difference in shadowing length. Botttom right: Shadowing ratios between PDA and 4DVAR (green) and between 4DVAR and the observations (red) as a function of F. The error bars represent 95 percent resampling intervals of the shadowing ratio in each case.

The results of the experiment are shown in figure 7.5. In the top left panel, the mean length of the linear regime is shown as a function of F. Putting more forcing into the system decreases the length of the linear regime and thus the level of nonlinearity is increased. In the top right panel, the mean shadowing lengths achieved using PDA (blue), 4DVAR (red) and when no data assimilation is used (orange), i.e. initialising with the observations, are shown as a function of F. The error bars represent 95 percent resampling intervals of the mean. In the bottom left panel, the mean difference in shadowing length between PDA and 4DVAR (green) and between 4DVAR and using no data assimilation (red) are shown. In each case, the error bars represent 95 percent resampling intervals of the mean difference. In the bottom right panel, shadowing ratios between trajectories formed using PDA and 4DVAR (green) and between those formed using 4DVAR and using no data assimilation (red) are shown. From these results, it is clear that, although both techniques are affected by increasing the nonlinearity in the system, the effect on PDA is much smaller than it is on 4DVAR. Whilst, for $F = 12$, zero lies within the resampling interval for the mean difference in shadowing length between PDA and 4DVAR and there is thus no significant difference, for larger values of F, the difference is significant. The same conclusion can be drawn from the shadowing ratio between trajectories formed using these techniques since 1 lies within the resampling interval for $F = 12$ but not for larger values of F. We can also compare the performance of 4DVAR and using no data assimilation. Looking at both the mean difference in shadowing length and the shadowing ratio, as the value of F is increased, the benefit from using 4DVAR appears to fall quickly suggesting that, if there is enough nonlinearity in the system, 4DVAR may eventually have no benefit in terms of increasing the mean shadowing length. Although more investigation would be invaluable as to the effect of the length of the assimilation window, the effect of model imperfection and

the number of variables etc, these encouraging results suggest that PDA can be a much more effective algorithm than 4DVAR when the underlying system is highly nonlinear.

We can speculate as to why PDA outperforms 4DVAR when the nonlinearity in the system is increased. 4DVAR aims to find the maximum likelihood model trajectory over the assimilation window. In theory, increasing the assimilation window increases the information available and thus the maximum likelihood trajectory is expected, when the model is perfect, to be closer to the system trajectory. In practice, a known limitation of 4DVAR is the existence of local minima. As the assimilation window is increased, the number of local minima and thus the probability of finding one also increases [123]. 4DVAR is thus generally most effective when the assimilation window is short. Whilst maximum likelihood model trajectories would likely provide long shadowing lengths, 4DVAR is unlikely to find them in practice. The large difference in performance between PDA and 4DVAR when the nonlinearity is increased may be caused by an increase in the probability of finding a local minimum far from the maximum likelihood trajectory in 4DVAR. Whilst the PDA cost function also has local minima, it can be shown that these minima tend to provide model trajectories closer to the observations than 4DVAR [39].

Chapter 8

Improving Probabilistic Prediction with Imperfect Models

In this chapter, we demonstrate how some of the methods introduced in previous chapters can be used to improve prediction of dynamical systems with imperfect models.

In section 8.1, we return to the topic of constructing forecast densities from ensembles. In chapter 5, we demonstrated how dynamic kernel dressing can outperform simple kernel dressing in both the perfect model scenario, in which the system dynamics are known perfectly, and the perfect forecast scenario, in which the ensemble is drawn from the same distribution as the outcome. In this chapter we show, using two different system-model pairs, that dynamic kernel dressing can outperform simple kernel dressing when the model is structurally imperfect.

We then make use of the boosted probability approach introduced in section 4.3 to compare the performance of ensembles formed using inverse noise and PDA. In both

cases, we also compare the performance of forecast densities formed using simple and dynamic kernel dressing. We conclude that using both PDA ensembles and dynamic kernel dressing can yield improved boosted probability times.

We then revisit the techniques considered in chapter 6 which aim to improve skill by combining forecasts from multiple lead times. We apply these methods to both a perfect and an imperfect model of the Moore-Spiegel system and demonstrate that similar conclusions can be made.

Finally, we show how shadowing can be used to assess the performance of ensembles and to help identify whether skillful forecast densities can be expected to be formed at a given lead time. We use this approach to demonstrate how shortcomings in a forecast density formation scheme can be identified, in particular the simple approach of fitting Gaussian distributions and estimating the parameters from the ensembles. We then show that dynamic kernel dressing yields forecast densities more consistent with the performance of the ensembles.

The new contributions of this chapter are:

1. Comparison of the performance of simple and dynamic kernel dressing in the imperfect model scenario.

2. The use of boosted probability to compare the performance of forecast densities constructed using inverse noise and PDA ensembles with simple and dynamic kernel dressing.

3. Comparison of the performance of the methods described in chapter 6 in the context of both a perfect and an imperfect model of the Moore-Spiegel system.

4. The use of shadowing to compare the performance of sets of ensembles with the skill of forecast densities.

8.1 Improving forecast densities using dynamic kernel dressing

In chapter 5, we described three kernel dressing methods that take the dispersion of individual ensembles into account when fitting kernel widths. Due to its ability to improve forecast skill, its ease of use and the fact that large training sets are not required, we concluded that out of those described, dynamic kernel dressing is likely to be the most effective method. Whilst we tested this method in both the perfect model scenario and the perfect forecast scenario, we have not yet tested its performance in the more realistic imperfect model scenario. In this section, we show how dynamic kernel dressing can improve forecast performance when the model is imperfect. To do this, we use two different system-model pairs, the first, a model of the simple 2 dimensional Henon map and the second, the higher dimensional Lorenz '96 system.

8.1.1 A low dimensional system - Henon Map

In this experiment, which we refer to as experiment 8.A, we compare the performance of forecast densities of the x variable of the Henon map formed using simple and dynamic kernel dressing. Using the imperfect model defined in appendix A.1.2 with imperfection parameter $c = 24$, inverse noise ensembles consisting of 64 members each are used to form forecast densities at 10 lead times ranging from 20 steps to 2 steps ahead. The training set, over which the parameters are optimised, consists

of 1024 ensemble-outcome pairs whilst the performance of the forecasts is assessed over a test set of the same size. Details of the experiment are listed in table B.11. Similarly to the ensembles formed using the perfect model of the Lorenz '63 system in experiment 5.A, the standard deviation of the ensembles is highly variable. This is illustrated in figure 8.1 in which the standard deviation of each individual ensemble along with the mean standard deviation are shown for each lead time. Again, whilst, on average, the standard deviation of the ensembles increases with lead time, there is significant variation at each one, suggesting that simple kernel dressing is at risk of yielding uninformative forecast densities in some cases.

The results of the experiment are shown in figure 8.2. The blue and green lines represent the mean ignorance of forecast densities at each lead time formed using simple and dynamic kernel dressing respectively. The error bars represent 95 percent bootstrap resampling intervals of the mean ignorance relative to climatology. At most of the lead times considered, the forecast densities formed using dynamic kernel dressing perform better, on average, than those formed using simple kernel dressing. Dynamic kernel dressing appears to be particularly successful in improving forecast skill at long lead times (14 steps and above) where simple kernel dressing yields very little skill. This suggests that dynamic kernel dressing has the potential to increase the maximum lead time in which forecast skill can be found. Note that the resampling intervals give an indication of whether each method performs significantly better than the climatology rather than an indication of their performance relative to each other. We attempt to determine whether dynamic kernel dressing performs significantly better than simple kernel dressing at the end of section 8.1.2.

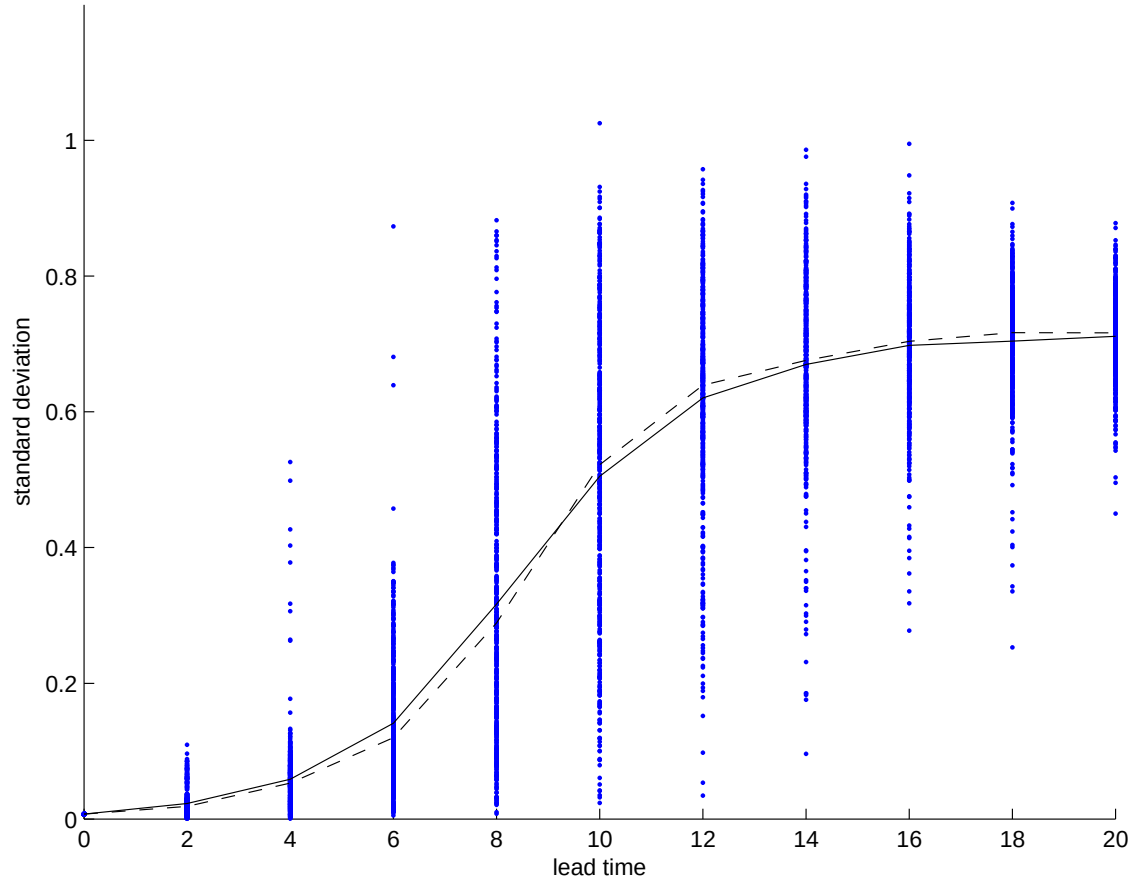

Figure 8.1: The standard deviation (blue points) along with the mean (black solid line) and the median (black dashed line) standard deviation at each lead time of the ensembles in experiment 8.A. Although, on average, the standard deviation of the ensembles increases with lead time, there is significant variation at each one.

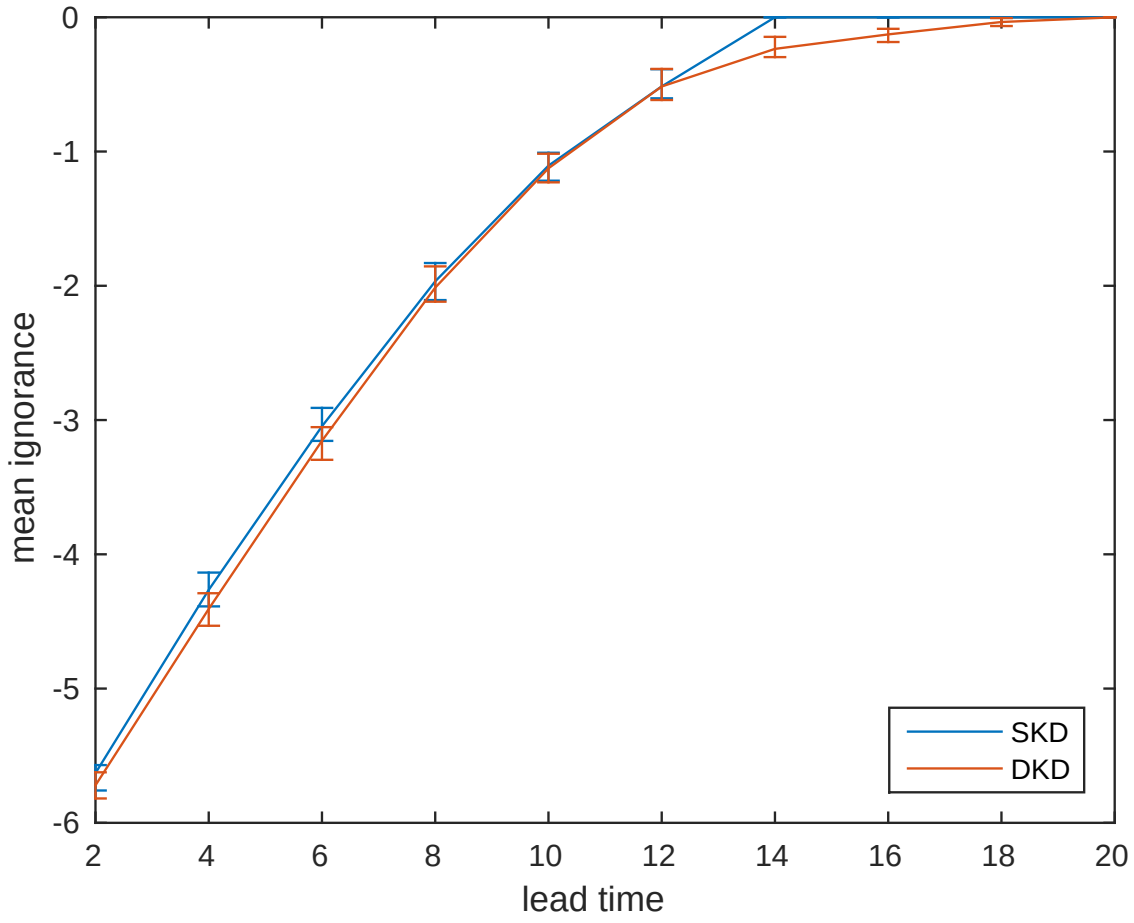

Figure 8.2: The mean ignorance scores of forecast densities formed using simple (blue) and dynamic (green) kernel dressing for the 10 different lead times in experiment 8.A. The error bars represent 95 percent bootstrap resampling intervals of the mean ignorance. At most lead times considered, dynamic kernel dressing yields more skillful forecast densities, on average, than simple kernel dressing.

8.1.2 A higher dimensional experiment: Lorenz '96

In experiment 8.B, we compare the performance of dynamic and simple kernel dressing using an imperfect model of the higher dimensional Lorenz '96 system. We define the system to be model 2 with $m = 4$ and $n = 8$, as described in appendix A.2.3, whilst we use model 1 to generate ensembles at 10 different lead times, 1 day apart from each other. Details of the experiment are listed in table B.12. The standard deviation of each 64 member ensemble as well as the mean standard deviation at each lead time are shown in figure 8.3. As in experiment 8.A, there is significant variation in the dispersion of ensembles at each lead time.

The mean ignorance scores for each method are shown in figure 8.4. Again, the blue and green lines represent the mean ignorance of forecast densities formed using simple and dynamic kernel dressing at each lead time whilst the error bars represent 95 percent bootstrap resampling intervals of the mean ignorance relative to climatology. At shorter lead times, dynamic kernel dressing performs better, on average, than simple kernel dressing whilst, beyond 6 days, the value of the slope parameter b in dynamic kernel dressing is very close to zero and hence the forecast densities formed using each method are very similar. At lead times of 3 and 4 days, dynamic kernel dressing appears particularly successful in improving forecast skill. The performance of the forecasts densities at longer lead times demonstrates how, whilst dynamic kernel dressing can often be successful in improving forecast skill, it easily reduces down to simple kernel dressing when there is no value to be added in this way. The results of this experiment demonstrate how dynamic kernel dressing can be used to improve the skill of forecast densities with an imperfect model of a higher dimensional system.

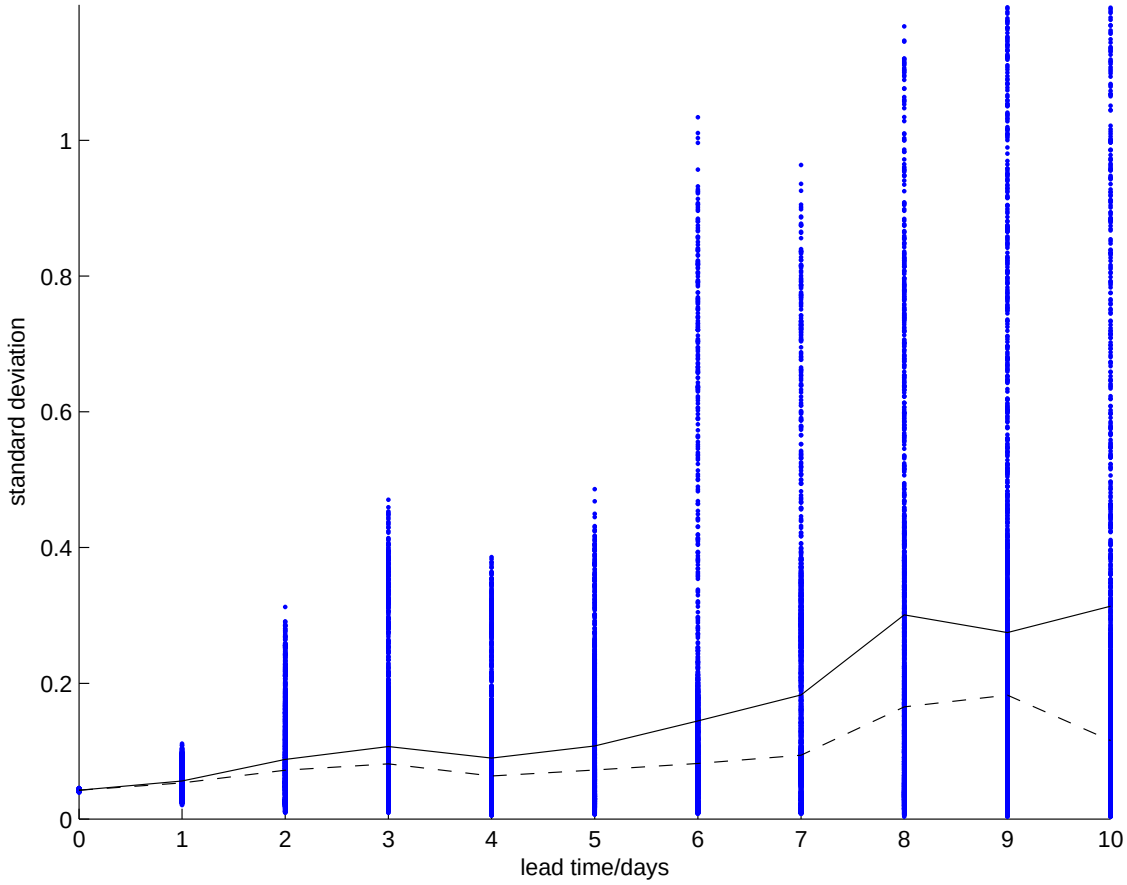

Figure 8.3: The standard deviation (blue points) along with the mean (black solid line) and the median (black dashed line) standard deviation of the ensembles at each lead time in experiment 8.B. Although, on average, the standard deviation of the ensembles increases with lead time, there is significant variation at each one.

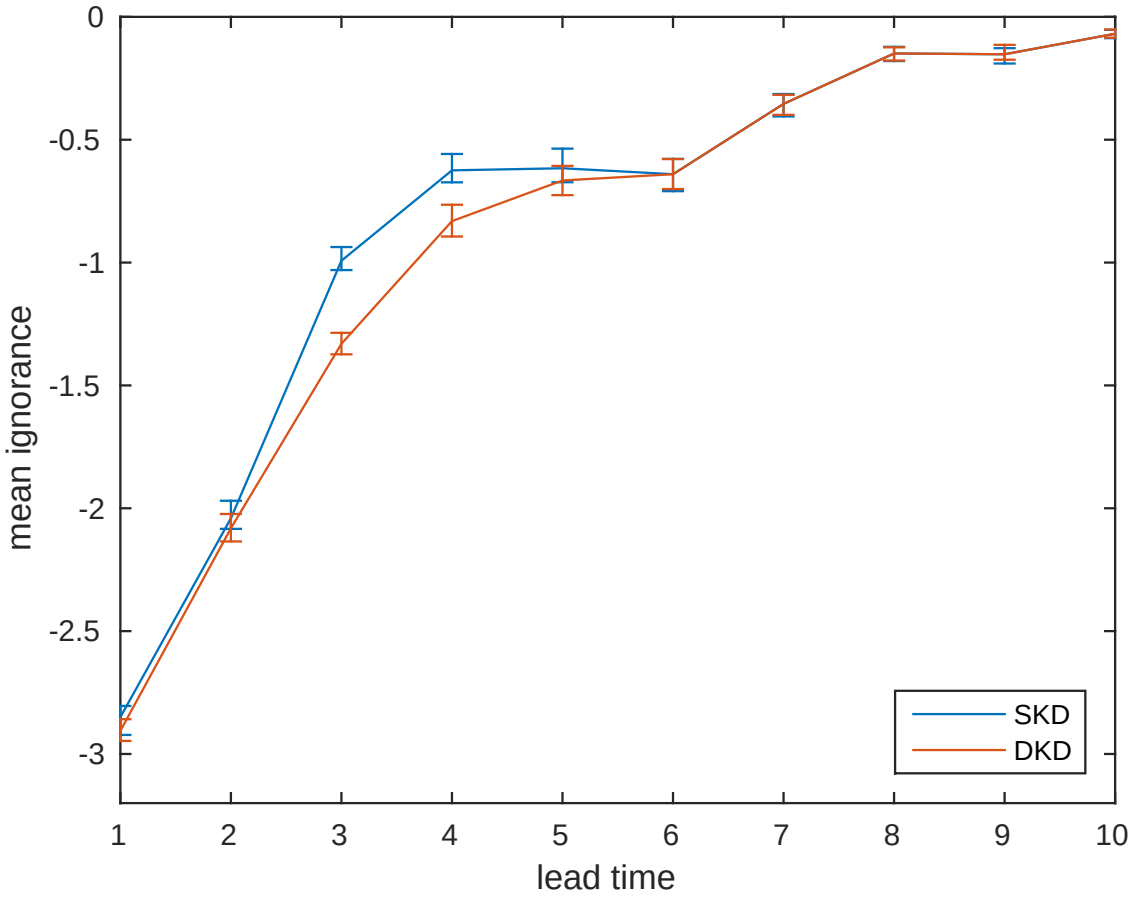

Figure 8.4: The mean ignorance scores of forecast densities formed using simple (blue lines) and dynamic (green lines) kernel dressing for each of 10 different lead times in experiment 8.B. For the shortest lead times, dynamic kernel dressing improves forecast skill whilst for longer lead times, very little improvement is made over simple kernel dressing.

To determine whether the differences in the performance of dynamic and simple kernel dressing are significant, we find bootstrap resampling intervals of the ignorance of the former relative to the latter. These are shown in figure 8.5 for experiments 8.A and 8.B. Shown in blue are the mean relative ignorance scores between the two methods with 95 percent bootstrap resampling intervals for experiment 8.A, in which ensembles are formed using an imperfect model of the Henon map. Here, dynamic kernel dressing performs significantly better than simple kernel dressing at all lead times considered except for 5 and 6 steps ahead. Shown in red is the same but for the imperfect model of the Lorenz '96 system in experiment 8.B. Here, at lead times of 5 days and shorter, dynamic kernel dressing performs significantly better than simple kernel dressing. At longer lead times, the performance of the two methods is very similar since the value of the parameter b in dynamic kernel dressing is close to zero. Combined, the results of experiments 8.A and 8.B support the argument that dynamic kernel dressing has potential to provide improved forecast densities in real life forecasting scenarios.

8.2 Using Boosted Probability Times to Compare Forecasting Systems

The adoption of ensembles as a forecasting tool has proven to be invaluable in the formation of probabilistic forecast densities [157, 149]. The way in which ensembles are formed, however, can make a large difference to the quality of a set of ensemble members and, consequently, the skill of a forecast density. The quality of the forecast densities also depends strongly on the approach taken to their formation, as we have shown in chapter 4.

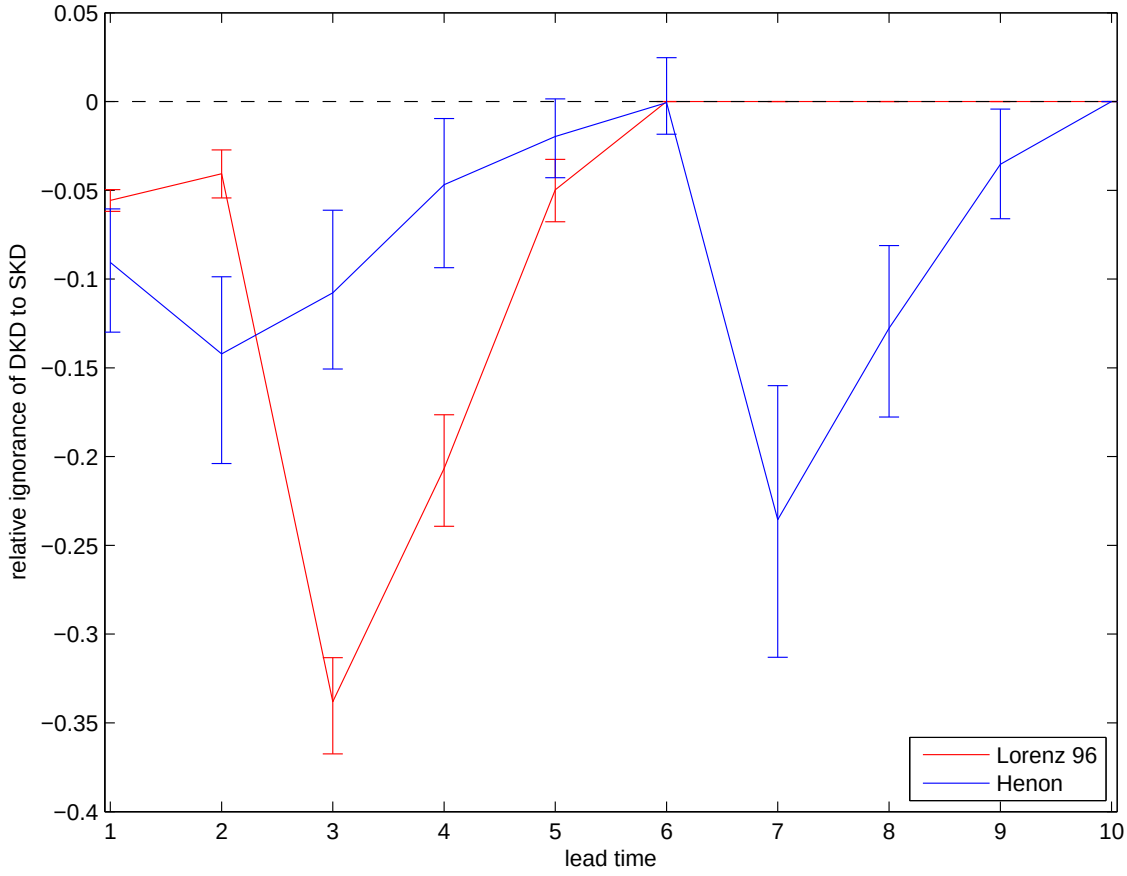

Figure 8.5: Mean ignorance of forecast densities formed using dynamic kernel dressing relative to the ignorance of forecast densities formed using simple kernel dressing as a function of lead time for experiment 8.A (blue), in which ensembles are formed using an imperfect model of the Henon map, and experiment 8.B (red), in which ensembles are formed using an imperfect model of the Lorenz '96 system. The fact that zero does not fall within the resampling bars suggests that dynamic kernel dressing performs significantly better than simple kernel dressing at many of the lead times considered in these two examples.

In this section, we use the approach of boosted probability, defined in section 4.3, to compare the performance of sets of forecast densities constructed from ensembles generated using two different ensemble formation schemes. In each case, we also compare the performance of two approaches to density formation, resulting in a comparison of 4 forecasting systems in total. The two schemes used to form our ensembles are inverse noise and PDA which are both defined in section 2.4.1 whilst our two approaches to the formation of forecast densities are simple kernel dressing, defined in section 2.5.7 and the dynamic kernel dressing approach defined in section 5.5.

Boosted probability provides a good means of comparison between the different approaches in this case because it allows us to assess the length of time for which our forecast densities remain informative relative to the climatology. The different forecasting systems can thus be compared according to their average useful lifetime in terms of sequences of forecast densities.

In experiment 8.C, we use boosted probability to compare the performance of forecast densities formed using each of the forecasting systems described above. To do this, we randomly select 1024 trajectories from the Lorenz '63 attractor and form observations every 6 hours in Lorenz time by adding Gaussian noise at the 5 percent level. For each forecasting system, we then use a perfect model to form 32 member ensembles and construct forecast densities every 6 hours. Boosted probability times are calculated by comparing the density placed on the outcomes by the forecasts and the climatology. Details of the experiment are listed in table B.13.

The mean boosted probability times with their standard deviation in brackets are shown in table 8.1 for each forecasting system.

	Inverse Noise	PDA
Simple KD	43.4 (27.9)	53.5 (31.5)
Dynamic KD	45.3 (27.0)	55.5 (30.4)

Table 8.1: Mean boosted probability times using each combination of ensemble formation scheme and kernel dressing method in experiment 8.C. Standard deviations are shown in brackets in each case.

An alternative illustration of the results of experiment 8.C is shown in figure 8.6. Here the proportion of forecast densities that boost probability at each lead time are shown for each of the four forecasting systems. The red lines indicate that ensembles have been formed using PDA whilst blue lines indicate the use of inverse noise ensembles. Where the lines are solid, dynamic kernel dressing has been used whilst dashed lines indicate the use of simple kernel dressing. The error bars represent 95 percent confidence intervals of the proportions in each case (for clarity, these are only shown at a subset of lead times). Here, PDA ensembles result in a significantly larger proportion of forecast densities that boost probability than inverse noise for all but the longest and shortest lead times. Moreover, dynamic kernel dressing yields a significantly larger proportion that boost probability for moderately long lead times.

8.3 Combining Multiple Lead Time Forecasts

In chapter 6, we investigated whether forecasts launched at different lead times can be combined to yield improved forecast skill. We argued that a Bayesian approach, which we called the Pure Bayes method, is effective if forecast densities can be considered to represent probabilities of the system. In practice, we argued that this is unlikely to be the case even when the model is perfect, due to imperfections in other parts of the forecasting system such as the ensemble formation scheme. Nevertheless, we showed in experiment 6.B, using the Lorenz '63 system, that in

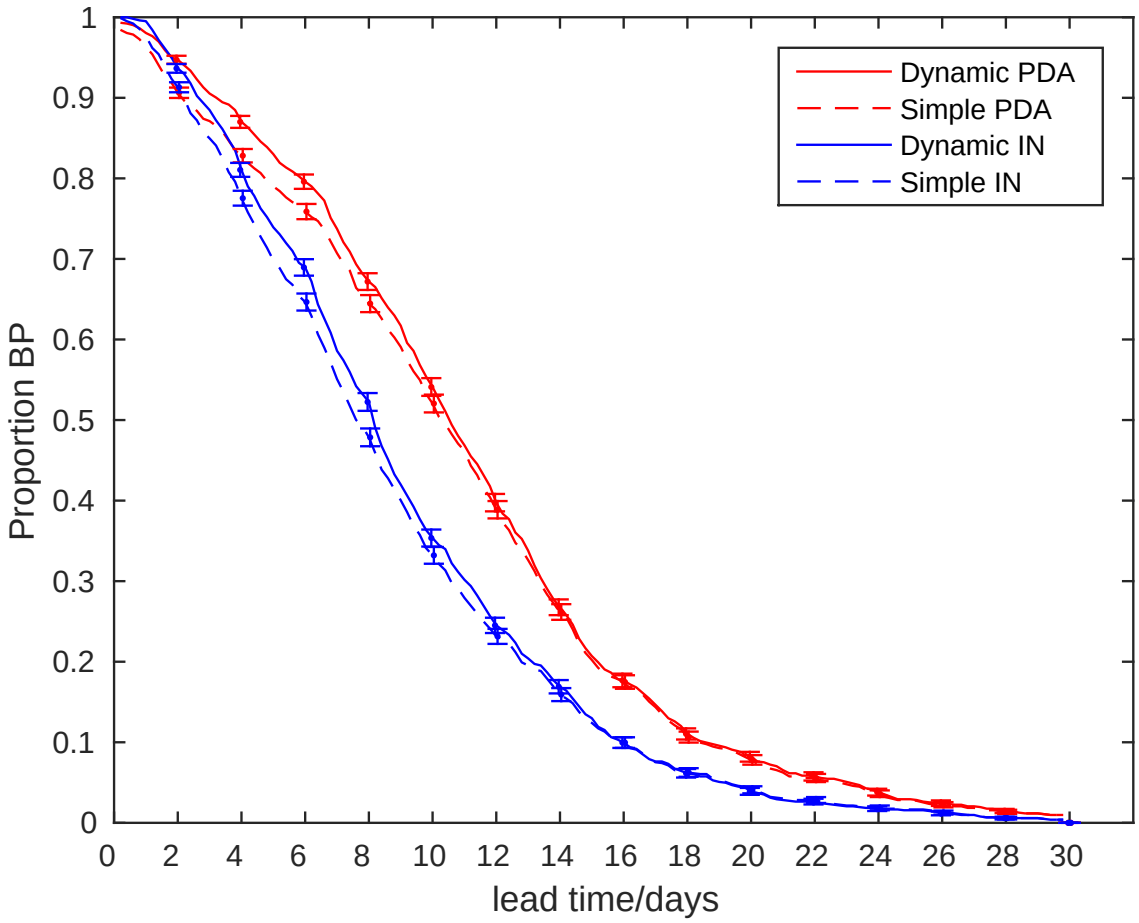

Figure 8.6: Proportion of instances in which probability is boosted as a function of lead time using dynamic (solid lines) and simple (dashed lines) kernel dressing with inverse noise (blue) and PDA ensembles (red) on the ensembles formed from the perfect model of the Lorenz '63 system in experiment 8.C. The error bars represent 95 percent confidence intervals of each proportion at a subset of lead times.

the perfect model scenario, we can form forecast densities that are close enough to representing probabilities for the Pure Bayes method to be successful in improving forecast skill. We showed, however, that in the more realistic scenario in which the model is imperfect, since it is not possible to calibrate the forecasts perfectly, the Pure Bayes method can lead to significant losses of forecast skill. We also showed that two alternative approaches to the combination of forecasts, namely the time-weighted method and sequential blending, can yield significantly improved skill even when the model is imperfect. In this section, we show that similar results can be found for a perfect and an imperfect model of the Moore-Spiegel system. As with experiment 6.B, we begin by considering an imperfect model.

8.3.1 Moore-Spiegel IMS

We make use of the Moore-Spiegel system-model pair defined in appendix A.2.2 setting the value of c that defines the level of imperfection in the model to 8. We form 64 member inverse noise ensembles at a total of 10 different lead times, optimising the parameters over a training set consisting of 1024 ensemble-outcome pairs, and test over a set of the same size. Details of the experiment, which we call experiment 8.D, are listed in table B.14. The results are shown in figure 8.7. In the top panel, the black circles represent the mean ignorance of the original forecasts formed using kernel density estimation with the kernel width chosen using leave one out cross validation. The blue lines represent the mean ignorance scores obtained using the Pure Bayes Method where the beginning of the lines indicate where Pure Bayes was first applied. The red crosses represent the mean ignorance of the time weighted method and the light blue diagonal crosses represent the mean ignorance of sequential blending. In the bottom left panel, the mean ignorance of the Pure Bayes Method is shown again but with the addition of coloured dots indicating whether

the mean ignorance relative to that of the original forecasts is significantly better than (green), worse than (red) or not significantly different from (yellow) the standard method where significance is tested using 95% resampling intervals of the mean difference. In the bottom right panel, the mean relative ignorance scores between sequential blending and the original forecasts (blue), the time weighted method and the original forecasts (red) and the time weighted method and sequential blending (orange) are shown with 95% resampling intervals of the mean difference. As for the Lorenz '63 case, demonstrated in chapter 6, the Pure Bayes Method appears to be counterproductive, in general, yielding significantly less skillful forecast densities on average. Both the time-weighted method and sequential blending, on the other hand, yield significantly improved forecast densities at a number of lead times. Out of the time-weighted method and sequential blending, neither appears to be consistently better than the other.

8.3.2 Moore-Spiegel PMS

We now show that, as in experiment 6.B, replacing the imperfect model with a perfect model can make the forecasting system perform well enough for the Pure Bayes method to yield improved forecast densities at some lead times. We perform the same experiment as in the previous section but with a perfect model. The results are shown in figure 8.8 which is the same as figure 8.7 but with the imperfect model replaced with the perfect model. Here, as suspected, there is a large improvement in the performance of the Pure Bayes method. At the shortest lead times, the forecasts formed using this approach outperform, on average, the original forecasts whilst this is not the case at longer lead times. The reason there is not consistent improvement at longer lead times is likely due to other imperfections in the forecast system, such as the ensemble formation scheme, which tend to be magnified over time. Both the

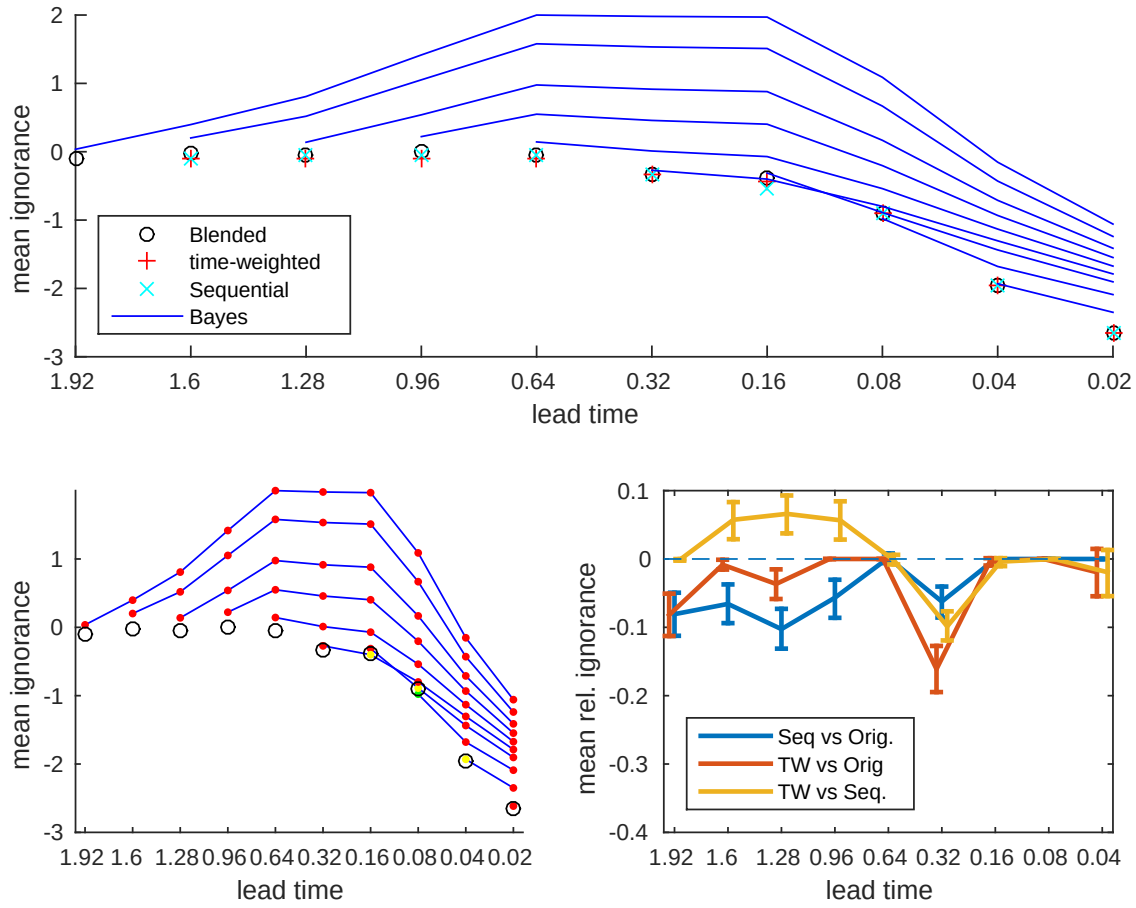

Figure 8.7: Top: The mean ignorance scores obtained by applying the Pure Bayes method (blue lines), the time-weighted method (red crosses) and sequential blending (cyan diagonal crosses) for the imperfect model of the Moore-Spiegel system in experiment 8.D. The black circles represent the mean ignorance scores of the original forecast densities. Bottom left: The mean ignorance scores of the Pure Bayes method with coloured dots indicating whether the forecasts are significantly than (better), worse than (red) or not significantly different (yellow) from the original forecasts. Bottom right: Mean relative ignorance scores between sequential blending and the original forecasts (blue), the time weighted method and the original forecasts (red) and the time weighted method and sequential blending (orange) with 95% resampling intervals of the mean difference. Whilst the Pure Bayes method tends to yield forecast densities that are significantly worse than the original forecasts, the time-weighted method and sequential blending yield improved forecast densities in around half of the lead times considered.

time-weighted method and sequential blending, as in the imperfect model scenario, yield improved forecast densities at all but the two shortest lead times.

8.4 Linking shadowing and forecast skill

In our forecasting framework, forecast densities are formed by evolving forward an ensemble of initial conditions and converting them into density functions. The skill of the forecast densities, therefore, depends both on the quality of the ensembles and the methodology used to convert the ensembles into densities. If we can differentiate between these two factors, we can identify whether there is room for improvement in the way in which forecast densities are formed. This requires us to evaluate the quality of our ensembles separately from our forecast densities.

For a model trajectory to give a reliable prediction, we would hope for it to stay close to all observations up to the forecast target time. Shadowing is a useful indication of whether this is the case. The proportion of ensemble members that shadow thus gives an indication (though not a necessary condition) of whether the ensembles contain useful information at a given lead time.

In this experiment, we demonstrate how we can identify shortcomings in our methodology by evaluating the performance of the raw ensembles. Using the imperfect model of the Ikeda map defined in appendix A.1.4 with imperfection parameter $c = 8$, we form 1024 member ensembles using inverse noise at lead times ranging from 1 to 24 time steps ahead. We form forecast densities by using the simple approach described in section 2.4.4 of fitting Gaussian distributions in the form $N(m(\boldsymbol{x}), v(\boldsymbol{x}))$ where $m(\boldsymbol{x})$ and $v(\boldsymbol{x})$ are the ensemble mean and variance respectively. To evaluate the performance of the ensembles, we find the mean proportion

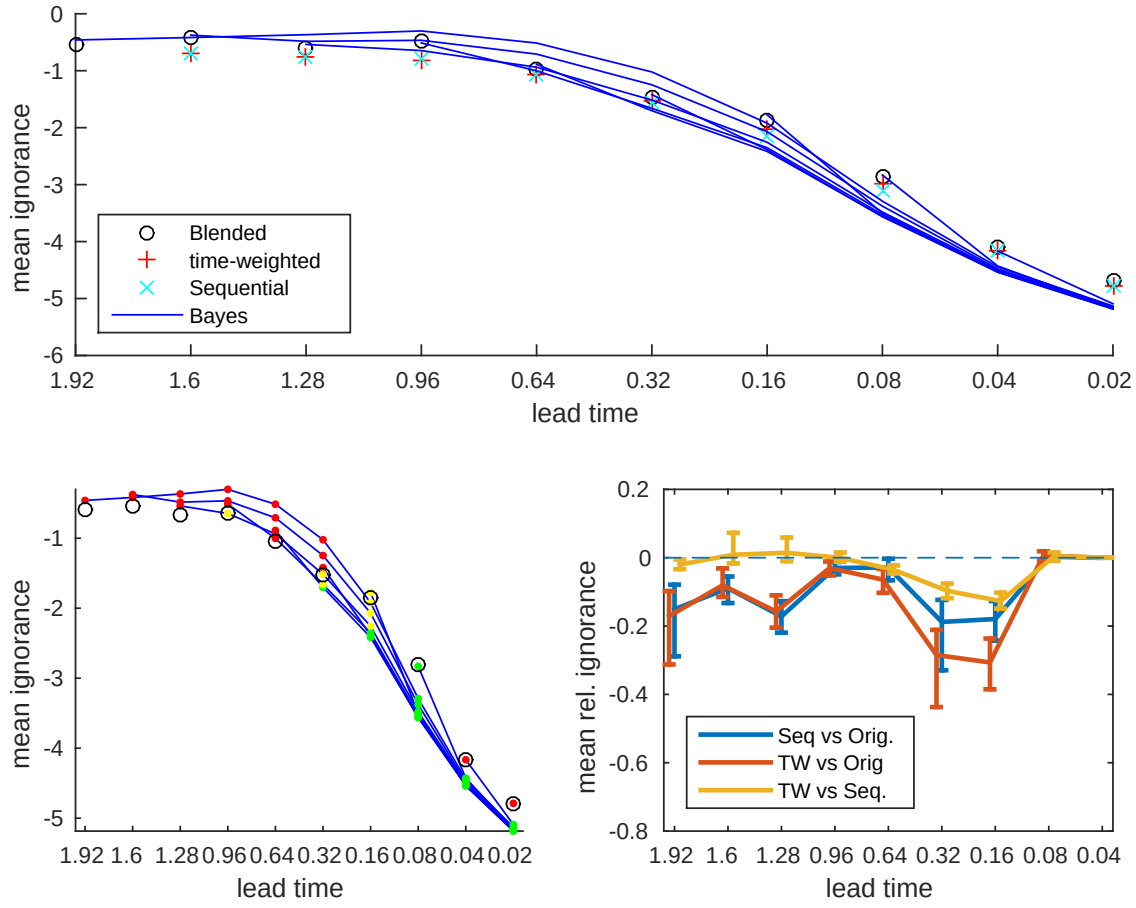

Figure 8.8: Top: The mean ignorance scores obtained by applying the Pure Bayes method (blue lines), the time-weighted method (red crosses) and sequential blending (cyan diagonal crosses) for the perfect model of the Moore-Spiegel system in experiment 8.D. The black circles represent the mean ignorance scores of the original forecast densities. Bottom left: The mean ignorance scores of the Pure Bayes method with coloured dots indicating whether the forecasts are significantly than (better), worse than (red) or not significantly different from (yellow) the original forecasts. Bottom right: Mean relative ignorance scores between sequential blending and the original forecasts (blue), the time weighted method and the original forecasts (red) and the time weighted method and sequential blending (orange) with 95% resampling intervals of the mean difference. In this case, for shorter lead times, the Pure Bayes method tends to yield forecast densities that are significantly better than the original forecasts whilst at longer lead times they tend to be worse. Both the time weighted method and sequential blending are able to yield significantly improved forecasts for all but the two shortest lead times.

of members that shadow at each forecast lead time. Details of the experiment, which we call experiment 8.E, are listed in table B.15.

The results of the experiment are shown in figure 8.9. The blue line, corresponding to values on the left hand y axis, shows the mean proportion of ensemble members that shadow the observations, whilst the red line, corresponding to the values on the right hand y axis, shows the mean ignorance of the forecast densities, each as a function of forecast lead time. The error bars represent 95% resampling intervals of the mean ignorance. At short lead times, up to 6 steps ahead, the forecast densities perform better, on average, than climatology whilst for all lead times longer than this, they perform worse, albeit not generally significantly. Looking at the performance of the ensembles, however, it appears that, at some lead times, the ensembles contain useful information yet this is not reflected in the skill of the forecast densities. For example, over 20 percent of ensemble members shadow the observations up to 7 time steps ahead, yet our forecast densities perform worse than climatology at this lead time. This suggests that there is room for improvement in the way the forecast densities are constructed.

In building forecast densities, our aim is very different to that of estimating the underlying distribution of a set of data points. This is because, due to model error and other imperfections in the forecasting system, the ensemble and the outcome cannot be considered to be drawn from the same distribution. This means that some allowance should be made to account for these imperfections. Throughout this thesis, up to now, we have followed [22] in that we always blend our forecasts with climatology. We now show how this approach can improve our forecast densities. The mean ignorance scores of the Gaussian forecast densities blended with climatology are shown in magenta in figure 8.10 along with those of the original

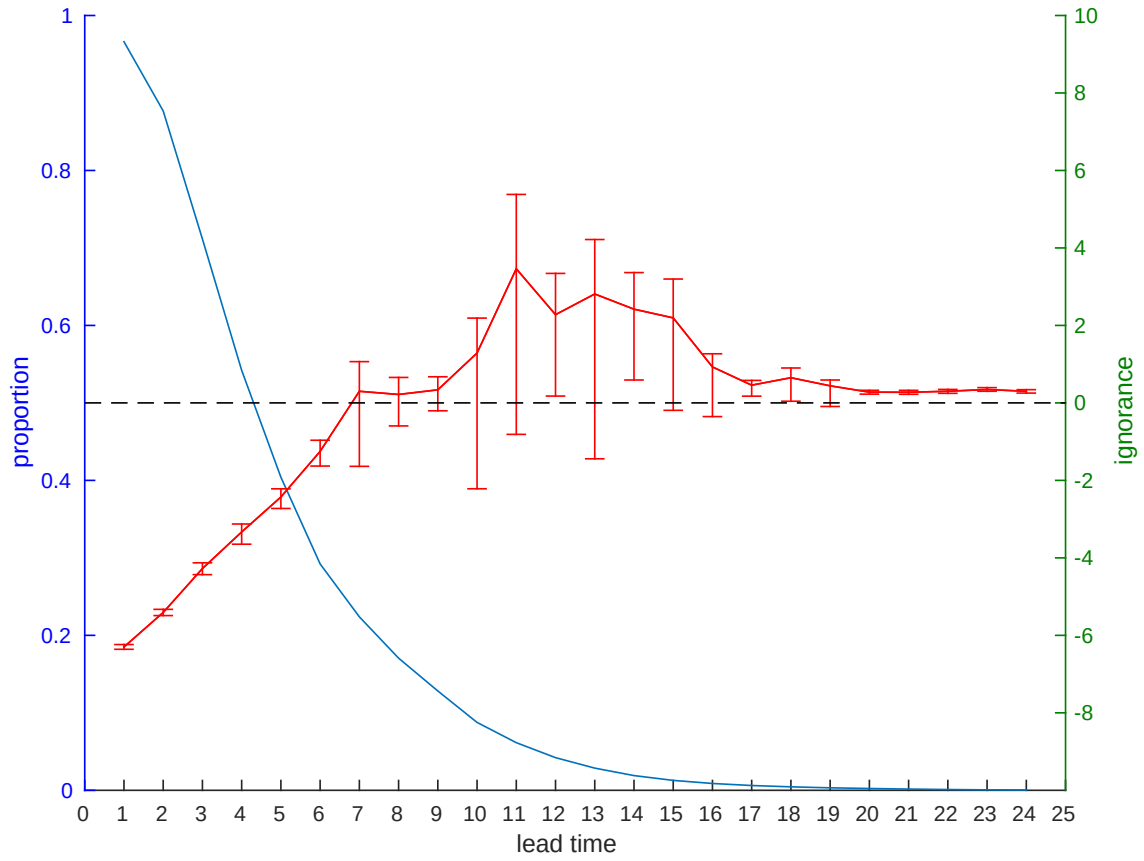

Figure 8.9: The mean proportion of ensemble members that shadow the observations (blue line, left axis) and the mean ignorance score obtained by fitting Gaussian distributions (red starred line, right axis) to the ensemble members as a function of forecast lead time for the ensembles in experiment 8.E. The error bars represent 95% resampling intervals of the mean ignorance. Points that lie below the dashed line indicate forecast skill with respect to climatology. For lead times greater than 6 time steps, a large proportion of ensemble members shadow whilst the forecast densities formed using this approach perform worse, though not generally significantly, than climatology. This suggests that a better approach to forecast density formation could be taken.

Gaussian forecasts, which, as in figure 8.9, are shown in red. Here, the maximum lead time at which we can build informative forecast densities increases from 6 to 13 time steps. This appears to be much more in line with the level of predictability suggested by the performance of the ensembles.

In fact, with the methods used throughout this thesis, we can improve our forecast densities further. In nonlinear systems, we do not expect our forecast densities to be well represented with Gaussian distributions. Kernel dressing makes no assumption about the underlying distribution and can therefore be expected to perform better. The mean ignorance scores obtained using the dynamic kernel dressing method described in section 5.5 are shown in green in figure 8.11 along with the ignorance scores for the blended and unblended Gaussian forecasts which are shown in magenta and red respectively. Here, whilst the maximum lead time at which our forecasts are informative does not increase, the skill of the forecast densities improves for lead times between 3 and 12 time steps. Note how kernel dressing fails to improve the skill of the forecasts at the very shortest lead times. This is easily explained by the approach taken to ensemble formation. Since the noise model is Gaussian, inverse noise ensembles are sampled from a Gaussian distribution centred around the observation. The distributions of the ensemble members are thus reasonably well represented with Gaussian distributions for several time steps until the nonlinear dynamics cause them to deviate from this distribution.

A similar experiment, using the imperfect model of the Henon map defined in appendix A.1.2 with imperfection parameter $c = 12$, is shown in figure 8.12. Details of the experiment, which we call experiment 8.F, are shown in table B.16. Here, again, the mean proportion of ensemble members that shadow the observations gives a measure of the potential predictability of forecast densities. Whilst applying the

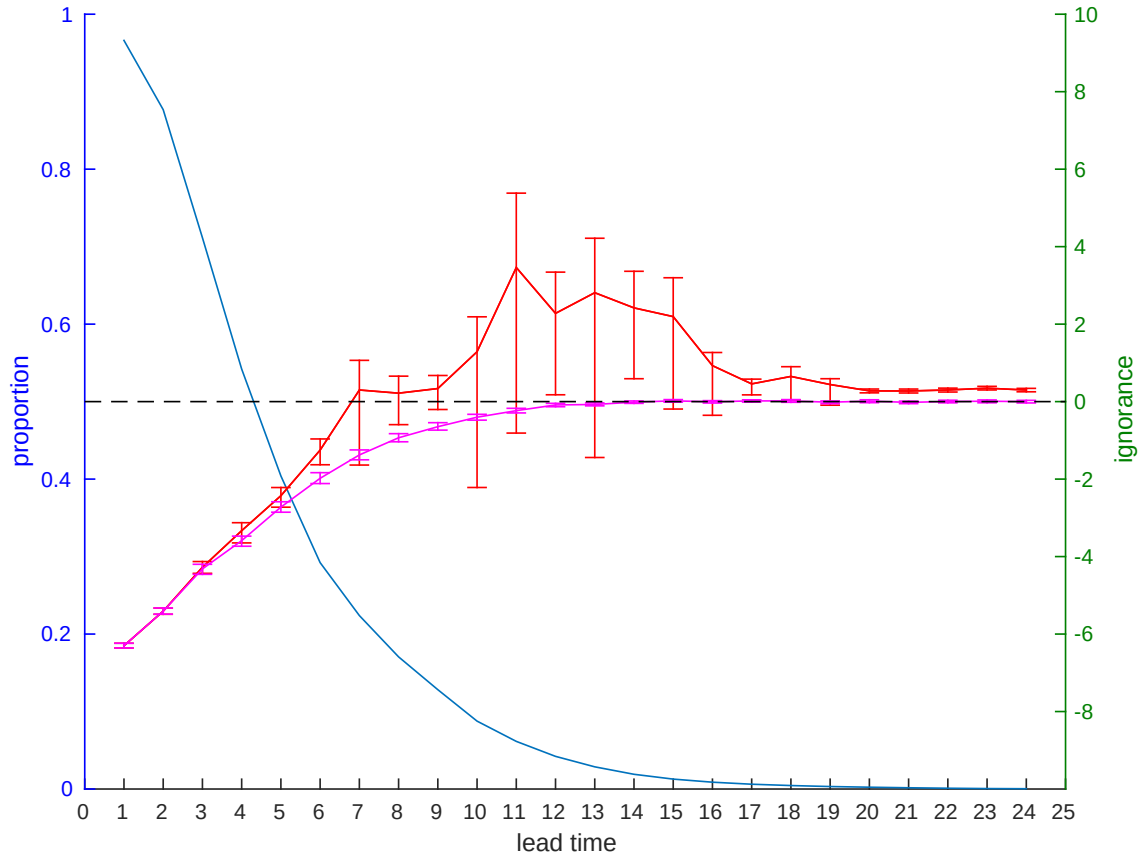

Figure 8.10: The mean proportion of ensemble members that shadow the observations (blue line, left axis) and the mean ignorance score obtained by fitting Gaussian distributions with (magenta starred line, right axis) and without (red starred line, right axis) blending as a function of forecast lead time. The error bars represent 95% resampling intervals of the mean ignorance. Forecasts formed without blending only yield significant skill with respect to climatology up to 6 steps ahead whilst blending increases this time to 13 steps.

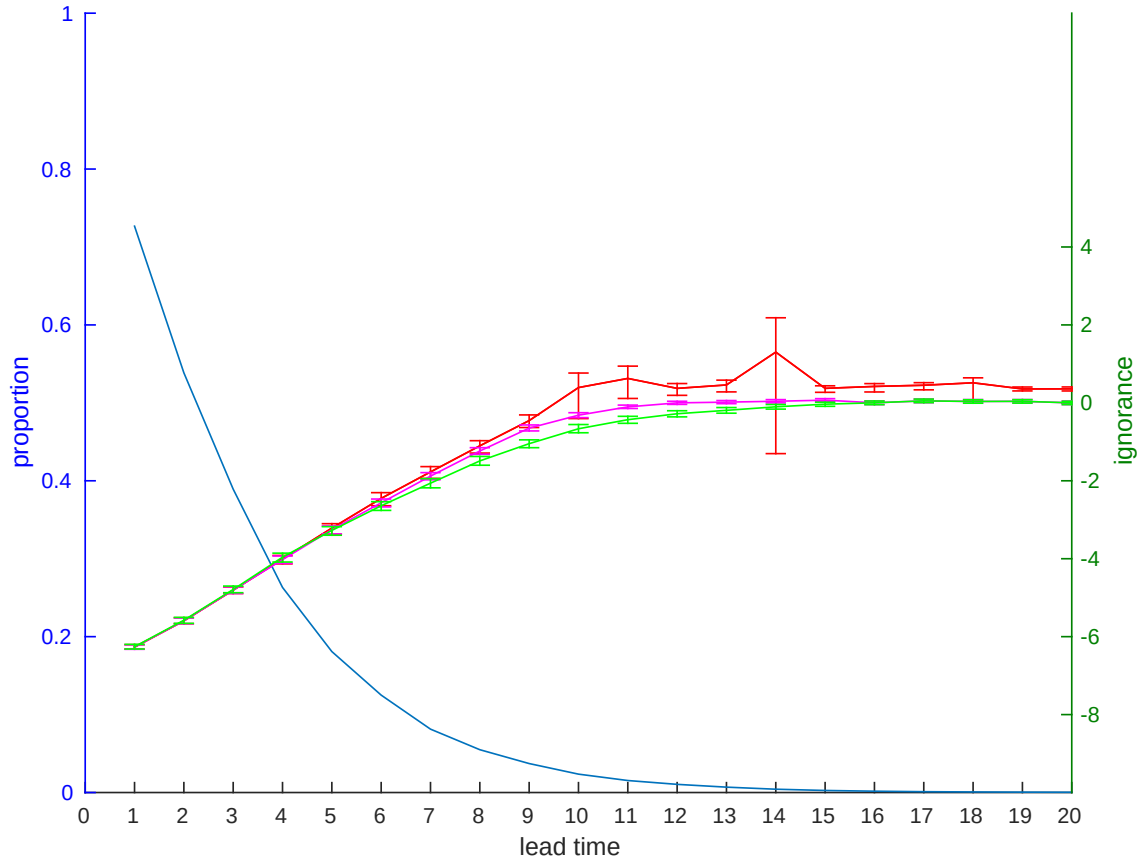

Figure 8.11: The mean proportion of ensemble members that shadow the observations (blue line, left axis), the mean ignorance score obtained by fitting Gaussian distributions with (magenta starred line, right axis) and without (red starred line, right axis) blending and the mean ignorance score obtained using dynamic kernel dressing (green starred line, right axis) as a function of forecast lead time in experiment 8.E. The error bars represent 95% resampling intervals of the mean ignorance. Dynamic kernel dressing makes no assumption about the underlying distribution and thus yields better forecast skill than the Gaussian approach.

Gaussian approach without blending yields informative forecast densities up to 10 steps ahead, on average, a small proportion of ensemble members still shadow at this time and beyond. This suggests that it might be possible to find informative forecast densities beyond this lead time. This is confirmed by the performance of dynamic kernel dressing which provides informative forecast densities up to 15 steps ahead.

In these experiments, we demonstrated how calculating shadowing lengths of ensemble members can help to identify shortcomings in the methodology used to form forecast densities. This is potentially very useful in deciding where to allocate resources in order to improve the skill of a set of forecast densities. Moreover, we have shown how great care should be taken in the formation of forecast densities so that the best use is made of the information contained in the ensembles.

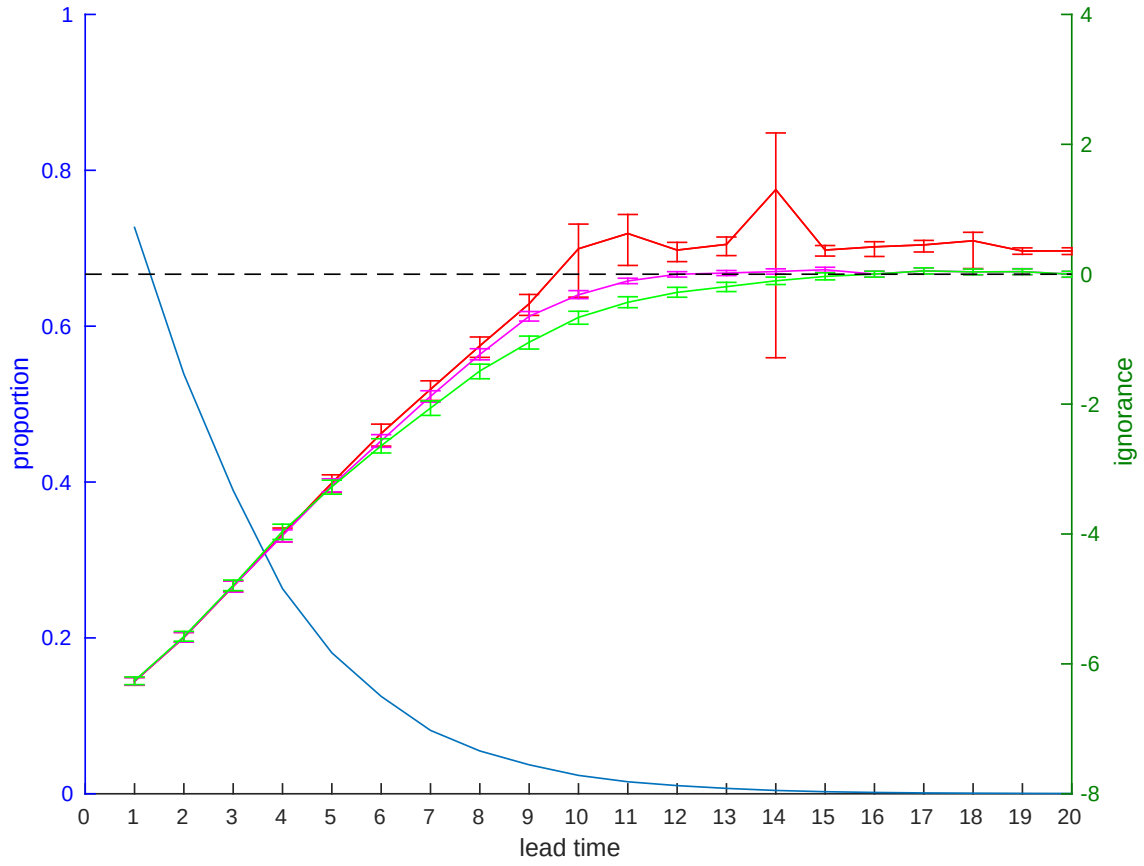

Figure 8.12: The mean proportion of ensemble members that shadow the observations (blue line, left axis), the mean ignorance score obtained by fitting Gaussian distributions with (magenta starred line, right axis) and without (red starred line, right axis) blending and the mean ignorance score obtained using dynamic kernel dressing (green starred line, right axis) as a function of forecast lead time in experiment 8.F. The error bars represent 95% resampling intervals of the mean ignorance. For the first 9 lead times, the approach of fitting Gaussian distributions without blending yields forecast densities that perform significantly better than climatology. The fact that a small proportion of ensemble members shadow longer than this time suggests that informative forecast densities may be found. This is confirmed by the performance of the forecasts formed using dynamic kernel dressing which yield better skill than the climatology up to 15 steps ahead.

Chapter 9

Conclusion

In this chapter, we discuss the implications of the most important contributions in this thesis.

The quality of probabilistic forecasts in the imperfect model scenario (in which all real life applications fall) depends, not only on the quality of the model, but on other aspects such as the ensemble formation scheme, the size of the ensemble and the quality of the density formation scheme. It is therefore important to evaluate the entire forecasting system. Not only have we shown how to improve many aspects of a forecasting system, we have described new approaches to evaluating forecasts.

Evaluating point forecasts is fraught with difficulties since, when the underlying system is nonlinear, many traditional metrics fail in that they cannot be expected to favour a perfect model over an imperfect one. Shadowing is a logical and intuitive method of comparing the performance of forecast trajectories. Shadowing ratios, introduced in this thesis, give a different perspective on the shadowing lengths of two forecasting systems. These approaches have potentially far reaching benefits in

the evaluation of both point forecasts and individual components of the probabilistic forecasting framework. Moving forward, it would be interesting to use shadowing and shadowing ratios to compare the performance of forecasts of real life dynamical systems such as the weather.

A shortcoming of simple kernel dressing is that the dispersion of an individual ensemble is not accounted for when the kernel width is fitted. We introduced two new approaches to kernel dressing and described another that take the ensemble standard deviation into account and showed that each one can improve forecast skill. These results have important implications for the building of forecast densities. Some measure of the dispersion of an ensemble should always be taken into account. We thus suggest that dynamic kernel dressing should routinely be considered as a method of building forecast densities. Many real world systems exhibit nonlinear properties and hence ensembles are likely to vary in their dispersion. It would thus be interesting to assess the performance of dynamic kernel dressing with ensembles formed using models of such systems.

We showed in two different imperfect model scenarios that a Bayesian approach, in which forecast densities formed at different lead times are combined, is not successful in improving forecast skill. This has important implications for the way in which we treat forecast densities. Were our forecast densities able to be treated as probabilities, a Bayesian approach would be expected to make optimal use of the information in terms of maximising the likelihood and thus minimising the ignorance and would thus be expected to outperform all other approaches with the same information available to them. The fact that better performing forecasts can be obtained by not treating the original forecast densities as probabilities calls into question the entire probability calculus. This has implications for the way in which forecast densities

should be interpreted. Probabilities derived from a forecast density, whilst useful in terms of information, can not, in general, be considered to represent probabilities of the system.

Nomenclature

α	Blending parameter that governs the weighting placed on model based forecast information and the climatology
α_h	Hyperparameter in Bayesian inference
α_i	Blending parameter of the i_{th} group in K groups kernel dressing
$\boldsymbol{\theta}$	Parameter vector
δ	Euler step size in PDA
η	Ensemble Shadowing Ratio
$\mathbf{s}_t$	observation of true state at time t
$\mathbf{x}_c$	Multiple lead time ensemble
ϕ	Shadowing Ratio
σ	Kernel width in simple kernel dressing
σ_d	Kernel width in dynamic kernel dressing
σ_i	Kernel width parameter of the i_{th} group in K groups kernel dressing

σ_{KL} Kernel width that minimises the KL divergence between the model and system densities

$\tau_{i,new}$ i_{th} shadowing length formed using a 'new' method in shadowing ratios

$\tau_{i,ref}$ i_{th} shadowing length formed using a 'reference' method in shadowing ratios

$\tilde{\tau}_{i,new}$ i_{th} shadowing time of a 'new' forecasting system in ensemble shadowing ratios

$\tilde{\tau}_{i,ref}$ i_{th} shadowing time of a reference forecasting system in ensemble shadowing ratios

$\tilde{F}$ System Dynamics

$\tilde{F}$ System dynamics

$\tilde{p}$ Ensemble shadowing proportion

$\tilde{x}_t$ Point forecast calibrated using linear regression

υ_i Offset parameter of the i_{th} group in K groups kernel dressing

a Intercept term in dynamic kernel dressing

$A(\boldsymbol{u})$ Adjoint matrix

b Gradient term in dynamic kernel dressing

B_p Boosted Probability Time

d Dimension of a system

F Model dynamics

G_i i_{th} group in K groups kernel dressing

K Number of groups in K groups kernel dressing

$K(.)$ Kernel Function

l lead time

$L(u)$ PDA mismatch cost function

M Number of ensemble-outcome pairs selected from the training set in fixed window kernel dressing

m number of ensemble members

$m(.)$ ensemble mean

N_i Size of the i_{th} group in K groups kernel dressing

N_{tr} Number of ensemble-outcome pairs in a training set

$p_{clim}(.)$ Density of the climatological distribution at x

$p_i^{bayes}(.)$ Forecast density formed using the Bayesian approach to combining forecast densities at the i_{th} lead time.

$p_i^s(.)$ Sequentially blended forecast at the i_{th} lead time.

P_{win} Subset of training set used to optimise parameters in fixed window kernel dressing

$p_w(.)$ Forecast density formed using the time-weighted method

q System density

r_i Weighting placed on the forecast density launched at the i_{th} lead time in sequential blending.

S	Scoring rule
t	time
$v(.)$	ensemble variance
w_i	Weighting on the forecast launched at the i_{th} lead time in the time-weighted method
Y	Forecast outcome
s	standard deviation of a sample or ensemble

Appendix A

Dynamical Systems

A.1 Maps

A.1.1 Logistic Map

The logistic map is a one dimensional map introduced in [104] as a simple model of changing animal populations. Each system state lies on the range [0,1] and represents the existing size of a population as a proportion of the maximum possible population size. The behaviour of the system states is described by the difference equation

$$x_{i+1} = \lambda x_i(1 - x_i) \tag{A.1}$$

where λ is a system parameter. The logistic map is chaotic for most values of λ on the range $[3.57, 4]$. The logistic map is often used as a demonstration of how simple nonlinear equations can show chaotic behaviour.

A.1.2 Henon map

The Henon map [73] is a two dimensional discrete time map originally introduced as a simplifed model of the Lorenz '63 system. It is defined by the difference equations

$$\begin{aligned} X_{n+1} &= 1 - aX_n^2 + Y_n \\ Y_{n+1} &= bX_n \end{aligned} \tag{A.2}$$

where a and b are parameter values. The parameter values are commonly set to $a = 1.4$ and $b = 0.3$ for which the system shows chaotic behaviour. The system attractor for these parameter values is shown in figure A.1.

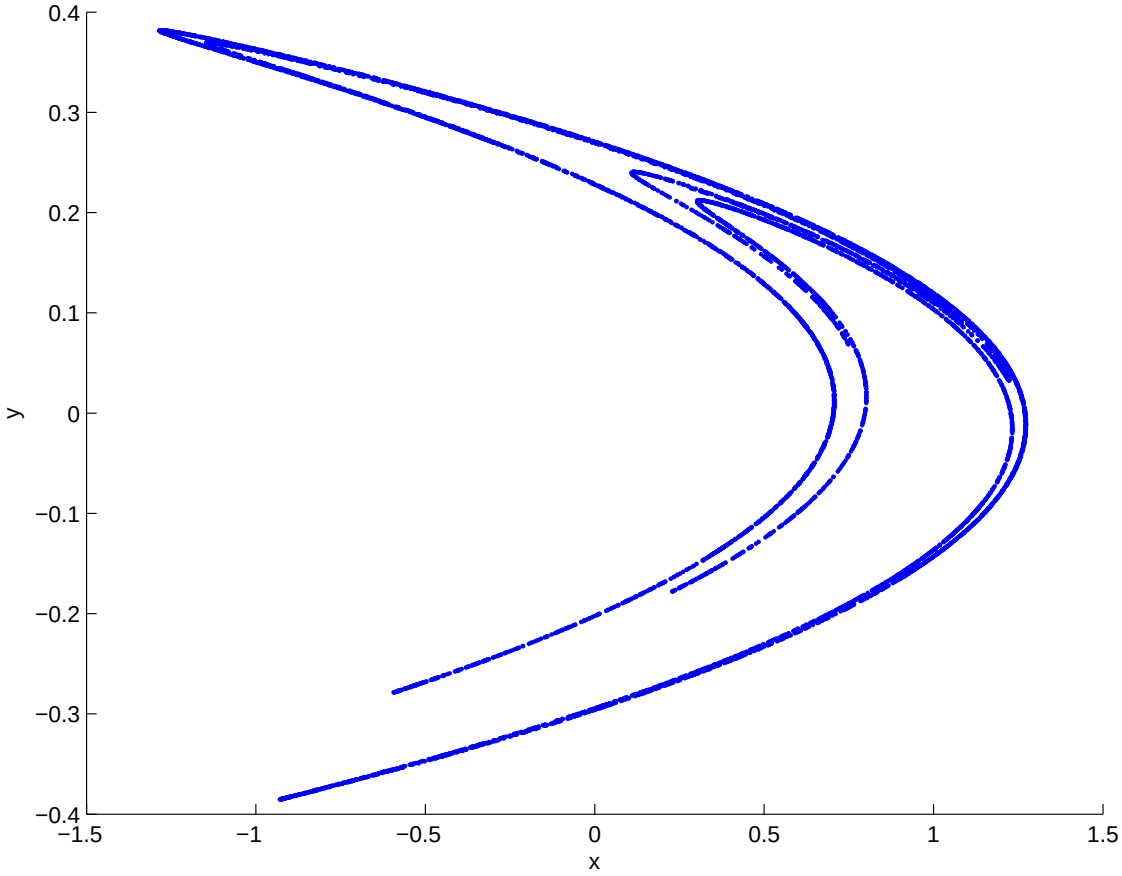

Figure A.1: The Henon map attractor with parameter values $a = 1.4$ and $b = 0.3$.

Imperfect Model

We form an imperfect model of the Henon map by replacing x in equation A.2 with $x' = c\sin(\frac{x}{c})$ where the parameter c governs the level of imperfection in the model.

A.1.3 Duffing Map

The Duffing map is a 2 dimensional map, originally used to model the motion of a damped oscillator, described by the equations

$$\begin{aligned} X_{n+1} &= y_n \\ Y_{n+1} &= -bx_n + ay_n - y_n^3 \end{aligned} \tag{A.3}$$

where a and b are parameters. The Duffing map can be shown to exhibit chaotic behaviour when the parameters are set to $a = 2.75$ and $b = 0.2$. The attractor for these parameter values is shown in figure A.2.

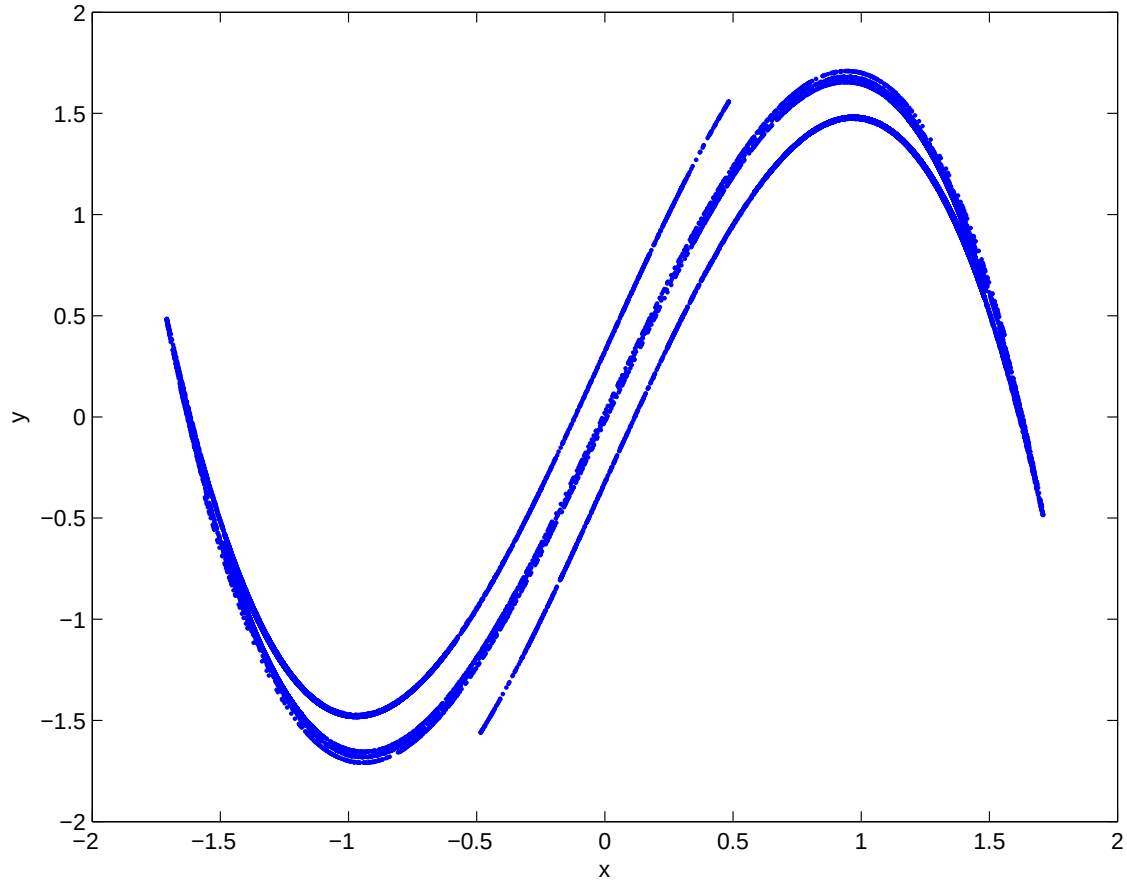

Figure A.2: The Duffing map attractor with parameter values $a = 2.75$ and $b = 0.2$.

Imperfect Model

We form an imperfect model of the Duffing map by replacing x in equation A.3 with $x' = c\sin(\frac{x}{c})$ where the parameter c governs the level of imperfection in the model.

A.1.4 Ikeda map

The Ikeda map [74] is a two dimensional discrete dynamical system originally designed as a model of light going around a nonlinear optical cavity. The equations

are given by

$$x_{n+1} = \gamma + u(x_n \cos\phi - y_n \sin\phi) \tag{A.4}$$

$$y_{n+1} = u(x_n \sin\phi + y_n \cos\phi) \tag{A.5}$$

where $\phi = \beta - \alpha/(1 + x_n^2 + y_n^2)$ In this thesis, the parameter values used are $\alpha = 6$, $\beta = 0.4$, $\gamma = 1$ and $u = 0.83$ for which the system is chaotic. The attractor for these parameter values is shown in figure A.3.

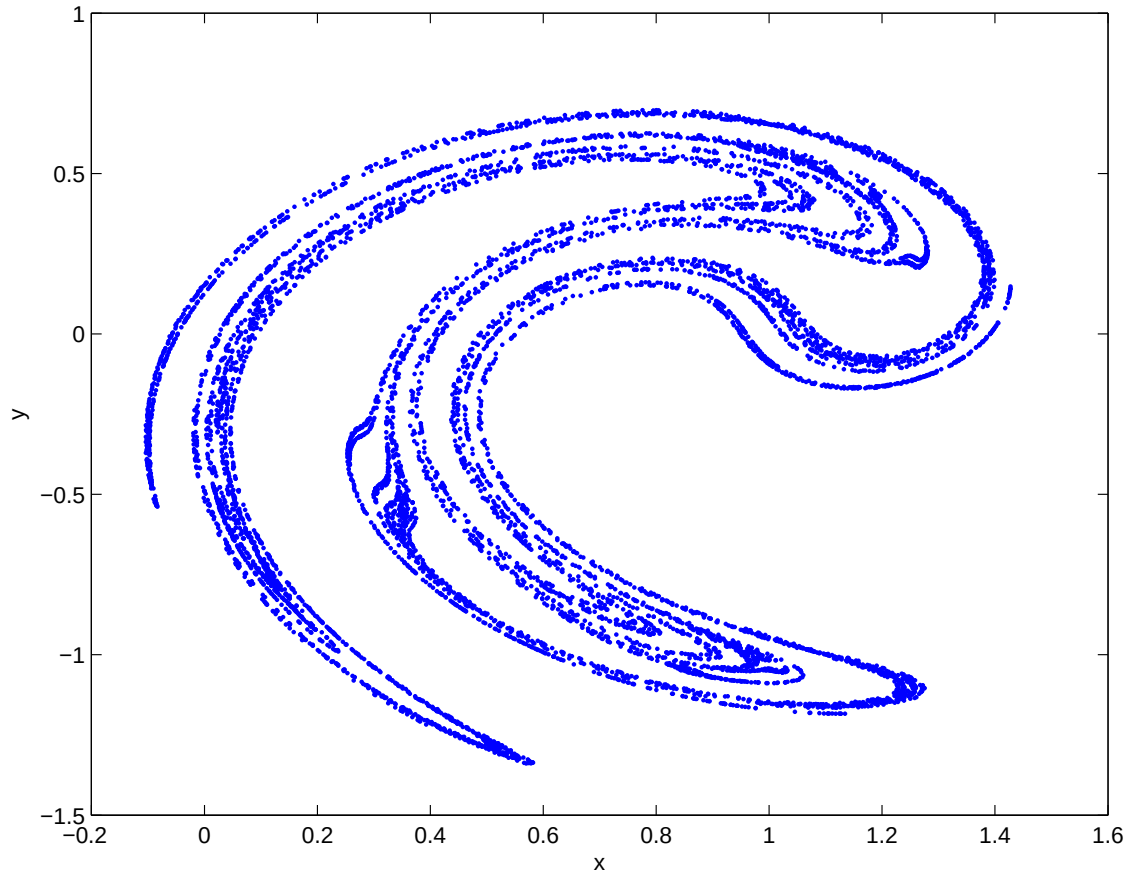

Figure A.3: The Ikeda map attractor with parameter values $\alpha = 6$, $\beta = 0.4$, $\gamma = 1$ and $u = 0.83$.

Imperfect Model

An imperfect model of the Ikeda map is formed by replacing the x variable in the system equations with $x' = c\sin(\frac{x}{c})$ where the parameter c governs the level of imperfection in the model.

A.2 Flows

A.2.1 Lorenz '63

The Lorenz '63 system [99] is a 3 dimensional system of ordinary differential equations proposed by Lorenz in 1963 as a simplified model of atmospheric convection. The equations, often known as the Lorenz equations, are

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y - x) \\ \frac{dy}{dt} &= x(\rho - z) - y \\ \frac{dz}{dt} &= xy - \beta z \end{aligned} \tag{A.6}$$

In this thesis, we use the parameter values $\rho = 28$, $\sigma = 10$ and $\beta = \frac{8}{3}$ for which the system is chaotic. The Lorenz '63 attractor for these parameter values is shown in figure A.4. Since the doubling time (the average time taken for initial condition error to double in magnitude) for this system is approximately one unit and the doubling time of state of the art weather models is around 5 days, 0.2 units of time in the Lorenz '63 system can be considered to be represent one day in real time.

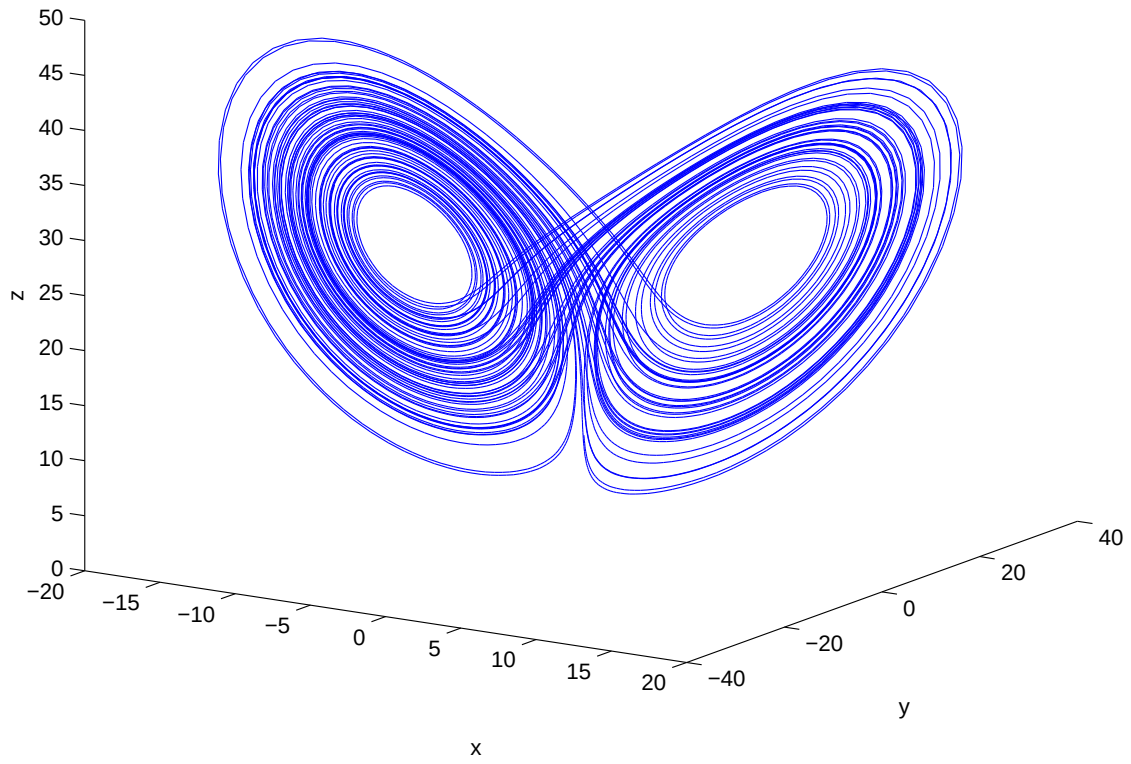

Figure A.4: The Lorenz attractor with parameter values $\rho = 28$, $\sigma = 10$ and $\beta = \frac{8}{3}$

Imperfect Model

To form an imperfect model of the Lorenz '63 system, we replace the x in the system equations with $x' = c\sin(\frac{x}{c})$ where c is a parameter that governs the level of imperfection in the model.

A.2.2 Moore-Spiegel system

The Moore-Spiegel system [107] is a 3 dimensional system, originally proposed to model the height of ionised gas in the atmosphere of a star, defined by the differential

equations

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= z \\ \frac{dz}{dt} &= -x - (T - R + Rx^2)y - Tx \end{aligned} \tag{A.7}$$

In this thesis, we use the parameter values $R = 100$ and $T = 35$ for which the system is chaotic.

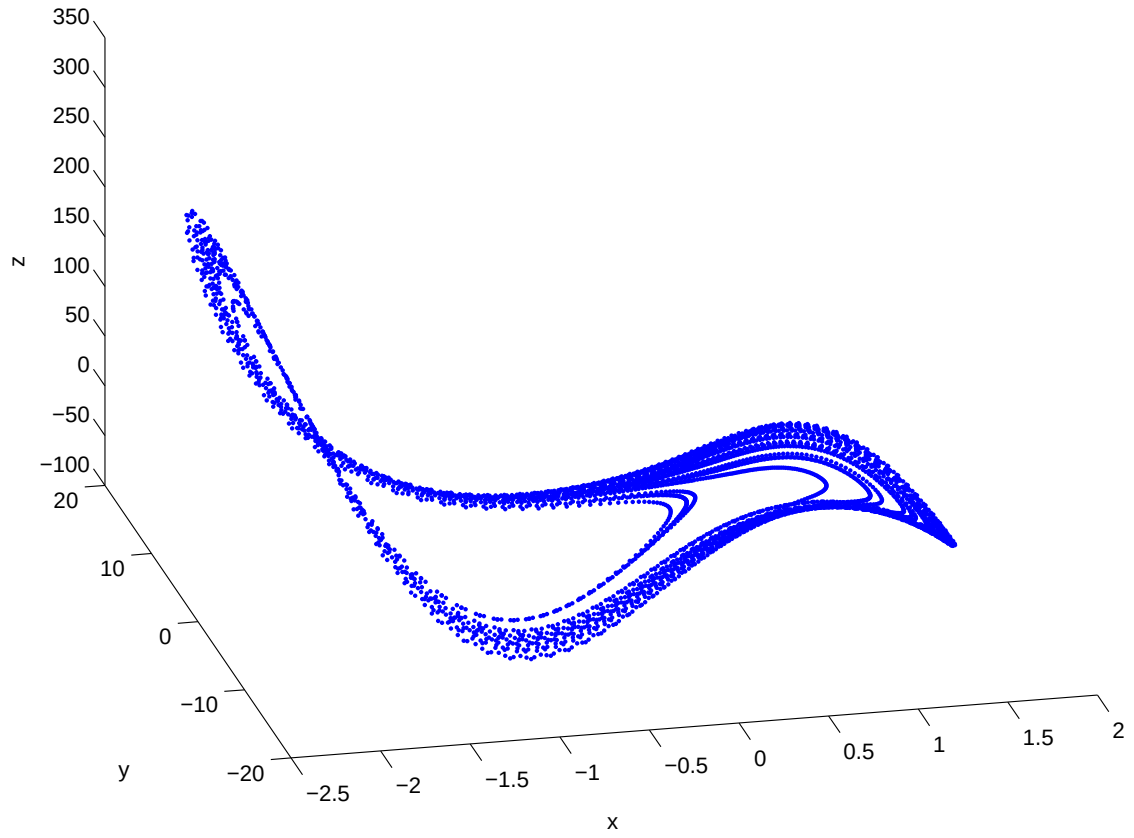

Figure A.5: The Moore Spiegel attractor with parameter values $R = 100$ and $T = 35$.

Imperfect Model

To form an imperfect model of the Moore-Spiegel system, we replace the x variable in the system equations with $x' = c\sin(\frac{x}{c})$ where c is a parameter that governs the level of imperfection in the model.

A.2.3 Lorenz '96

With the aim of obtaining a dynamical system analogous to climatic variables spatially distributed around the Earth, Lorenz [98] introduced 2 sets of nonlinear ODEs in 1995 calling them model 1 and model 2. The Lorenz '96 system is deigned so that model 2 consists of m 'slow' large scale observed variables coupled to mxn 'fast' small scale unobserved variables. Model 1 is then used to approximate the system dynamics of model 2.

Model 1

Consisting of m variables $x_1, ..., x_m$ which have cyclical boundary conditions (i.e. $x_{m+1} = x_1$) The equations are

$$\frac{dx_i}{dt} = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + F \tag{A.8}$$

for $i = 1, ..., m$ where the parameter F is a forcing parameter. For $F = 8$, the doubling time is around one unit and hence Lorenz considered 0.2 time units to be roughly equivalent to one day in real time for this parameter value.

Model 2

In model 2, each of the variables $x_1, ..., x_m$ are coupled to another set of variables $y_{1,i}, ..., y_{n,i}$ which also have cyclical boundary conditions (i.e. $y_{j+1,i} = y_{1,i}$). The

ODE's for model 2 are

$$\frac{dx_i}{dt} = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + F - \frac{h_x c}{b}\sum_{j=1}^{n} y_{j,i} \tag{A.9}$$

$$\frac{dy_{j,i}}{dt} = cby_{j+1,i}(y_{j-1,i} - y_{j+2,i}) - cy_{j,i} - \frac{h_y c}{b}x_i \tag{A.10}$$

In this thesis, the constants b and c are set to 10 so that the x variables are 10 times as fast as the y variables. The coupling coefficents h_x and h_y are both set to 1.

A.2.4 PST system

The PST system is a 5 dimensional model devoloped by Platt, Spiegel and Tresser [124] which qualitatively resembles the behaviour of solar cycles. The equations are given by

$$\begin{aligned}
\frac{dx}{dt} &= \beta_0(X - x_0)x - wy - (x^2 + y^2)x \\
\frac{dy}{dt} &= wx + \beta_0(X - x_0)y - (x^2 + y^2)xy \\
\frac{dX}{dt} &= Y \\
\frac{dY}{dt} &= aX - X^3 - bY + Z + qx^2 \\
\frac{dZ}{dt} &= -\delta(Z - cX(X^2 - 1))
\end{aligned} \tag{A.11}$$

The x and y variables represents a dynamo process in the Tachocline whilst the X, Y and Z represent a chaotic system of the solar convection zone. In this thesis, the parameter values are set to the values $a = 0.7$, $b = 0$,$c = -0.38$,$d = 0.03$,$q = 0$,$w = 2$, x_0 and $\beta_0 = 1.1$ used in [124].

A.2.5 Rössler Map

The Rössler Map [128] is a 3 dimensional system of equations, originally intended to exhibit similar behaviour to the Lorenz '63 system whilst being easier to study analytically. The system equations are given by

$$\begin{aligned} \frac{dx}{dt} &= -y - z \\ \frac{dy}{dt} &= x + ay \qquad Lorenz \\ \frac{dz}{dt} &= b + z(x - c) \end{aligned} \tag{A.12}$$

In this thesis, we set the parameters to the commonly used values $a = 0.1$ and $b = 0.1$ and $c = 14$. The system attractor with these parameter values is shown in figure A.6.

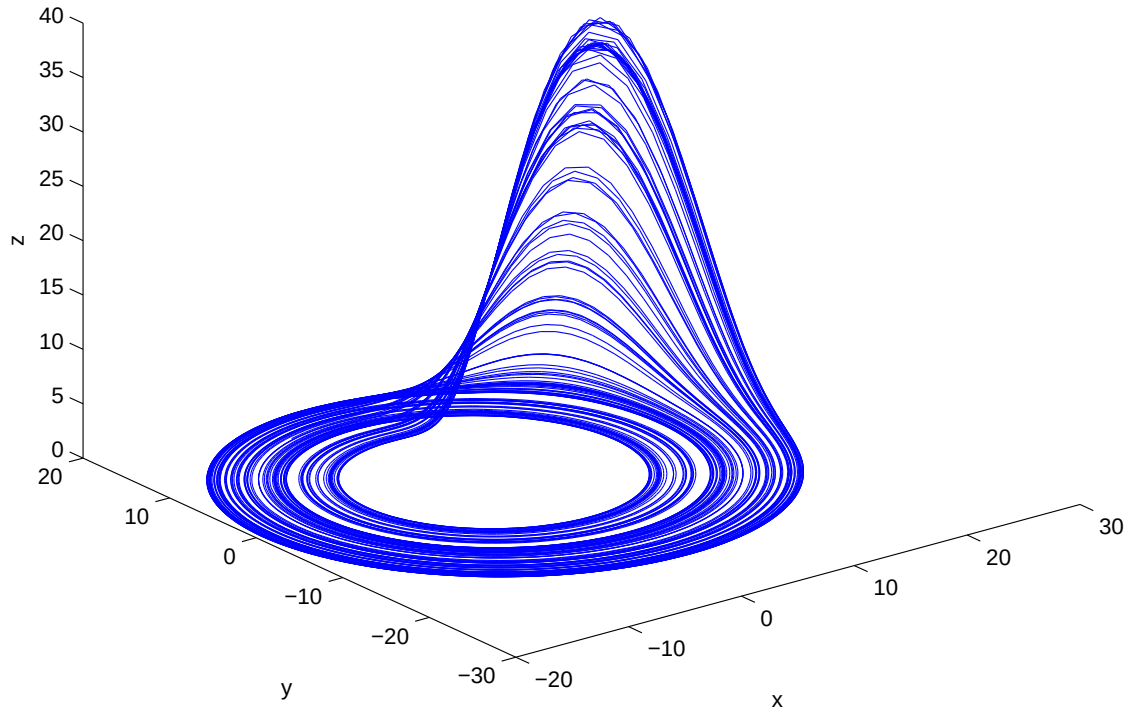

Figure A.6: The Rössler attractor with parameter values $a = 0.1$ and $b = 0.1$ and $c = 14$.

Appendix B

Details of Experiments

System	Lorenz '63
Step size	0.01
Noise model	3 σ trimmed Gaussian
Noise level	10 percent
N_{tr}	2048
N	2048
Sample rate	6 hours
Data assimilation scheme	none

Table B.1: Details of experiment 4.A

System	Duffing Map
Noise model	3 σ trimmed Gaussian
Noise level	5 percent
N_{tr}	1024
N	1024
Sample rate	6 hours
PDA stopping criterion	0.001

Table B.2: Details of experiment 4.B

System	Lorenz '63
Model	Lorenz '63 imperfect with $c = 8$
Noise model	Gaussian
Noise level	10 percent
N_{tr}	2048
N	2048
Sample rate	6 hours
Data assimilation	none

Table B.3: Details of experiment 4.C

System	Lorenz '63
Step size	0.01
Noise model	Gaussian
Noise level	5 percent
N_{tr}	2048
N	2048
Ensemble formation scheme	Inverse noise
Ensemble size	64
Data assimilation scheme	PDA
Assimilation window	1.6 days
PDA stopping criterion	0.00001
Sample rate	1

Table B.4: Details of experiment 5.A

System	Lorenz '63
Model	Perfect
Step size	0.01
Sample rate	5
Noise model	Gaussian
Noise level	5 percent
N_{tr}	2048
N	2048
Ensemble formation scheme	Inverse noise
Ensemble size	64
Data assimilation scheme	None

Table B.5: Details of experiment 6.A

System	Lorenz '63
Step size	0.01
Sample rate	5
Noise model	Gaussian
Noise level	5 percent
N_{tr}	1024
N	1024
Ensemble formation scheme	Inverse noise
Ensemble size	64
Data assimilation scheme	None

Table B.6: Details of experiment 6.B

System	Duffing map
Noise model	3 σ trimmed Gaussian
Noise level	5 percent
N_{tr}	1024
N	1024
PDA stopping criterion	0.0001
Sample rate	2

Table B.7: Details of experiment 7.A

x	0.9614
y	0.5235
X	0.9292
Y	-0.4947
Z	0.0079

Table B.8: Initial conditions for the system trajectory in experiment 7.B shown in figure 7.2 and 7.3.

System	PST system
Noise model	Gaussian
Noise level	10 percent
N	1024
Number of PDA iterations	100
Step size	0.01
Sample rate	10

Table B.9: Details of experiment 7.B

System	Duffing map
Noise model	Gaussian
Noise level	25 percent
N	1024
Number of PDA iterations	100
Sample rate	1

Table B.10: Details of experiment 7.C

System	Henon Map
Noise model	Gaussian
Noise level	1 percent
N_{tr}	512
N	512
Sample rate	1
Ensemble scheme	Inverse noise
Ensemble size	64
Data assimilation scheme	None

Table B.11: Details of experiment 8.A

System	Lorenz '96 model 2
Model	Lorenz '96 model 1
Noise model	Gaussian
Noise level	1 percent
N_{tr}	2048
N	2048
Step size	0.01
Sample rate	1
Ensemble scheme	Inverse noise
Ensemble size	64
Data assimilation scheme	None

Table B.12: Details of experiment 8.B

System	Lorenz '63
Model	Perfect
Noise model	Gaussian
Noise level	10 percent
N_{tr}	1024
N	1024
Step size	0.01
Sample rate	5
Ensemble size	32
Mismatch limit	0.005
Data assimilation scheme	None

Table B.13: Details of experiment 8.C

System	Moore-Spiegel
Noise model	Gaussian
Noise level	5 percent
N_{tr}	1024
N	1024
Step size	0.01
Sample rate	5
Ensemble formation scheme	Inverse noise
Ensemble size	64
Data assimilation scheme	None

Table B.14: Details of experiment 8.D

System	Ikeda Map
Model	Ikeda imperfect with $c = 8$
Noise model	Gaussian
Noise level	1 percent
N_{tr}	1024
N	1024
Sample rate	1
Ensemble formation scheme	Inverse noise
Ensemble size	1024
Data assimilation scheme	None

Table B.15: Details of experiment 8.E

System	Henon Map
Model	Henon imperfect with $c = 12$
Noise model	Gaussian
Noise level	1 percent
N_{tr}	1024
N	1024
Sample rate	1
Ensemble formation scheme	Inverse noise
Ensemble size	1024
Data assimilation scheme	None

Table B.16: Details of experiment 8.F

Appendix C

Details of Logarithmic Spiral

C.1 Logarithmic Spiral

A logarithmic-spiral can be described by the equations

$$\begin{aligned} x(t) &= ae^{bt}\sin(\theta) \\ y(t) &= ae^{bt}\cos(\theta) \end{aligned} \tag{C.1}$$

where a, b and θ are parameters that govern its behaviour. A set of 16 point forecasts $x_1, .., x_{16}$ and corresponding verifications $y_1, .., y_{16}$ are derived from a logarithmic-spiral with parameters $a = 1$ and $b = \frac{\pi}{18}$, sampling on the range $\theta \in (-16\pi + \frac{\pi}{4}, 7\pi - \frac{\pi}{4})$ at intervals of $\frac{3\pi}{2}$.

Bibliography

[1] 70th anniversary of the d-day landings and the role of the met office. http://www.metoffice.gov.uk/news/in-depth/D-Day-70th-anniversary. Accessed: 19/04/2015.

[2] Ensemble forecasting. http://www.metoffice.gov.uk/research/areas/data-assimilation-and-ensembles/ensemble-forecasting. Accessed: 19/04/2015.

[3] Flashback: August 1992-montana snow. http://www.krtv.com/news/flashback-august-1992-montana-snow. Accessed: 12/12/2014.

[4] Measure of skill - the anomaly correlation coefficient. http://old.ecmwf.int/products/forecasts/guide/Measure_of_skill_the_anomaly_correlation_coefficient.html. Accessed: 25/04/2015.

[5] Which city has the most unpredictable weather? http://fivethirtyeight.com/features/which-city-has-the-most-unpredictable-weather. Accessed: 12/12/2014.

[6] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974.

[7] S. Amari, N. Murata, K.-R. Muller, M. Finke, and H. H. Yang. Asymptotic statistical theory of overtraining and cross-validation. *Neural Networks, IEEE Transactions on*, 8(5):985–996, Sep 1997.

[8] american meteorological society committee on the compendium of meteorology. *Compendium of meteorology*. American Meteorological Society, 1951.

[9] M. B. Arajo and M. New. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22(1):42 – 47, 2007.

[10] J. S. Armstrong. *Principles of Forecasting. A Handbook for Researchers and Practitioners*. Kluwer Academic Publishers, 2002.

[11] G. A. Barnard. New methods of quality control. *Journal of the Royal Statistical Society. Series A (General)*, 126(2):pp. 255–258, 1963.

[12] T. Bayes and R. Price. An essay towards solving a problem in the doctrine of chances. 1763.

[13] S. Bell. *A Beginner's Guide to Uncertainty of Measurement*. Measurement good practice guide. National Physical Laboratory, 2001.

[14] A. Bellucci, R. Haarsma, S. Gualdi, P. Athanasiadis, M. Caian, C. Cassou, E. Fernandez, A. Germe, J. Jungclaus, J. Kröger, D. Matei, W. Mller, H. Pohlmann, D. Salas y Melia, E. Sanchez, D. Smith, L. Terray, K. Wyser, and S. Yang. An assessment of a multi-model ensemble of decadal climate predictions. *Climate Dynamics*, pages 1–20, 2014.

[15] J. M. Bernardo. Expected Information as Expected Utility. *The Annals of Statistics*, 7:686–690, 1979.

[16] R. Binter. *Applied Probabilistic Forecasting.* PhD thesis, London School of Economics, 2011.

[17] M. Bonavita, L. Isaksen, and E. Hólm. On the use of eda background error variances in the ecmwf 4d-var. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1540–1559, 2012.

[18] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Annals of Statistics*, 2010.

[19] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1, 1950.

[20] E. Britton, P. Fisher, and P. Whitley. The ination report projections: Understanding the fan chart. *Bank of England Quarterly Bulletin*, 1998.

[21] J. Bröcker. Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1611–1617, 2012.

[22] J. Bröcker and L. A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A*, 60(4):663–678, 2008.

[23] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting.* Springer, 2nd edition, Mar. 2002.

[24] J. G. Charney, R. Fjörtoft, and J. Von Neumann. Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254, 1950.

[25] A. Chatterjee and A. Michalak. Comparison of ensemble kalman filter and variational approaches for co_2 data assimilation. *Atmospheric Chemistry and Physics*, 13:12825–12865, 2013.

[26] W. Conover. *Practical nonparametric statistics*. Wiley series in probability and statistics. Wiley, New York, NY [u.a.], 3. ed edition, 1999.

[27] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[28] R. D. Cook. Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174, 1979.

[29] S. Corti, A. Weisheimer, T. N. Palmer, F. J. Doblas-Reyes, and L. Magnusson. Reliability of decadal predictions. *Geophysical Research Letters*, 39(21):n/a–n/a, 2012.

[30] G. D'Agostini. Probability and measurement uncertainty in physics - a bayesian primer, 1995.

[31] A. Dalcher, E. Kalnay, and R. N. Hoffman. Medium range lagged average forecasts. *Monthly Weather Review*, 116(2):402–416, 1988.

[32] T. DelSole, X. Yang, and M. K. Tippett. Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological Society*, 139(670):176–183, 2013.

[33] L. Descamps, C. Labadie, A. Joly, E. Bazile, P. Arbogast, and P. Cbron. Pearp, the mto-france short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, pages n/a–n/a, 2014.

[34] E. T. DeWeaver. *Arctic Sea Ice Decline: Introduction*, pages 1–5. American Geophysical Union, 2013.

[35] F. J. Doblas-Reyes, R. Hagedorn, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting ii. calibration and combination. *Tellus A*, 57(3):234–252, 2005.

[36] M. G. Donat and L. V. Alexander. The shifting probability distribution of global daytime and night-time temperatures. *Geophysical Research Letters*, 39(14), 2012.

[37] Y. Dong, J. Wang, C. Li, G. Yang, Q. Wang, F. Liu, J. Zhao, H. Wang, and W. Huang. Comparison and analysis of data assimilation algorithms for predicting the leaf area index of crop canopies. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(1):188–201, Feb 2013.

[38] N. R. Draper. The cambridge dictionary of statistics, fourth edition by b. s. everitt, a. skrondal. *International Statistical Review*, 79(2):273–274, 2011.

[39] H. Du. *Combining statistical methods with dynamical insight to improve non-linear estimation.* PhD thesis, London School of Economics, 2009.

[40] H. Du and L. Smith. Pseudo-Orbit Data Assimilation. Part I: The Perfect Model Scenario. *Journal of the Atmospheric Sciences*, 71:469–482, 2014.

[41] H. Du and L. Smith. Pseudo-Orbit Data Assimilation. Part II: Assimilation with Imperfect Models. *Journal of the Atmospheric Sciences*, 71:483–495, 2014.

[42] W. Ebisuzaki and E. Kalnay. Ensemble experiments with a new lagged average forecasting scheme. *Research activities in atmospheric and oceanic modeling*, pages 6.31–6.32, 1991.

[43] L. E.N. Predictability; does the flap of a butterfly's wings in brazil set off a tornado in texas?, 1972. Speech at the 139th meeting of the American Association for the Advancement of Science.

[44] E. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987, 1969.

[45] E. Epstein. Stochastic dynamic prediction. *Tellus A*, 21(6), 2011.

[46] L. Euler. *Institutionum calculi integralis*. Number Bd. 1 in Institutionum calculi integralis. imp. Acad. imp. Saènt., 1768.

[47] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

[48] D. Fairbairn, S. R. Pring, A. C. Lorenc, and I. Roulstone. A comparison of 4dvar with ensemble data assimilation methods. *Quarterly Journal of the Royal Meteorological Society*, 140(678):281–294, 2014.

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

[50] E. M. Fischer and C. Schär. Future changes in daily summer temperature variability: driving processes and role for temperature extremes. *Climate Dynamics*, 33(7-8):917–935, 2009.

[51] J. Galbraith and S. van Norden. The calibration of probabilistic economic forecasts. Departmental Working Papers 2008-05, McGill University, Department of Economics, Nov 2008.

[52] A. Germe, M. Chevallier, D. Salas y Mlia, E. Sanchez-Gomez, and C. Cassou. Interannual predictability of arctic sea ice in a global climate model: regional contrasts and temporal evolution. *Climate Dynamics*, 43(9-10):2519–2538, 2014.

[53] I. Gilmour. *Nonlinear Model Evaluation: Iota -shadowing, Probabilistic Prediction and Weather Forecasting*. University of Oxford, 1998.

[54] R. Gilmour. I. Smith, L. Buizza. Linear regime duration: Is 24 hours a long time in synoptic weather forecasting? *Journal of the Atmospheric Sciences*, 58(651):3525–3539, 2001.

[55] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[56] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[57] T. Gneiting, A. H. Westveld, A. E. Raftery, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. Technical report, Monthly Weather Review, 2005.

[58] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14:107–114, 1952.

[59] P. I. Good and J. W. Hardin. *Common Errors in Statistics: (and How to Avoid Them)*. Wiley-Interscience, July 2003.

[60] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, Apr 1993.

[61] C. Grebogi, S. M. Hammel, J. A. Yorke, and T. Sauer. Shadowing of physical trajectories in chaotic dynamics: Containment and refinement. *Phys. Rev. Lett.*, 65:1527–1530, Sep 1990.

[62] J. Gribbin and M. Gribbing. *FitzRoy: The Remarkable Story of Darwin's Captain and the Invention of the Weather Forecast.*

[63] T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560, 2000.

[64] K. Harper, L. W. Uccellini, L. Morone, and E. Kalnay. 50th anniversary of operational numerical weather prediction. *Bulletin of the American Meteorological Society*, 2(4):639–650, 2007.

[65] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., 2001.

[66] E. Hawkins and P. D. Jones. On increasing global temperatures: 75 years after callendar. *Quarterly Journal of the Royal Meteorological Society*, 139(677):1961–1963, 2013.

[67] J. M. Henderson, R. N. Hoffman, S. M. Leidner, T. Nehrkorn, and C. Grassotti. A 4d-var study on the potential of weather control and exigent weather forecasting. *Quarterly Journal of the Royal Meteorological Society*, 131(612):3037–3051, 2005.

[68] R. Hoffman, J. Henderson, S. Leidner, C. Grassotti, and T. Nehrkorn. Using 4d-var to move a simulated tropical cyclone in a mesoscale model. *Computers and Mathematics with Applications*, 52(89):1193 – 1204, 2006. Variational Data Assimilation and Optimal Control.

[69] R. N. Hoffman and E. Kalnay. Lagged average forecasting, an alternative to monte carlo forecasting. *Tellus A*, 35A(2):100–118, 1983.

[70] R. N. Hoffman and E. Kalnay. Lagged average forecasting, some operational considerations. *AIP Conference Proceedings*, 106(1):141–147, 1984.

[71] A. Hollingsworth, K. Arpe, M. Tiedtke, Capaldo, and H. M. Savijarvi. The performance of a medium-range forecast model in winter impact of physical parameterizations. *Monthly Weather Review*, 108:1736, 1980.

[72] B. A. Huberman. A model for dysfunctions in smooth pursuit eye movementa. *Annals of the New York Academy of Sciences*, 504(1):260–273, 1987.

[73] M. Hnon. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, 50(1):69–77, 1976.

[74] K. Ikeda. Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system, opt. *Comm*, pages 257–261, 1979.

[75] J.Bröcker and L. Smith. Scoring probabilistic forecasts: the importance of being proper. *Tellus A*, 22(2), 2007.

[76] H. Jeffreys. *Theory of probability*. 1939.

[77] I. Jolliffe and D. Stephenson. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 2003.

[78] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 90, 1995.

[79] K. Judd. Nonlinear state estimation, indistinguishable states, and the extended kalman filter. *Physica D: Nonlinear Phenomena*, 183(3-4):273–281, 2003.

[80] K. Judd. Forecasting with imperfect models, dynamically constrained inverse problems, and gradient descent algorithms. *Physica D: Nonlinear Phenomena*, 237(2):216–232, 2008.

[81] K. Judd, C. A. Reynolds, T. E. Rosmond, and L. A. Smith. The geometry of model error. *Journal of the Atmospheric Sciences*, 65(6):1749–1772, 2008.

[82] K. Judd, L. Smith, and A. Weisheimer. Gradient free descent: shadowing, and state estimation using limited derivative information. *Physica D*, 190:153–166, 2003.

[83] E. Kalnay and S.-C. Yang. Accelerating the spin-up of ensemble kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1644–1651, 2010.

[84] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[85] V. M. Kattsov, V. E. Ryabinin, J. E. Overland, M. C. Serreze, M. Visbeck, J. E. Walsh, W. Meier, and X. Zhang. Arctic sea-ice change: A grand challenge of climate science. *Journal of Glaciology*, pages 1115–1121, Dec. 2010.

[86] D. Kelsey. The economics of chaos or the chaos of economics. *Oxford Economic Papers*, 40(1):pp. 1–31, 1988.

[87] A. N. Kolmogorov. Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

[88] Y. Kosaka and S.-P. Xie. Recent global-warming hiatus tied to equatorial pacific surface cooling. *Nature*, 501:403–407, 2013.

[89] T. Krishnamurti, K. Rajendran, T. Kumar, S. Lord, Z. Toth, X. Zou, S. Cocke, J. Ahlquist, and I. Navon. Improved skill for the anomaly correlation of geopotential heights at 500 hpa. *Monthly Weather Review*, 131(6):1082–1102, 2003.

[90] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[91] W. Kutta. *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*. B.G Teubner, 1901.

[92] P. Laplace. Théorie analytique des probabilités. 1812.

[93] I. Law and A. Stuart. Evaluating data assimilation algorithms. *Monthly Weather Review*, 140(11):3757–3782, 2012.

[94] E. Leamer. *Specification searches: ad hoc inference with nonexperimental data*. Wiley series in probability and mathematical statistics. Wiley, 1978.

[95] C. E. Leith. Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, 102:409, 1974.

[96] M. Leutbecher and T. N. Palmer. Ensemble forecasting. *J. Comput. Phys.*, 227(7):3515–3539, Mar. 2008.

[97] R. W. Lindsay, J. Zhang, A. J. Schweiger, and M. A. Steele. Seasonal predictions of ice extent in the arctic ocean. *Journal of Geophysical Research: Oceans*, 113(C2), 2008.

[98] E. Lorenz. Predictability-a problem partly solved, 1995.

[99] E. N. Lorenz. Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, 20(2):130–141, Mar. 1963.

[100] E. N. Lorenz. *The growth of errors in prediction. In M. Ghil: in Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*. North-Holland, 1985.

[101] C. Lu, H. Yuan, B. E. Schwartz, and S. G. Benjamin. Short-range numerical weather prediction using time-lagged ensembles. *Weather and Forecasting*, 22(3):580–595, 2007.

[102] P. Lynch. The origins of computer weather prediction and climate modeling. *J. Comput. Phys.*, 227(7):3431–3444, Mar. 2008.

[103] J. E. Matheson and R. L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

[104] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467, 1976.

[105] F. Molteni, R. Buizza, T. N. Palmer, and T. Petroliagis. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119, 1996.

[106] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. Wiley series in probability and statistics. Wiley, third edition edition, 2001.

[107] D. Moore and E. Spiegel. A thermally excited nonlinear oscillator. *Astrophysical Journal*, 143:871, 1966.

[108] T. Morimoto. Markov Processes and the H-Theorem. *Journal of the Physical Society of Japan*, 18:328, Mar. 1963.

[109] P. A. Morris. Combining expert judgments: A bayesian approach. *Management Science*, 23(7):679–693, 1977.

[110] J. Moskaitis and J. Hansen. Deterministic forecasting and verication: A busted system? *Weather and Forecasting*, 2006.

[111] A. Murphy. The coefficients of correlation and determination as measures of performance in forecast verification. *Weather and Forecasting*, 10(4):681–688, 1995.

[112] A. Murphy and E. Epstein. Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117(3):572–582, 1989.

[113] A. H. Murphy. The Early History of Probability Forecasts: Some Extensions and Clarifications. *Weather and Forecasting*, 13:5–15, Mar. 1998.

[114] H. Murphy. On the ranked probability score. *Journal of Applied Meteorology*, 8(6):988–989, 1969.

[115] R. Murphy, A. Winkler. Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79(387):489–500, 1984.

[116] T. Palmer, F. Doblas-Reyes, R. Hagedorn, and A. Weisheimer. Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1463):1991–1998, 2005.

[117] T. N. Palmer, R. Buizza, M. Leutbecher, R. Hagedorn, T. Jung, M. Rodwell, F. Vitart, J. Berner, E. Hagel, A. Lawrence, F. Pappenberger, Y. Y. Park, L. von Bremen, and I. Gilmour. The ensemble prediction system - recent and ongoing developments. *ECMWF Technical Memorandum*, 2007.

[118] D. E. Parker, T. P. Legg, and C. K. Folland. A new daily central england temperature series, 17721991. *International Journal of Climatology*, 12(4):317–342, 1992.

[119] E. Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[120] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In S. Kotz and N. Johnson, editors, *Breakthroughs in Statistics*, Springer Series in Statistics, pages 11–28. Springer New York, 1992.

[121] G. Peng, L. Leslie, and Y. Shao. *Environmental Modelling and Prediction*. Springer, 2002.

[122] P. Peng, A. Kumar, H. van den Dool, and A. G. Barnston. An analysis of multimodel ensemble predictions for seasonal climate anomalies. *Journal of Geophysical Research: Atmospheres*, 107(D23):ACL 18–1–ACL 18–12, 2002.

[123] C. Pires, R. Vautard, and O. Talarand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A*, 48(1):96–121, 1996.

[124] N. Platt, E. Spiegel, and C. Tresser. Geophysical and astrophysical fluid dynamics. *Physical Review Letters*, 70(279), 1993.

[125] B. Powell, H. Arango, A. Moore, E. D. Lorenzo, R. Milliff, and D. Foley. 4dvar data assimilation in the intra-americas sea with the regional ocean modeling system (roms). *Ocean Modelling*, 23(34):130 – 145, 2008.

[126] S. Rahmstorf and D. Coumou. Increase of extreme events in a warming world. *Proceedings of the National Academy of Science*, 108:17905–17909, Nov. 2011.

[127] M. Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837, Sept. 1956.

[128] O. Rössler. An equation for continuous chaos. *Physics Letters A*, 57:397–398, 1976.

[129] M. Roulston and L. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55(1), 2011.

[130] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660, 2002.

[131] M. S. Roulston and L. A. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55(1):16–30, 2003.

[132] C.-A. S and S. von Holstein. Measurement of subjective probability. *Acta Psychologica*, 34(0):146 – 159, 1970.

[133] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[134] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[135] S. S. P. Shen, A. N. Basist, G. Li, C. Williams, and T. R. Karl. Prediction of sea surface temperature from the global historical climatology network data. *Environmetrics*, 15(3):233–249, 2004.

[136] J. Shukla, D. A. Paolino, D. M. Straus, D. DeWitt, M. Fennessy, J. L. Kinter, L. Marx, and R. Mo. Dynamical seasonal predictions with the cola atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, 126(567):2265–2291, 2000.

[137] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Apr. 1986.

[138] A. Silverstein, V. Silverstein, and L. Silverstein Nunn. *Weather and climate*. Twenty-First Century Books, 1998.

[139] L. Smith, C. Cuéllar, H. Du, and H. Judd. Exploiting dynamical coherence:a geometric approach to parameter estimation in nonlinear models. *Physics letters A*, 374(26):2618–2623, 2010.

[140] L. Smith and K. Judd. Indistinguishable states i. perfect model scenario. *Physica D*, 151:125–141, 2001.

[141] L. Smith and K. Judd. Indistinguishable states ii. imperfect model scenario. *Physica D*, 196:224–242, 2004.

[142] L. A. Smith. Predictability past, predictability present. In T. Palmer and R. Hagedorn, editors, *Predictability of Weather and Climate*, pages 217–250. Cambridge University Press, 2006. Cambridge Books Online.

[143] L. A. Smith, H. Du, E. B. Suckling, and F. Niehrster. Probabilistic skill in ensemble seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141(689):1085–1100, 2015.

[144] T. Stemler and K. Judd. A guide to using shadowing filters for forecasting and state estimation. *Physica D: Nonlinear Phenomena*, 238(14):1260 – 1273, 2009.

[145] I. Strub, P. J., O. Tossavainen, and A. Bayen. Comparison of two data assimilation algorithms for shallow water flows. *Networks and Heterogeneous Media*, 4(2):409–430, 2009.

[146] E. Suckling and L. Smith. An evaluation of decadal probability forecasts from state-of-the-art climate models. GRI Working Papers 150, Grantham Research Institute on Climate Change and the Environment, 2014.

[147] O. Talagrand and P. Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. i: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328, 1987.

[148] J. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. A series of books in physics. University Science Books, 1997.

[149] J. Taylor, P. McSharry, and R. Buizza. Wind power density forecasting using ensemble predictions and time series models. *Energy Conversion, IEEE Transactions on*, 24(3):775–782, Sept 2009.

[150] M. K. Tippett, T. DelSole, and A. G. Barnston. Reliability of regression-corrected climate forecasts. *J. Climate*, 27(9):3393–3404, Jan. 2014.

[151] Z. Toth and E. Kalnay. Ensemble forecasting at nmc: The generation of perturbations. *Bulletin of the American Meteorological Society*, 74:2317–2330, Dec. 1993.

[152] M. S. Tracton and E. Kalnay. Operational ensemble prediction at the national meteorological center: Practical aspects. *Weather Forecasting*, 8:379–398, 1993.

[153] G. J. van Oldenborgh, F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger. Decadal prediction skill in a multi-model ensemble. *Climate Dynamics*, 38(7-8):1263–1280, 2012.

[154] H. Vogel, J. Förstner, B. Vogel, T. Hanisch, B. Mühr, U. Schättler, and T. Schad. Time-lagged ensemble simulations of the dispersion of the eyjafjallajökull plume over europe with cosmo-art. *Atmospheric Chemistry & Physics*, 14:7837–7845, aug 2014.

[155] E.-J. Wagenmakers and S. Farrell. Aic model selection using akaike weights. *Psychonomic Bulletin and Review*, 11(1):192–196, 2004.

[156] R. C. Wajsowicz. Seasonal-to-interannual forecasting of tropical indian ocean sea surface temperature anomalies: Potential predictability and barriers. *Journal of Climate*, 20(13):3320–3343, 2007.

[157] D. S. Wilks. Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128:2821–2836, Oct. 2002.

[158] M. Wilson. *Structure and Method in Aristotle's Meteorologica*. Cambridge University Press, 2013. Cambridge Books Online.

[159] T. Wilson and M. Bell. Probabilistic regional population forecasts: The example of queensland, australia. *Geographical Analysis*, 39(1):1–25, 2007.

[160] R. Winkler. *An Introduction to Bayesian Inference and Decision*. International series in decision processes. Holt, Rinehart and Winston, 1972.

[161] L. Wu, V. Mallet, M. Bocquet, and B. Sportisse. A comparison study of data assimilation algorithms for ozone forecasts. *Journal of Geophysical Research: Atmospheres*, 113(D20), 2008.

[162] A. Zellner. Optimal information processing and bayes theorem. *The American Statistician*, 42(4):278–280, 1988.