The London School of Economics and Political Science

On the Provision, Reliability, and Use of Hurricane Forecasts on Various Timescales

ALEXANDER S. JARMAN

London, July 17, 2014

A thesis submitted to the Department of Statistics of the London School of Economics and Political Science for the degree of Doctor of Philosophy

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 identified in it).

I declare that my thesis consists of 91,165 words.

Abstract

Probabilistic forecasting plays a pivotal role both in the application and in the advancement of geophysical modelling. Operational techniques and modelling methodologies are examined critically in this thesis and suggestions for improvement are made; potential improvements are illustrated in low-dimensional chaotic systems of nonlinear equations.

Atlantic basin hurricane forecasting and forecast evaluation methodologies on daily to multi-annual timescales provide the primary focus of application and real world illustration. Atlantic basin hurricanes have attracted much attention from the scientific and private sector communities as well as from the general public due to their potential for devastation to life and property, and speculation on increasing trends in hurricane activity. Current approaches to modelling, prediction and forecast evaluation employed in operational hurricane forecasting are critiqued, followed by recommendations for best-practice techniques. The applicability of these insights extends far beyond the forecasting of hurricanes.

Hurricane data analysis and forecast output is based on small-number count data sourced from a small-sample historical archive; analysis benefits from specialised statistical methods which are adapted to this particular problem. The challenges and opportunities arising in hurricane statistical analysis and forecasting posed by small-number, small-sample, and, in particular, by serially dependent data are clarified. This will allow analysts and forecasters alike access to more appropriate statistical methodologies. Novel statistical forecasting techniques are introduced for seasonal hurricane prediction. In addition, a range of linear and non-linear techniques for analysis of hurricane count data are applied for the first time along with an innovative algorithmic approach for the statistical inference of regression model coefficients.

A real-time outlook for the 2013 hurricane season is presented, along with a methodology to support a running (re)analysis for National Hurricane Center 48 hour forecasts in 2013; the focus here is on if, and if so how, to improve

forecast effectiveness by "recalibrating" the raw forecasts in real time. In this case, it is revealed that recalibration does not improve forecast performance, and that, across years, it can be detrimental.

In short, a new statistical framework is proposed for evaluating and interpreting forecast reliability, forecast skill, and forecast value to provide a sound basis for constructing and utilising operational event predictions. This novel framework is then illustrated in the specific context of hurricane prediction. Proposed methods of forecast recalibration in the context of both a low-dimensional dynamical system and operational hurricane forecasting are employed to illustrate methods for improving resource allocation distinguishing, for example, scenarios where forecast recalibration is effective from those where resources would be better dedicated towards improving forecast techniques. A novel approach to robust statistical identification of the weakest links in the complex chain leading to probabilistic prediction of nonlinear systems is presented, and its application demonstrated in both numerical studies and operational systems.

Acknowledgements

Contents

List of Figures

List of Tables

List of Variables

Forecast evaluation and recalibration with dynamical systems

Forecast evaluation under serial dependence

Hurricane modelling and forecasting

Chapter 1

Introduction

Since the pioneering of modern-day weather forecasting by Robert Fitzroy in 1861 [138], the accuracy and efficiency of predictions of weather including extreme events such as tropical storms have progressed significantly. Probabilistic forecasting, in particular, has emerged as an essential tool in operational weather forecasting since the U.S. Weather Bureau began issuing subjective probabilistic forecasts of precipitation in 1965 [141]. Indeed, understanding and quantifying uncertainties about the future evolution of a complex system such as the Earth's ocean-atmosphere system is best addressed using the probabilistic approach. Probabilistic forecasts contrast with point forecasts which only provide a single value prediction of an outcome, and hence do not communicate uncertainty. Furthermore, reliable probabilistic forecast information is generally more informative for forecast users, allowing them to optimise their decision-making [191].

The benefits of probabilistic forecasting have been recognised at least as long ago as the 1940's [19]; it is now operational at forecasting centres around the world [150, 151]. Moreover, the practice of probabilistic forecasting has now become widespread, and is now commonplace in fields such as economics [22], health [130], and insurance [117, 116]. Application of probabilistic forecasting to impactful geophysical events such as Atlantic basin hurricanes has significant

1

potential value for insurance, policy-making, and civil planning [43, 117, 156]. Operating within a robust statistical framework for best-practice forecasting is therefore important to maximise the benefit of predictive information.

This chapter is structured as follows: the fundamental topics of probabilistic forecasting and Atlantic basin hurricanes are briefly introduced in Sections 1.1 and 1.2, followed by an overview of this thesis in Section 1.3.

Sections 1.4 to 1.9 provide background on many of the theoretical conventions, terminology, and notation covered in this thesis, allowing later chapters to focus on what is new. Other than the presentation itself, there is little new material in this chapter.

The sections in this chapter are summarised as follows: a review of some basic concepts of forecasting, a fundamental topic of this thesis, is given in Section 1.4. The types of dynamical systems, and forecast models that are involved in a forecasting situation are briefly described in the context of perfect and imperfect model scenarios [89, 90], followed by the definition of a forecasting framework. An explanation of probabilistic forecasting using dynamical and statistical models is given in Section 1.5. A method for producing forecasts from a dynamical model called ensemble forecasting [150, 114] is discussed along with forecast density construction [26, 81]. Statistical techniques for constructing forecasts are also briefly described, along with the concept of a climatological reference forecast which is used as a benchmark for forecast performance. Much of the work in this thesis is focuses on forecast evaluation and forecast recalibration which are discussed in Section 1.6 and Section 1.7. Measures of forecast quality such as forecast skill [142, 133], and other key attributes of forecast performance are described along with a brief introduction to recalibration techniques. Finally, a brief overview of the characteristics of Atlantic basin hurricanes, and types of hurricane data is given in Section 1.9.

1.1 Probabilistic forecast framework

Probabilistic forecasts are typically constructed from a collection, or *ensemble*, of point forecasts produced from dynamical model output (e.g. numerical weather models [114]), but can also be constructed using statistical models based on past observational data [200], or even from a hybrid of the two [205]. Whichever method is employed, either a finite set of probabilities of discrete outcomes or a probability density function for continuous outcomes is usually the resulting output information. These probability distributions express the uncertainty in the forecasts, reflecting the predictability of the future evolution of a system.

There are difficulties which are unique to probabilistic forecasts, however, in part because more sophisticated methods are necessary to produce probabilistic information, and also because they are perceived to be more challenging to communicate than point forecasts. There are, in fact, a number of stages typically involved in the process of operational forecasting, whether probabilistic or not. These are listed in a typical chronological order as:

1. data retrieval and transformation;

2. current system state estimation;

3. ensemble construction;

4. ensemble post-processing;

5. forecast evaluation.

The focus of this thesis falls on the last two stages, specifically forecast recalibration, which is typically part of the post-processing stage, and forecast evaluation. The work herein provides recommendations for best-practice forecast recalibration and evaluation in the context of hurricane prediction.

Forecast evaluation

Forecast evaluation plays a critical role not only in monitoring forecast quality, but also in increasing the effectiveness of predictive weather information for decision-making support. The evaluation stage includes the collection of *forecast-outcome* pairs; consisting of a single forecast of a given outcome in the future and the actual outcome observed. The performance of single forecast is assessed by assigning it a value according to a numerical performance measure which is a mathematical function of the forecasts and the outcomes. The purpose of the performance measure is to determine the degree of "correspondence" between the forecasts and outcomes. A common type of performance measure is called a *scoring rule* [60, 87], which is defined for individual pairs of forecasts and outcomes, and quantifies forecast accuracy, or "skill". An example of a scoring rule for point forecasts is Root Mean Square Error (RMSE) [136]. RMSE is a measure of the "distance" between a forecast and the corresponding outcome. While this is an intuitively appealing measure of performance, it is doubtful whether the true quality of a point forecast can be expressed with RMSE. The problem is that the point forecast and RMSE only hold information about the statistical expectations of the forecasts and forecast performance. An important requirement of measuring the performance of a forecast is that the information contained in the full joint distribution of forecasts and outcomes is included in the measurement [142]. This is the basis for robust forecast evaluation, and is a guiding principle behind the best-practice evaluation techniques discussed in this thesis.

Another challenge for achieving robust forecast evaluation is posed by serial dependence in outcome data. Scoring rules are typically evaluated for sequential pairs of forecasts and outcomes over some time period. For many physical systems, sufficiently short intervals between outcomes results in them becoming serially dependent. Given that performance measures quantify the degree of correspondence between forecasts and outcomes, the serial dependence in the

data can be replicated in the forecasts, and hence, the performance measure statistics. Consequently, the variances of the sampling distributions of the performance measure can become "inflated", resulting in erroneous estimates of forecast skill.

Forecast recalibration

During the post-processing stage of the forecasting process, statistical methods are commonly employed to improve the quality of forecasts that contain systematic (i.e. persistent) errors. Typically, a procedure called *calibration* is performed where biases in the mean and variance of forecast probability distributions are corrected by simply adding or multiplying by constants. A superior approach which utilises the joint forecast-outcome distribution (i.e. measures forecast skill) by, for example, using a linear regression to model past sets of forecasts on their corresponding outcomes, is *recalibration*. Recalibration can be used to improve a particular attribute of forecast performance called *reliability*. Specifically, reliability is a measure of the statistical consistency between the forecasts and the conditional expectations of the outcomes given a forecast probability. In short, it measures the "closeness" between forecast probability values and the observed relative frequency of an outcome. Forecast reliability is generally improved using recalibration where models suffer from systematic bias, making it a useful and relatively straightforward technique in the forecast post-processing stage. A number of statistical techniques are employed in this thesis to assess the effectiveness of recalibration when applied to forecast models.

1.2 Hurricane forecasting

Extreme weather events such as Atlantic basin hurricanes are responsible for some of the worlds greatest economic losses due to natural hazards. The extensive and increasing impacts on life and property [158, 117] have focused efforts

to understand the physical mechanisms of hurricanes and improve predictions of hurricane activity on various timescales [64, 50, 28, 181]. Forecast information can potentially be of significant value to policy-makers, the commercial and insurance sectors, and for the general public [43]. Skilful predictions of coastal hurricane strikes on annual to decadal timescales are of huge potential value for applications such as land use planning, hazard mitigation, emergency management, and (re)insurance pricing [149]. There is a large degree of uncertainty in hurricane predictions on climatic timescales [164], however, and the out-of-sample skill of seasonal (i.e. up to a year ahead) forecasts is still yet to be proven [46, 156]. These limitations are due to the inadequacy of numerical models for simulating the climate system, and the relatively short length of a reliable historical hurricane data archive [106]. In this thesis, a number of novel and easily deployable statistical forecast systems are presented for constructing predictions of seasonal hurricane counts (i.e. annual numbers of hurricanes). While these forecast systems are potentially skilful, the key purpose is to demonstrate robust practice in statistical forecast construction and forecast evaluation.

Hurricanes are often analysed, modelled, and predicted as count data which are typically small in value, for example, there are on average approximately 6 hurricanes forming in the Atlantic basin every year. A key forecast quantity of interest to the (re)insurance and risk management sectors is the fractions of hurricanes that make landfall over the coast of the U.S. each year, obviously because they are responsible for the worst inflicted economic losses. Since U.S. landfall hurricanes make up a subset of the total number of hurricanes forming in the Atlantic basin, the counts are usually very small (around 1.5 hurricanes a year on average). Detection of trends in these small-count count data is prohibitively difficult, despite the efforts of research within the insurance sector [34]. Robust statistical analysis and inference with small-number count data, coupled with the limited size of the historical archive, require a number of adapted techniques. These methods are demonstrated both for inference of U.S. landfall fractions and in regression modelling of various categories of hurricane

activity in this thesis.

1.3 Thesis overview

The structure of this thesis is outlined as follows:

The effectiveness of forecast recalibration is investigated under perfect and imperfect model scenarios in Chapters 2 and 3 in the context of a well known dynamical system called Lorenz63 [118]. The performance of binary forecasts (i.e. forecasts of an event with two possible outcomes) of the state of the Lorenz63 system is compared before and after recalibration. These forecasts have been constructed with different forecast density construction methods which are also defined. The results of the recalibration experiments using a number of recalibration algorithms sourced from the literature are presented. Information-theoretical measures of forecast performance are defined, and are used to assess any improvements in forecast skill, forecast reliability, and forecast resolution. It is shown that recalibration is most effective where forecast performance is already poor. The investigation of recalibration comprises a new contribution in this thesis. The concept of "optimal skill" for binary forecasts, an upper bound on forecast skill for a particular performance measure, is also introduced for the first time.

Chapter 4 examines the complicating effects of serial dependence in outcome data on forecast evaluation. These important effects, which are often neglected in operational forecast evaluation, can result in inaccurate estimates of forecast skill. This research builds on the results of Wilks [216] who has demonstrated how the variances of the sampling distributions of a scoring rule become inflated where there is linear serial correlation in sequential forecasts, resulting in overconfidence in forecast skill. Wilks [216] also derives a mathematical function to make sample size corrections necessary to obtain accurate estimates of skill. These results are replicated here, but it is also shown for the first time that serial dependence is neither a sufficient nor necessary condition for esti-

mates of forecast skill to be inaccurate. A new empirical method for sample size corrections is also proposed.

Chapters 6, 7, and 5 review and extend the current methodology for hurricane modelling and prediction, and provide insights into best practice when evaluating and utilising hurricane forecasts.

In Chapter 6, a case-study of forecast recalibration is applied to short-term (i.e. 48-hour ahead) forecasts of tropical cyclone formation issued by the National Hurricane Center based in the U.S.. Recalibration is performed both out-of-sample and with leave-one-out cross-validation to assess whether the performance of the tropical cyclone forecasts can be improved. While the latter technique increased the reliability of the forecasts, the out-of-sample approach generally led to a deterioration of reliability. This is explained by year-to-year variability in patterns of hurricane formation. It is also shown that the assessment of the reliability of these forecasts is complicated by variation in the time between forecast issuance and the occurrence of tropical cyclone formation, referred to as "Time Until Event".

In Chapter 7, Poisson and logistic regressions are used to model annual counts of various categories of hurricanes, and fractions of subset categories of hurricanes using a variety of predictor variables. Various techniques are employed to fit and select the models so that nonlinear dependencies between the response variable (i.e. hurricane variable) and the predictors, as well as collinearity between predictors, are accounted for. An innovative computational "sliding linear" root-finding algorithm for constructing confidence intervals for regression coefficients where sample sizes are small is presented for the first time.

In Chapter 5, various challenges to robust hurricane forecast construction and forecast evaluation presented by small-number count data and limited sample sizes are highlighted, and are followed by best-practice solutions to address these challenges. Two new forecast systems, based on univariate and bivariate statistical predictive techniques called "synoptic conditioning" and "conditional

analogue" are introduced. These techniques have been designed to exploit the data available in the relatively short hurricane data archive. The limitations of statistical inference with small samples of small-number count data are also discussed, followed by suggested methods which are specialised for these types of data. In addition, the relationship between forecast skill and forecast value is examined in a monetary betting scenario to demonstrate that a forecast user need not wait to establish statistical confidence in the skill of a forecast system before putting it to use. This concept is aptly titled "profit before proof".

Finally, the forecast systems based on the techniques discussed in Chapters 7 and 5 are fitted to the historical hurricane dataset, and are deployed in Chapter 8 to construct a real-time outlook for the 2013 Atlantic basin hurricane season. The skill of these forecasts is evaluated and compared to other operational predictions.

The key new contributions and innovations in this thesis are summarised as follows:

1. critique of existing recalibration algorithms for binary probabilistic forecast recalibration. Kernel density estimation [23] and beta-transform linear pool [165] algorithms are shown to perform the best out of all the algorithms.

2. examination of the relationship between forecast skill and forecast reliability in the context of recalibration using the decomposition of the ignorance score

3. surveyance of the conditions where recalibration is effective for increasing forecast reliability and forecast skill

4. introduction, discussion and quantification of "optimal skill" of binary forecasts

5. discussion of the limitations of forecast binning/categorisation for forecast

recalibration, and review and critique of binning/categorisation methods in the literature

6. identification of the conditions where recalibration has a detrimental effect on forecast resolution

7. derivation of the analytical sampling variance of ignorance score estimates for binary forecasts

8. demonstration of how misleading estimates of forecast skill can result from the presence of serial correlation in evaluation data (with both a stochastic and nonlinear dynamical system)

9. explanation of how the presence of serial correlation in evaluation data is neither a necessary nor sufficient condition for misleading estimates of forecast skill (with stochastic systems)

10. illustration of how misleading estimates of forecast skill can occur where serial correlation is not present in evaluation data but is present in forecast evaluation statistics (with a nonlinear dynamical system)

11. investigation of the *time until convergence* of score estimates to their asymptotic "true" value

12. proposal and illustration of an empirical method for effective sample size corrections where serial correlation is present in evaluation statistics

13. evaluation of NHC 2012 short term TC genesis forecasts using reliability diagrams both with consistency bars and on probability paper to quantify forecast reliability

14. recalibration of NHC 2012 short term TC genesis forecasts using simple translation out-of-sample and with leave-one-out cross-validation

15. examination of the relationship between NHC short term TC genesis forecast reliability and "Time Until Event"

16. proposal of supplementary diagrams/tables to reliability diagrams which provide additional information about the effect of time until event on forecast reliability where relevant

17. presentation of an innovative "sliding linear" root-finding algorithm for constructing confidence intervals for regression model coefficients where sample sizes are small

18. tests for overdispersion of tropical cyclone count data for Poisson and logistic regression models

19. introduction and demonstration of new "synoptic conditioning" and "conditional analogue" hurricane forecast systems

20. investigation of the limitations on statistical inference of hurricane data analysis where storm counts are small, and data are sparse

21. description of a new statistical empirical conditional analogue hurricane forecast system using temporal single and series analogues

22. introduction of a novel "top-hat" kernel dressing method designed for forecast PDF smoothing with count data

23. examination of the relationship between forecast skill and forecast value in an evaluation/betting scenario

1.4 Forecasting

Decision-makers are constantly faced with uncertainty about the future, and rely on predictions to quantify this uncertainty, and guide the planning process. Efforts to predict many physical (e.g. the motion of a planet) or non-physical (e.g. financial markets) dynamical systems evolves in time are hampered, however, due to their *nonlinear*, and sometimes *chaotic*, behaviour (as well as imperfect observational and computational capabilities). In reality, the best that

one can hope to achieve is to construct a model which imperfectly describes the underlying rules governing a dynamical system, and issue a predictive statement about the probabilities of given states of the system occurring.

The principle concern of this thesis is the prediction, or *forecasting* [6], of hurricanes which are extreme storm weather phenomena forming in the Atlantic ocean basin. Strictly speaking, weather is defined as the state of the Earth's atmosphere, which is itself considered a highly complex nonlinear dynamical system with a defined set of fixed rules, or physical laws [150]. Moreover, the ocean-atmosphere system is chaotic [118], implying that it has *sensitive dependence* on initial conditions. Sensitive dependence describes the scenario where the distance between two nearby initial states can grow rapidly and exponentially-on-average over time[1] [184]. Two forecasts of the same future state of the weather, or say, the formation or non-formation of a hurricane, can also differ significantly. Producing a useful forecast of a complex, chaotic system such as weather is a formidable task, yet, due to its direct impact on many fields, including (re)insurance, agriculture, transport, etc., weather forecast information has large potential value.

Forecasts come in two different forms: *point* forecasts and *probabilistic* forecasts. Point forecasts consist of a single "best guess" value while probabilistic forecasts aim to quantify forecast uncertainty by providing probability statements about the chances of occurrence of certain future events. Probabilistic forecasting has the crucial advantage over point forecasting in that uncertainty about the future evolution and state of a system is expressed. Unless a point forecast is accompanied by some measure of its quality, no indication of forecast uncertainty is provided.

Before an explanation of the process of forecasting, it is useful to describe

---

[1]coincidentally, this scenario is linked to hurricanes. It is sometimes referred to as the "butterfly effect", a phrase which has been credited to Edward Lorenz who used the metaphor of a butterfly flapping its wings, resulting in the eventual formation of a hurricane several weeks later [212]

the types of dynamical systems (i.e. the target objects of which predictions are produced) which are studied in this thesis.

1.4.1 Dynamical systems

A dynamical system describes the evolution of a physical or non-physical *state* in time according to a fixed behavioural rule. Let the evolution of the state $\mathbf{x}$ of a system over time in state space $\mathbb{S}$ be denoted by $\mathbf{x}_t = \Psi_t(\mathbf{x}_0)$, where Ψ represents the dynamics of the system, $\mathbf{x} \in \mathbb{S}$ where $\mathbb{S} \subset \mathbb{R}^n$, and $t \in (-\infty, \infty)$ is the time of the evolution. $\mathbf{x}_0$ denotes what is commonly referred to as the *initial conditions*. Dynamical systems are mathematically classified as either *deterministic* or *stochastic*. The evolution of deterministic systems is determined by the system's dynamics and the initial conditions (IC) without any effects of randomness (i.e. their current state defines their future state unambiguously). An example of a deterministic system is a simple pendulum [188] where the fixed rule is expressed with respect to Newton's second law as

$$\frac{d^2x}{dt^2} + b\frac{dx}{dt} + sinx = Asin\Omega t, \tag{1.1}$$

where x is the angular displacement, t is time, and $Asin\Omega t$ is a driving force.

Stochastic systems, on the other hand, are governed by a rule that has a random component, although it may also involve a fixed (non-random) component. Instead of describing a unique evolution of a state variable, its future state must be determined probabilistically. Cases of both nonlinear dynamical systems and stochastic systems are considered in this thesis. A common example of a stochastic system is a financial stock market modelled by Brownian motion [127].

The evolution of a dynamical system in time can can either be continuous or discrete. In the first case, a change in the state of the system x, is referred to as *flow*, is usually represented by a set of first order differential equations of the form

$$\frac{d\mathbf{x}(t)}{dt} = \Psi(\mathbf{x}), \tag{1.2}$$

where the dynamics Ψ are defined for all real values of time $t \in \mathbb{R}$, and $\{x_t\}_{t=0}^{T}$ forms an unbroken *trajectory* in state space. In the second case, a change in x, referred to as *map*, occurs at regular intervals, and assumes the mathematical form

$$\mathbf{x}_{t+1} = \Psi(\mathbf{x}_t), \tag{1.3}$$

where $t \in \mathbb{Z}$.

Many of the studied dynamical systems in this thesis are *nonlinear*, meaning that they have have nonlinear dynamics so that the response of the system is not directly proportional to its input [184].

Precisely determining the current state, and accurately predicting the future state, of a nonlinear dynamical system is challenging due to inherent uncertainties in the current state of the system. To deal with the problem of forecasting a system's future state, a mathematical model is usually constructed in the form of either a

(a) **dynamical model:** a mathematical description of the underlying rule(s) (e.g. a set of differential equations) governing the system to simulate its evolution, or a

(b) **statistical model:** formalising the relationships between the state variable of a system, or *predictand*, and a set of *predictor* variables based on the assumption that historically observed relationships are preserved.

A given system may be described by different models; where a physicist uses a dynamical model which incorporates differential equations, a statistician may opt for a statistical model based on regression analysis. Even under the former approach, however, there might be different sets of differential equations describing the same system. The obvious reason for having more than a single mathematical description of a system is that many physical laws are "useful approximations in restricted circumstances" [90]. In reality, no model of a physical dynamical system is a perfect description of the system at hand, simply because

forecasters do not possess a perfect knowledge of all the laws of nature. Moreover, models, like physical theories, are always unprovable, however, and can only be falsified [188]. Both dynamical and statistical models are explained in the next two sections.

1.4.2 Dynamical models

Dynamical models are constructed to describe deterministic and stochastic dynamical systems, and so the models themselves can also be categorised as deterministic and stochastic. Physicists typically use deterministic models based on a set of differential equations, whereas a stochastic model is more the domain of statisticians. Both types of dynamical model are employed in this thesis to demonstrate various properties of forecast construction, evaluation, and recalibration. Several examples of deterministic nonlinear dynamical systems are used in this thesis (e.g. Lorenz63 [118], logistic map). Constructing predictions from dynamical models is generally a complex task which involves several stages such as inputting observational data, estimating initial conditions, running the model simulation, correcting *systematic* model error, and interpreting the model output. These concepts are explained in more depth in Section 1.5.

1.4.3 Statistical models

The aim of statistical modelling of a system is to quantify the relationships between the system predictand and a set of predictors. This set-up may include univariate relationships (i.e. a describe of the relationship between the current value of the predictand and values observed in the past). The usefulness of a statistical model is reliant on both the preservation of these statistical relationships and a sufficiently large and high-quality (i.e. accurate) datasets of independent observations [125]. A common approach is linear regression analysis where a single "best guess" prediction is constructed on the basis that a given change in the value of a predictor results in a constant change in the

expected value of the predictand regardless of the value of the predictor. Non-linear relationships can also be modelled by modifying linear models or using alternative statistical techniques. The validity of regression models is based on a number of assumptions, however, which can only be justified through robust testing of the model. Producing forecasts from statistical models is typically more straightforward than dynamical models, and they are used throughout this thesis. For example, both linear and nonlinear regression models are employed in Chapter 7 to model and make predictions of long-term hurricane activity. Combined statistical-dynamical are also possible, and are used more commonly in the modern era of weather prediction (see, for example, Wilks [215], Vecchi et al. [205]).

1.4.4 Perfect and imperfect model scenario

As explained in Section 1.4.1, there is no such thing as a *perfect model* in "real-world" forecasting [90]. By considering the idealised situation of a perfect model of a dynamical system referred to as *Perfect Model Scenario*, however, it is possible to isolate and understand the effects of properties of a forecast model on its ability to produce accurate forecasts. PMS is exploited in Chapter 2 to investigate the effectiveness of forecast recalibration. *Imperfect Model Scenario* is first explained to separate limitations in practice from limitations in principle.

Imperfect model scenario

In an imperfect model scenario (IMS), the model consists of an imperfect description of the dynamics of the system, since the assumption is that a forecaster has incomplete knowledge of both the underlying rule of the system, and the exact initial conditions $\mathbf{x_0}$. The ability of a model to accurately simulate the evolution in time of a dynamical system and predict its future state is impaired by both its structural imperfections and uncertainty in initial conditions [90]. Forecasts produced from imperfect models are by definition imperfect, mean-

ing that they are subject to *forecast error* (i.e. they are inaccurate predictions of the future state of a dynamical system). IMS applies to every forecasting situation in the real world, for example, a weather forecaster can only hope to construct a crude dynamical model which contains an incomplete set of differential equations describing the ocean-atmosphere system [150].

Perfect model scenario

In a perfect model scenario (PMS), a model provides an exact description of the dynamics of the system, so forecast error is attributable solely to IC and model parameter uncertainty, and other sources of error can be ignored when evaluating the forecast model. In that sense, PMS is the opposite to IMS, but should that imply that a forecaster can issue a probabilistic perfect forecast using a perfect model? Unless the true state of the system is known, a perfect probabilistic forecast is not obtainable. Even with a perfect model at hand and access to infinite past observations of the state of the system, it is not possible to identify the initial "true" state due to uncertainty in the observations [89]. It follows that a single "best guess" prediction is not an optimal approach to accurate estimation of the initial state. Instead, an *ensemble* of initial conditions provides a better account of the uncertainty in the observations. Ensemble forecasts produced from a perfect model are assumed to have independent and identically distributed (i.i.d.) errors, and the observed outcome at a given time in the future can be considered a dynamically consistent member of the forecast ensemble, distinct from the other members only by the sampling of the initial conditions.

Given that all real world forecasting cases fall under IMS, do there exist any examples of PMS? The answer is that PMS can only be "artificially" constructed. This can be achieved by simply letting the model act as both the model and the system. For example, consider a perfect statistical model assuming the form of a standard normal distribution, $\mathcal{N}(0,1)$, which produces forecasts of a system which has state x_t determined at time t by $x_t \overset{iid}{\sim} \mathcal{N}(0,1)$.

Even though the forecast is imperfect, it is drawn from the same distribution as the system states, and will be a useful, and valuable prediction. By studying model-system configurations under PMS, we can glean important insights about the properties of a forecast model, as is the case in Chapter 2.

1.4.5 Forecasting framework

A brief overview of the forecasting process is given in this section. There are generally a number of stages involved in the process, although these may vary depending on the application and scale of the operation. Figure 1.1 shows a flowchart which describes the general framework for a forecasting process that is typically implemented at, for example, an operational weather forecasting centre.

The first three stages come under the process of *data assimilation* whereby observational information is mapped onto the model's state space (i.e. state estimation). The role of data assimilation (DA) is essentially to provide the best possible initial conditions, hence the collection and quality control of this data is an important task. In nonlinear dynamical (chaotic) systems such as weather and climate, small errors in a current state estimation lead to badly degraded forecasts due to sensitive dependence [118].

The fourth stage of the forecasting framework, and perhaps one of the most important is forecast generation. The model is *initialised* by integrating the initial conditions determined in the state estimation stage to produce a "raw" ensemble of single-value *point* forecasts for a given lead time. There are often differences between the model output and the observed state of the system, which leads to substantial forecast errors. These particular type of errors are referred to as model systematic bias. The purpose of the post-processing stage is to remove these biases using relatively simple statistical techniques. One such technique is called forecast *recalibration* which uses previous observed outcomes and forecasts.
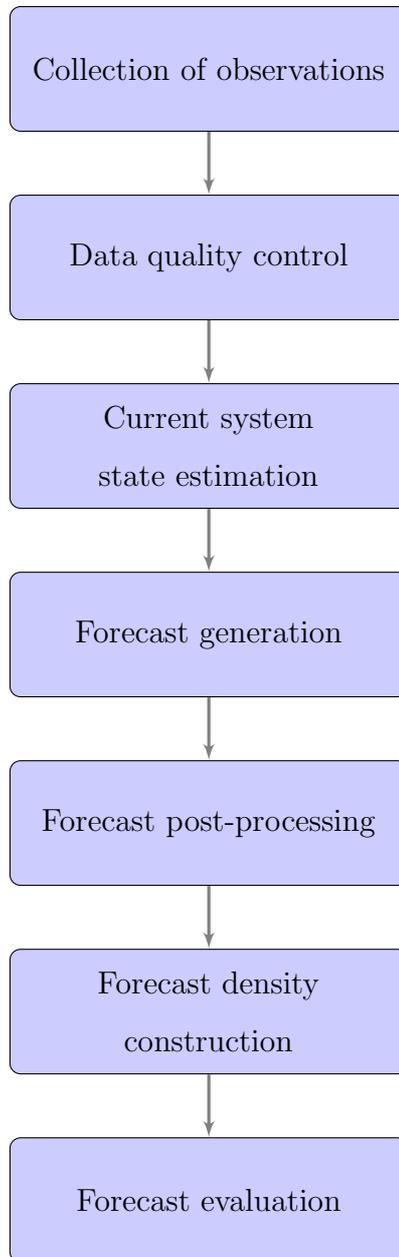
Figure 1.1: Schematic of a typical operational forecasting framework

After the post-processing stage, an ensemble needs to be converted into a usable probabilistic forecast so that probability densities are assigned to all possible future system states, this is the density construction stage. A probabilistic forecast typically comes in the form of a forecast PDF. Following this, forecast evaluation is an important stage for assessing the quality of the forecast PDF

ex ante and *ex post*. Forecast evaluation and forecast recalibration are two important aspects of the work in this thesis, and are described in more detail in Sections 1.6 and 1.7, respectively.

This very brief overview provides a simplified and perhaps incomplete summary of the forecasting framework; other processes may be involved depending on the scale of the forecast operation. For example, "downscaling" procedures are used in weather forecasting to translate the resolution of large-scale forecasts to a smaller and more practically useful scale.

1.5 Probabilistic forecasting

The uncertainty inherent in the current and future states of a nonlinear dynamical system should be reflected in a forecast of the future observable state of that system. Forecast uncertainty is best quantified by issuing probability statements about future observable outcomes based on model output and/or past observed outcomes and forecasts (i.e. *probabilistic* forecasting). Probabilistic forecasts can come in various forms: as a set of probabilities of discrete events or counts of events; as quantiles of a continuous variable; or as full discrete or continuous *probability density functions* (PDF) or *cumulative distribution functions* (CDF). There are three good reasons for producing probabilistic forecasts:

- forecast uncertainty is communicated by assigning probabilities to the possible future states of the system

- probabilistic forecasts are more accountable than point forecasts, and are more robustly evaluated in probabilistic terms

- probabilistic forecasts allow forecast-users to quantify risk, and improve their decision-making

One of the simplest approaches to constructing a probabilistic forecast is to estimate the uncertainty on a single best-guess point forecast by fitting a PDF to

a sample of past forecast errors (i.e. the difference between pairs of forecasts and observed outcomes from the past). Not only do probabilistic forecasts provide information about the likelihoods of future possible states of the system, they also allow a user of a forecast to plan their decision-making contingencies based on these likelihoods [191].

1.5.1 Sources of forecast uncertainty and error

Classification of the different types of uncertainty endemic in the forecasting of dynamical systems is useful here. Uncertainty originates broadly from two different main sources: uncertainty in the initial conditions, and uncertainties arising from the mis-specification of the forecast model, (i.e. *model error*) [150, 88]. The combination of these two sources results in the degradation of the quality of the forecast (i.e. *forecast error*). Forecast error is defined as the discrepancy between a forecast and the "truth". The truth, however, is strictly speaking, indeterminable: it is, at best, unknown and arguably undefined [88]. To determine forecast error in practice, one must take an observation of the system state as a proxy for the truth which is problematic because observations are themselves also subject to error. In real-world forecasting (i.e. under IMS), it is not possible to distinguish between forecast error that arises from IC uncertainty and model error [114]. Formal definitions of these two sources of uncertainty are now given.

Initial condition uncertainty

Accurate estimation of the initial conditions is hindered by *observation error* under IMS. Observation error occurs as a result of sampling error or systematic error in measurements (e.g. biases in a measurement device). In the context of nonlinear dynamical systems, sensitive dependence on the initial conditions implies that an arbitrarily small error in the initial state can grow at an exponential rate up until the time a forecast is required for, referred to as the forecast

lead time. The error growth will eventually result in the forecast model reaching its own limit of predictability [182], a lead time beyond which the forecast is no longer "useful". To account for IC uncertainty, *data assimilation* techniques are usually employed. Data assimilation is the process of state estimation whereby a series of observational data are "mapped" into model state space [88]. This is a key part of the operational weather prediction framework (see Fig. 1.1).

Model error

Model error arises from a forecast model's inaccurate and incomplete representation of the dynamics of the system, referred to as *model inadequacy* [96]. Model inadequacy is characterised by two limitations of the formulation of the model. These are:

1. *Structural error:* the model contains an incomplete or incorrect mathematical description of the system dynamics (i.e. there is a missing variable or a mathematical equation is incorrectly specified).

2. *Ignored sub-space inadequacy:* a component of the system's dynamics is not included in the model. This deficiency occurs where computational devices are used to model complex physical dynamical systems requiring a discretisation of model space, and imposing a limit on the resolution this discretised space [90].

Model inadequacy inevitably leads to forecast error in which the projected state of the model is different to the actual system state at a given lead time. The error growth will eventually result in the forecast model reaching its own predictability limit. Model inadequacy is discussed in more detail in Chapter 3.

1.5.2 Ensemble forecasting

Section 1.5.1 explains how uncertainty in the estimation of the initial state of a dynamical system accumulates over the integrations of a forecasting model, and results in forecast error. Model error aside, a forecaster can attempt to at least quantify the IC uncertainty. A common approach in weather and climate forecasting is to sample the initial conditions using a Monte Carlo technique [91] called *ensemble forecasting* [114]. The sample, or "ensemble", of initial states is evolved in time with the forecast model to provide a representative sample of possible future states of a system. The spread among members of the ensemble at some future time gives an estimate of the flow-dependent predictability of the system. In weather and climate forecast models, a more sophisticated technique is used by perturbing the ensemble members to better account for model uncertainty [150].

Consider a *perfect ensemble* (PE) under PMS where it is assumed that IC uncertainty exists, but where the ensemble is drawn from the same distribution of initial states as the "truth". In this context, the reason for studying the properties of a forecast model under PMS can be properly understood. The assumption of knowing the true distribution allows for sampling initial state "candidates", so that an ensemble of "perfect initial conditions" can be constructed. Although it is not possible to produce a perfect forecast by sampling a perfect ensemble under PMS, the forecast distribution is equivalent to the distribution of the future system states. As described in Section 1.5.1, it is not possible to disentangle forecast error as a result of IC uncertainty from forecast error caused by model error under IMS. Disentangling the two sources of forecast error is, of course, possible under PMS. With the enhanced configuration of a perfect ensemble under PMS, any bias in a forecast (i.e. systematic forecast error) is attributable to the evaluation method [12].

The construction of a raw ensemble forecast is now illustrated. Consider a time series of observed states of a system $s_0, \ldots, s_t, \ldots, s_N$ plus some additive

observational noise, so that s_t is defined by

$$s_t = \tilde{x}_t + \epsilon, \tag{1.4}$$

where $\tilde{x}_t$ is the true state of the system variable, and $\epsilon \overset{iid}{\sim} F(\cdot)$ is the observational noise term determined by some noise model $F(\cdot)$.

Now consider a perfect model Ψ initialised at time $t = 0$ with an ensemble of initial conditions $\mathbf{x}_0 = x_{0,1}, \ldots, x_{0,N_{ens}}$ which is used to produce forecasts of the future state $\tilde{x}_{t+\tau}$ at different lead times given by τ. These forecasts are expressed as

$$\hat{x}_{t+\tau} = \Psi(x_t) \tag{1.5}$$

$$= x_{t_\tau} + \epsilon_{t+\tau}(x_t, \epsilon). \tag{1.6}$$

The IC ensemble, which represents the degree of uncertainty about the true state of the system at time t, is generated by perturbing the observed state s_t using the inverse of the model of observational noise i.e.

$$x_{t,k} = s_t + \epsilon. \tag{1.7}$$

At each initialisation time t, the raw ensemble forecast $\hat{\mathbf{x}}_{t+\tau}$ is produced using an *iterative* approach whereby the IC ensemble $\mathbf{x_t}$ is iterated as far as the lead time t_τ using the model Ψ so that

$$\hat{\mathbf{x}}_{t+1} = \Psi(\mathbf{x_t}) \tag{1.8}$$

$$\hat{\mathbf{x}}_{t+2} = \Psi(\hat{\mathbf{x}}_{t+1}) \tag{1.9}$$

$$\vdots$$

$$\hat{\mathbf{x}}_{t+l} = \Psi(\hat{\mathbf{x}}_{t+l-1}) \tag{1.10}$$

$$\vdots$$

$$\hat{\mathbf{x}}_{t_\tau} = \Psi(\hat{\mathbf{x}}_{t_\tau - 1}). \tag{1.11}$$

where $\hat{\mathbf{x}}_{t_\tau}$ is the raw ensemble iterated τ timesteps forward to time $t = t_\tau$.

1.5.3   Statistical forecasts

Construction of statistical forecasts is typically more straightforward than dynamical model forecasts. The process of initialising a dynamical model, post-processing its output and then converting it into a forecast PDF is not required. *Classical* statistical forecasts (i.e. those constructed from purely statistical methods, e.g. linear regressions) are generally confined to forecast lead times of less than a few hours, or beyond 10 days or so due to the steady improvement of dynamical models [217]. More often, classical statistical methods are incorporated into the forecasting process to improve on aspects of dynamical model forecasts at the post-processing stage (e.g. model-output statistics [215], or kernel dressing [12]). Statistical forecasting has featured often in hurricane forecasting [64, 50, 207], however, and a number of statistical forecasting techniques are employed in this thesis to predict annual hurricane activity.

Climatology as a forecast distribution

A very simple classical statistical approach to forecasting is to use the unconditional climatological distribution, that is, the historical distribution of observed outcomes of a system variable y. Usually, it is not possible to have a complete record of historical observed outcomes so a *sample climatology* provides an estimate of the unconditional climatological distribution $E[y]$. A climatological forecast p_{clim} is constructed from the sample climatology, henceforth referred to as simply *climatology*, with either parametric [81] or non-parametric methods [16]. Given that the climatology is unconditional, p_{clim} is issued irrespective of both the time it is issued and the forecast lead time. Under the assumption that a system's dynamics do not suddenly change, and are, at least to some extent, captured by the observational data, the climatological forecast can be considered robust to unexpected outcomes. For that reason, as well as being simple and quick to construct and deploy, it is often considered a benchmark, or *reference* forecast, against which forecasts produced from alternative models

can be compared and evaluated. In short, the climatological forecast provides a measure of *zero skill*. If another forecast cannot outperform it (i.e. achieve a higher degree of skill) then that forecast cannot be considered useful. Forecast skill, and methods for evaluating it, are described in more detail in Section 1.6. The unconditional climatology is employed as a reference forecast throughout this thesis.

As a robust forecast, the climatological forecast can serve another useful purpose. At shorter lead times, a forecast produced from a more complex statistical model or a dynamical model is expected to outperform a climatological forecast (or some other reference forecast). The skill of these model forecasts may deteriorate with lead time, however, as forecast errors tend to grow. At some point the climatological forecast may demonstrate a higher degree of skill. A simple procedure called *blending* [26] is utilised to reduce the deterioration of the skill of the model forecast at longer lead times. For a given lead time, the model forecast $p_{mod}(y)$ and climatological forecast p_{clim} are blended to produce a final density given by

$$p(y) = \alpha p_{mod}(y) + (1 - \alpha)p_{clim}(y), \tag{1.12}$$

where $\alpha \in [0, 1]$ is the blending parameter. By Eqn. (1.12), a parameter value of $\alpha \to 1$ implies that the model forecast outperforms the climatological forecast, and where $\alpha \to 0$ the opposite is true. The latter scenario tends to occur at longer lead times. The blending procedure ensures that the final forecast density p always outperforms or is comparable to the climatological forecast. Several forecast models deployed in this thesis make use of the blending procedure.

1.6 Forecast evaluation

Forecast evaluation is an important process within the forecasting framework (see Section 1.4.5). Background information on the role of forecast evaluation and the various methods for assessing forecast *performance* in this thesis are

described in this section.

1.6.1 The role of forecast evaluation

The purpose of forecast evaluation is to assess the quality of forecasts, and *forecast systems*[2]. Forecast quality is usually evaluated with some numerical performance measure such as a forecast *scoring rule.* Objective evaluations of forecast quality serve a variety of administrative, scientific, and economic purposes [21]. These are summarised as follows:

(a) *to monitor and improve the quality of a forecast system.* Rates in forecast improvement can be assessed and compared for different locations and lead times. Good forecast performance can also justify funding for research and improvement of a forecast system.

(b) *to compare the relative quality of different forecast systems.* Forecast performance can be compared so that forecast users can choose between competing forecast systems.

(c) *to provide forecast-users with effective decision-making support.* The performance a forecast system needs to be communicated in simple terms, but also ideally in terms of the *value* to the forecast user.

Forecast evaluation also plays a part in the forecast density construction stage. A performance measure can be used to *calibrate* (i.e. tune) and *validate* (i.e. assess the *ex ante* quality) a forecast system before it is deployed for operational use.

1.6.2 Measures of forecast quality

Forecast quality is really a term which encapsulates a variety of attributes of forecast performance. Strictly speaking, it is the combination of the statisti-

---

[2]forecast system is a term which encompasses a forecast model together with the set of techniques used to produce a forecast

cal characteristics of the forecasts p, the outcomes y, and their relationship represented by their *joint distribution* $P(p, y)$. Many researchers of forecast evaluation stress the importance of utilising the information contained in the joint distribution of forecasts and outcomes[3] [142, 87, 217]. There are a number of performance measures available to evaluate forecast quality, but a particular type which is defined for individual pairs of forecasts and outcomes is the *scoring rule*.

A scoring rule, or score, provides a summary measure of probabilistic forecast quality, or forecast *skill*[4], by assigning a numerical score S based on the forecast density $p(\cdot)$ assigned to the outcome y. The score is usually expressed as the average forecast performance over a set of N forecast-outcome pairs, given as

$$\langle S \rangle = \frac{1}{N} \sum_{i=1}^{N} S(p_i(y), Y_i), \tag{1.13}$$

where $p_i(y)$ is the ith forecast density assigned to the ith outcome Y_i. Forecasts cannot be robustly and meaningfully evaluated on the basis of a single forecast-outcome pair, however, so access to a large forecast sample of forecast-outcome pairs is important [182, 25]. There are two key scoring rules which are employed in this thesis; these are the Brier score and ignorance, and are defined below.

Brier score

The Brier Score [20] is the most commonly defined for binary event (i.e. $Y = 0$ or $Y = 1$), and is given by

$$S(p, Y) = (p - Y)^2, \tag{1.14}$$

where $p = P(Y = 1)$. The Brier Score is essentially the *mean-squared error* of the forecasts over a set of N forecast-outcome pairs.

---

[3]outcomes are sometimes referred to as observations or verifications in the literature

[4]skill is sometimes referred to as a measure of relative performance of two forecast systems, but it is consistently used here to define the absolute performance of a single forecast system

Ignorance

Ignorance is an *information-theoretical* [172, 113] scoring rule which measures the *information deficit* of a probabilistic forecast (i.e. the information that it lacks according to an optimal encoding scheme) before it is evaluated with the outcome. Ignorance is defined for binary, categorical, or continuous outcomes, and is expressed as

$$S(p(y), Y) = -\log_2(p(Y)).$$ (1.15)

Both of the above scores are negatively oriented, meaning that a smaller value of the score indicates a more *skilful* forecast. Ignorance and the Brier score are frequently used in the literature and share an ideal property of scoring rules called *propriety* [25]. In mathematical terms, a scoring rule is classified as proper if for any two probability densities p and q, the following inequality holds:

$$\int S(p(y))q(y)\mathrm{d}y \geq \int S(q(y))q(y)\mathrm{d}y.$$ (1.16)

The minimum (i.e. optimum) score, therefore, is obtained over all possible values of $p(y)$ if $p(y) = q(y)$. Furthermore, a scoring rule is *strictly* proper if the strict inequality $>$ holds. An interpretation of Eqn. (1.16) is that a proper score rewards a forecast density p that is as close as possible to the "true" density q. In short, a proper score encourages a forecaster to issue a forecast that reflects their attempt to achieve the most accurate forecast. For example, in the unlikely event that a forecaster needs to choose between a perfect forecast and an imperfect forecast, the former would be preferred by a proper scoring rule, at least over a sufficient number of forecast-outcome pair evaluations. In other words, if a perfect model is available, it should always be chosen over an imperfect forecast model.

While propriety may appear to be a minimum requirement for a scoring rule, there do exist scores which are *improper* (i.e. do not possess propriety). Bröcker and Smith [25] demonstrate that the *mean-square error* (MSE) score,

defined as

$$S(p(y), Y) = \int (Y - y)^2 p(y) \mathrm{d}y, \tag{1.17}$$

is improper. Another example of an improper scoring rule is the Pearson linear correlation coefficient for point forecasts which measures the *linear association* (i.e. the overall strength of the relationship between forecasts and outcomes) rather forecast accuracy as is done by the Brier and ignorance scores. The Pearson linear correlation coefficient is expressed as

$$S(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}. \tag{1.18}$$

where X denotes a point forecast (which could be a mean or median probabilistic forecast), and Y is the outcome. Evaluating forecasts with the Pearson linear correlation coefficient thus may lead a forecaster to choose an imperfect model over a perfect one. The impropriety of the Pearson linear correlation coefficient is demonstrated in Chapter 5.

It is important to note that use of a proper score does not guarantee that the best forecast model is preferred. For example, there may be cases where perfect ensemble forecasts are produced from a perfect model but the density construction method may be deficient, resulting in biased forecast densities. A proper score will then penalise the model for forecast error, failing to identify it as perfect.

Another property of the ignorance score is called *locality* [25], which implies that the score is only dependent on the forecast probability assigned to the actual outcome alone. An example of a non-local score is the MSE score. There are a number of other desirable properties of scoring rules, depending on the type of outcome variable (i.e. binary, categorical/discrete,continuous) [137], but propriety is considered a minimum requirement for robust forecast evaluation here.

As previously discussed in Section 1.5.3, a reference forecast is used to define zero skill. The arithmetic difference between the ignorance scores of a forecast system and a reference forecast is defined to measure relative skill. A climatological forecast is regularly used as the reference forecast throughout this thesis.

1.6.3 Imperfect forecast error

An example of how forecast error can arise from imperfect forecasts under either PMS or IMS is now illustrated. Consider a simple statistical forecast model of a "toy" system which has two states (i.e. the outcome $Y \in \{0, 1\}$ is binary). Suppose that a forecaster wishes to predict a series of system state outcomes using the forecast model which is defined by a PDF denoted p. Let the true PDF[5] be denoted by q, which belongs to a different class of distribution to p. Suppose also that q is a convex linear combination of p and a PDF g where $g \neq p$, so that

$$q = \alpha p + (1 - \alpha)g, \qquad (1.19)$$

where $0 \leq \alpha < 1$. Clearly, for all possible values of α, the forecast PDF is imperfect, but as $\alpha \to 1$ and the forecast PDF converges to the true PDF (i.e. $p \to q$), the forecast skill measured by some scoring rule would converge to the perfect score (i.e. $S(p(y)) - S(q(y)) = 0$).

Suppose that the forecaster issues probabilistic forecasts P_i of the binary outcome $Y_i = 1$ on $i = 1, \ldots, N$ occasions where the forecast is determined by an independent and identically-distributed (i.i.d.) random draw from a uniform distribution

$$P_i \sim \mathcal{U}(0, 1), \qquad (1.20)$$

and the true probability is

$$Q_i = \alpha P_i + (1 - \alpha)\rho, \qquad (1.21)$$

---

[5]this PDF purely represents the distribution of the two true states of the system in this case and is not subject to observational uncertainty - see Section 1.5.1

where $\rho \sim \mathcal{U}(0.5, 1)$. Hence, a forecast P_i is almost certainly different from the true probability Q_i, and hence imperfect, on every occasion i.

Figure 1.2 shows the results of evaluation of forecast performance using the ignorance score. The top plot shows the expected relative ignorance score $E[IGN]$ which measures the difference between the expected skill of p and q, expressed as

$$E[IGN] = \sum_{j=1}^{2} q_j \log_2 \left(\frac{q_j}{p_j} \right), \tag{1.22}$$

where q_j and p_j denote the true and forecast probabilities assigned to the jth outcome, respectively. The curve converges to the perfect score of 0 as $\alpha \to 1$. A perfect forecast model would achieve a perfect expected score. The bottom plot of *empirical* relative ignorance estimated over the $N = 2^{11}$ forecast-outcome pairs (see Eqn. (1.15)), however, shows that zero forecast error is not achievable. Even where $\alpha = 1$, so that the forecast and true PDFs are identical, the forecast model has an ignorance score value of 0.71 ± 0.03 bits simply because it produces incorrect probabilistic forecasts.

1.6.4   Forecast reliability

Forecast reliability is another key attribute of probabilistic forecast performance. Reliability describes the statistical consistency between the forecasts p and the conditional expectations of the outcomes given a forecast probability $E[y|p]$. A forecast system is considered reliable if, for a given forecast, the observed frequency of the predicted event is consistent with the forecast probability. In notational terms this is written as

$$P(Y = 1 | p = r) = r, \tag{1.23}$$

where p is the forecast probability, r is the realisation of p, and $Y = 1$ is the occurrence of an event. The left-hand side of Eqn. (1.23) is often referred to as the calibration function [217, 23], which is henceforth denoted as κ.

Forecast relative expected ignorance

Forecast ignorance
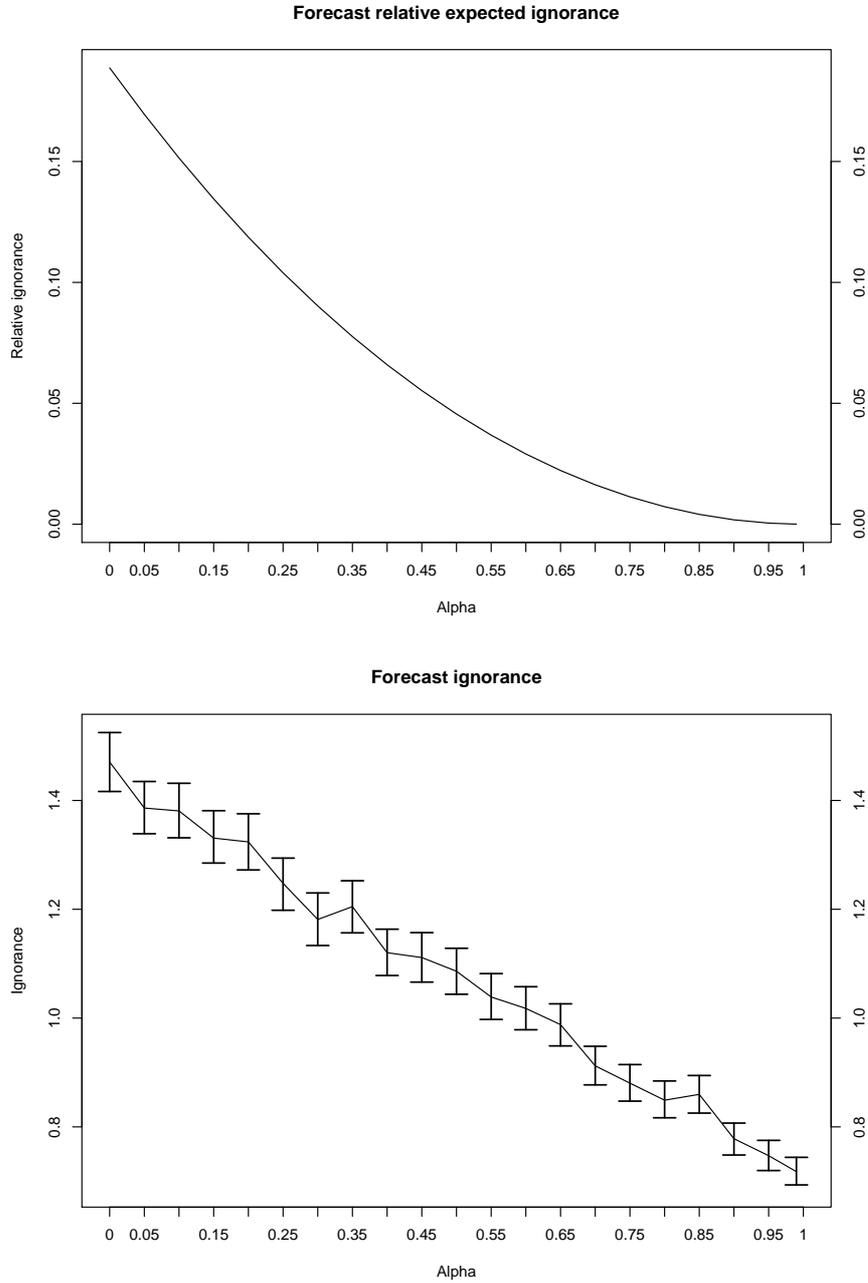
Figure 1.2: Skill of imperfect forecasts: the relative expected ignorance of the probabilistic binary forecasts drawn from the PDF p with respect to the true PDF q with increasing convergence α between the two PDFs (top), and the empirical ignorance score of the probabilistic forecast P_i with respect to the true probability Q_i for a series of $N = 2^{11}$ binary outcomes with 95% likelihood intervals (bottom). Forecast error results in larger values of relative empirical ignorance compared to relative expected ignorance even where the forecast PDF is perfect (i.e. $\alpha = 1$)

The calibration function κ is an important aspect of the examination of forecast recalibration presented in Chapters 2 and 3. Clearly, a forecast system would be perfectly reliable if all forecasts were issued as $\kappa(p)$, rather than p. Unfortunately, κ is generally unknown and must be estimated if forecasts are to be recalibrated. As with any estimation problem, this leads to residual errors (i.e. bias and variance) in the estimate of κ.

Reliability alone is a necessary but not sufficient condition for forecast quality. A forecast system would be reliable if the issued forecast probability was always equal to the climatological frequency. Eqn. (1.23) above would be satisfied in this case but there would be no predictive information provided about more extreme observed frequencies. Reliability therefore does not imply whether a forecast system is able to distinguish between system states that result in different observed events. This attribute is referred to as forecast resolution which is described in Section 1.6.6.

Reliability Diagrams

Reliability diagrams are employed throughout this thesis to assess the reliability of the binary probabilistic forecasts because they provide a quick and simple graphical representation of the overall performance of a forecast system [140, 24]. The format used here is based on that of Bröcker and Smith [24], which includes 5% to 95% consistency bars and the transposition of the reliability diagram onto probability paper with Bonferroni corrected[6] levels . These additional aspects quantitatively express the reliability of a set of forecasts by comparing them with the expected variance of observed relative frequencies given a theoretically reliable forecast system. The variance of observed relative frequencies given such a forecast system is attributable to the sampling effects of small-number counting statistics alone.

Reliability diagrams are constructed to assess the calibration or reliability of a set of forecasts $X_i \in [0, 1]$, $i = 1, ..., N$ given a corresponding set of outcomes

---

[6]these account for multiple hypothesis tests - see Bonferroni [14]

$Y_i \in \{0, 1\}$. Strictly speaking, the forecasts are reliable for all i for which X_i falls into a small interval B if the mean value of X_i over that interval is equal to the relative frequency of the event $Y_i = 1$. The forecasts are first grouped into bins B_k, $k = 1, ..., K$, which are ideally of equal width, or at least equally populated if the forecast values are non-uniform. The observed relative frequencies, f_k, are then given by

$$f_k = \frac{\sum\limits_{i \in I_k} Y_i}{\# I_k},$$ (1.24)

and averaged forecast values, r_k, given by

$$r_k = \frac{\sum\limits_{i \in I_k} X_i}{\# I_k},$$ (1.25)

where $\# I_k$ is the number of elements in a collection of indices $I_k := \{i; X_i \in B_k\}$ for which X_i falls into bin B_k. The diagram provides a measure of the correspondence between the forecasts and outcomes by showing the observed relative frequencies of the events, f_k, plotted against the averages of the binned forecast probability values, r_k for each bin, B_k. Collectively, these plotted points represent the estimate of the calibration function 1.23 for the forecast system, defined as

$$\hat{\kappa}(p) = P(Y = 1 | p = r_k)$$
$$= f_k,$$ (1.26)

which is important in forecast recalibration [23, 217]. The true calibration function κ of a perfectly reliable forecast is equal to the diagonal on the diagram. Even for such perfectly reliable forecasts, however, unbiased calibration function estimates $\hat{\kappa}(p)$ may exhibit deviations from the diagonal due to variance. It is important to note that the variance depends on the distribution of the forecast p. Hence, certain deviations of the observed relative frequencies from the diagonal might be typically exhibited by one (reliable) forecast system, but not by another.

Reliability Diagram
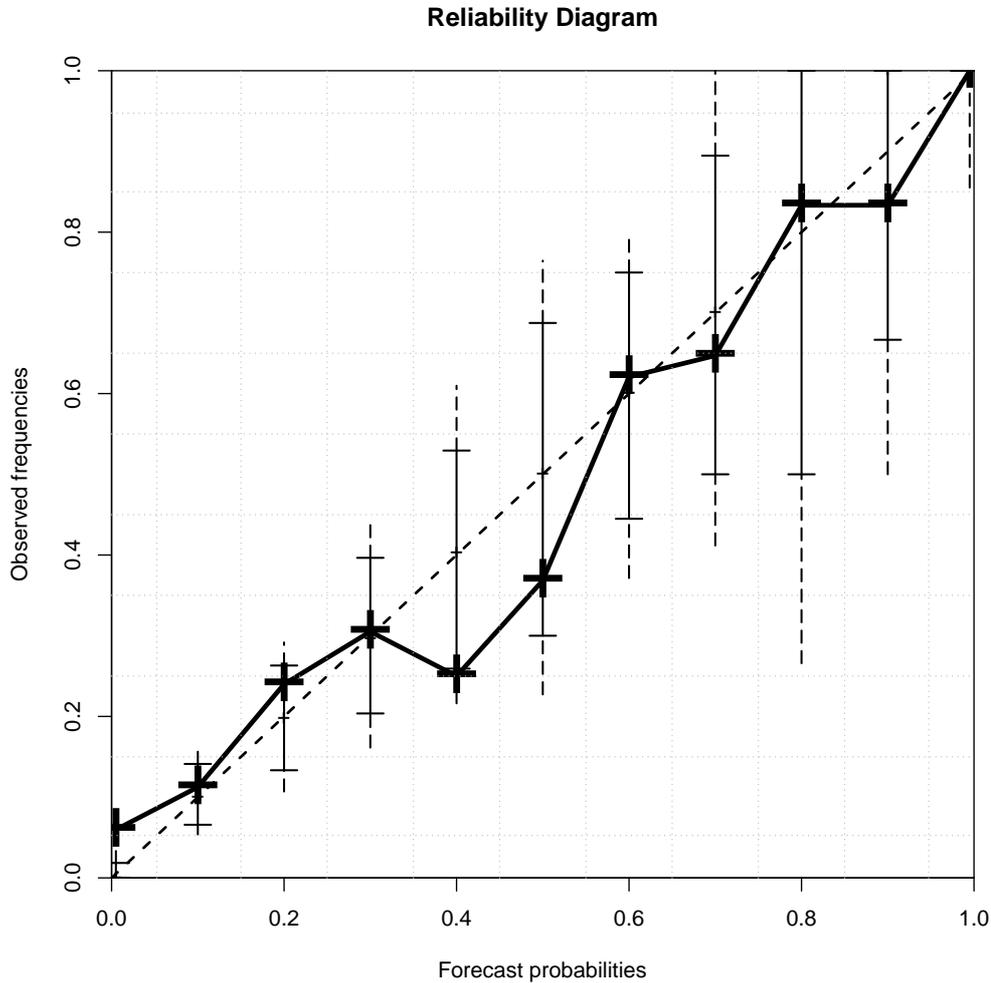
Figure 1.3: Forecast reliability: an example of a reliability diagram with 5% - 95% (1% - 99% vertical dashed line) consistency bars. All but three forecast categories lie within the consistency bars, indicating that the forecast system is mostly reliable. The forecast probability bin boundaries (grey dotted lines) have been determined by taking the mid-points between each probability category value.

Consistency Resampling

Several methods exist in the literature for presenting information about the variations expected of the observed relative frequencies consistent with a theoretically reliable forecast system on the reliability diagram. A common approach is to display the distributions of the forecast values X_i on additional inset plots

in the form of refinement distributions Wilks [217], or histograms [7]. Otherwise, bin populations can be communicated through the size of the plotted symbol [66], or by printing the population value adjacent to it [140]. These graphical methods help to convey the reliability expected of each forecast bin given the bin population, $\#I_k$ although they do not provide a direct quantification of the sampling error.

Bröcker and Smith [24] present a methodology for visualising the sampling error on a reliability diagram called *consistency resampling*. The range of variation expected of the observed relative frequencies conditioned on a set of reliable forecasts is computed through a bootstrapping method which accounts for uncertainties in both the bin forecast averages r_k and bin populations $\#I_k$. A simplified approach can be adopted if these two quantities are assumed to be fixed in each bin. The observed relative frequencies observe a binomial distribution with parameters r_k and $\#I_k$ in such a case. Consistency bars are constructed with the 5% to 95% quantiles of a set of outcomes $\hat{Y}_i$, $i \in I_k$ resampled from the binomial distribution i.e. $\hat{Y}_i \sim \mathcal{B}(\#I_k, r_k)$. If the observed relative frequencies fall within these *consistency bars* at each bin then the forecasts are calibrated to within 5% to 95% consistency. This is more informative than just measuring the distance between the point and the diagonal, which does not convey how consistent any deviations in the forecasts are with sampling error. Under the assumption of fixed parameters, consistency bars could be constructed with a variety of confidence intervals often used in categorical data analysis such as the *Wald*, or *inverted score-test* confidence intervals [2, 180].

The reliability diagram on probability paper serves to display the reliability of the forecasts with respect to the consistency bars. The y-axis represents the distance measured in probability from the 50% quantile of the consistency bar, and not the diagonal itself (although the difference between the two is generally minimal). For example, if, for a given bin, the observed relative frequency lies on the upper limit of the consistency bar then the corresponding point will lie on the 0.9 dashed line on probability paper since there is a 90% chance that
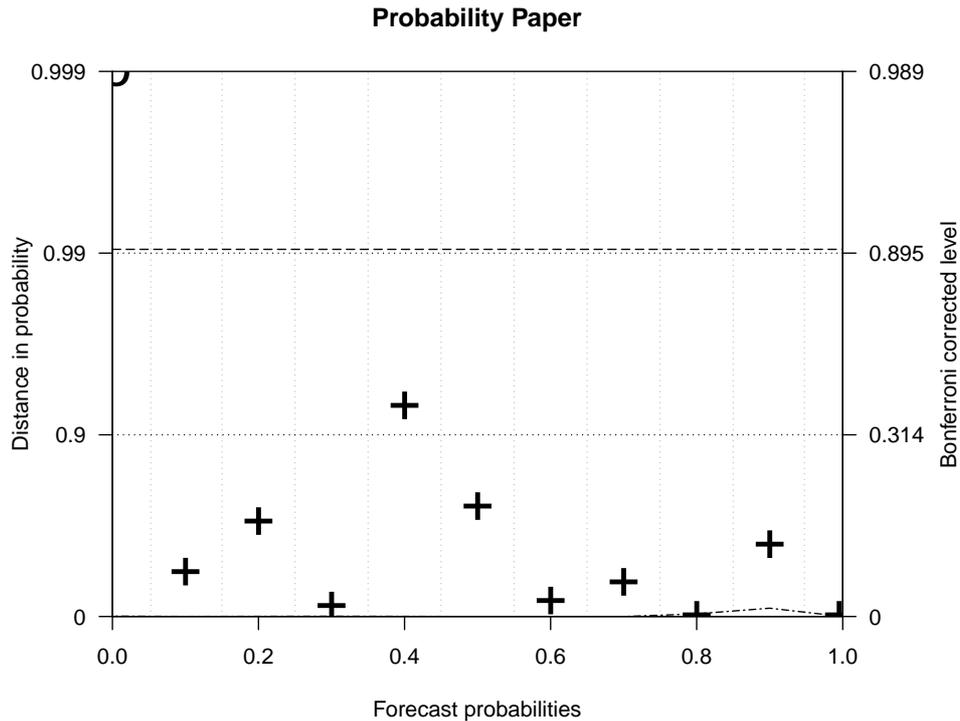
Figure 1.4: Forecast reliability on probability paper: an example of a reliability diagram on probability paper, corresponding to Fig. 1.3, showing that all but three forecast categories are consistent with forecast reliability. The dash–dotted line denotes the exact position of the diagonal. The right-hand axis indicates the equivalent Bonferroni corrected levels i.e. for a reliable forecast, all of the points (11 categories) would be expected to fall within the 0.99 probability distance band with an 88.6% chance. In addition, the dashed lines indicate where the entire diagram would be expected to fall within with a 90% chance. The forecast probability bin boundaries (grey dotted lines) have been determined by taking the mid-points between each probability category value.

any observed relative frequency would lie within the range of the consistency bar if the forecast was reliable. Bröcker and Smith [24] include the Bonferroni corrected probability distances on a secondary y-axis in the reliability diagram on probability paper. This helps to convey the overall reliability of the forecast system by denoting the chance of all of the forecast probability bins falling within a certain range, under the assumption of reliability, rather than just a single bin. For example, Fig. 1.4 shows that there is a 31.4% chance that all of the 11 bins fall within the 0.9 band indicated by the left hand axis if the forecast system is in fact reliable.

1.6.5 ROC curves

Relative Operating Characteristic (ROC) curves are, like reliability diagrams, a visual representation of forecast performance. Unlike reliability diagrams, however, they do not include the full information contained in the joint distribution of forecasts and outcomes. A ROC curve [124] consists of a plot of the *Hit Rate*, HR (i.e. the relative frequency of occurrences of the binary event $Y = 1$ that have been "successfully" forecast), against the *False Alarm Rate*, FAR (i.e. the fraction of "erroneously" forecast occurrences of the binary event). The definitions of "successful" and "erroneous" are justified in the probabilistic forecast setting by defining that the occurrence of the event is forecast when $p \geq p_k$ for $k = 1, \ldots, K$ forecast bins. These bins are determined in the same way as for reliability diagrams (see Section 1.6.4). Hence, HR and FAR are computed with respect to each forecast probability bin B_k as

$$HR(p_k) = \frac{1}{f} \int_{p_k}^{1} f_k \frac{\#I_k}{N} \mathrm{d}p, \tag{1.27}$$

and

$$FAR(p_k) = \frac{1}{1-f} \int_{p_k}^{1} (1 - f_k) \frac{\#I_k}{N} \mathrm{d}p, \tag{1.28}$$

where f is the overall relative frequency that the event occurs over the entire training set of forecast-outcomes of size N.

To fit the ROC curve, a *bi-normal model* is employed where a random binary *decision variable* ξ is defined for each outcome, and whose variations reflect the uncertainty of a binary event. There is supporting empirical evidence from a number of studies that suggests the bi-normal model performs well for fitting ROC curves [8]. A signal distribution $g_s(\xi)$ defines the *a posteriori* distribution of ξ given an event occurrence, while a noise distribution $g_n(\xi)$ defines the *a posteriori* distribution of ξ given an event non-occurrence. Under the assumption that g_s and g_n are both normal (i.e. *bi-normal*), HR and FAR can be formulated as integrations of the standard normal distribution, g, above a critical value of the decision variable ξ_c so that

$$HR(\xi_c) = \int\limits_{z_s(\xi_c)}^{\infty} g(x)\mathrm{d}x, \qquad (1.29)$$

and

$$FAR(\xi_c) = \int\limits_{z_n(\xi_c)}^{\infty} g(x)\mathrm{d}x, \qquad (1.30)$$

where $z_s(\xi_c) = z_{HR} = (\xi_c - \mu_s)/\sigma_s$ and $z_n(\xi_c) = z_{FAR} = (\xi_c - \mu_n)/\sigma_n$ are the *standardised normal deviates* of HR and FAR respectively, and μ_s and σ_s (μ_n and σ_n) are the mean and standard deviation of g_s (g_n)[7]. ROC curves are the basis for a forecast recalibration algorithm used in this thesis which is fully explained in Chapter 2.

1.6.6 Forecast resolution

The final attribute of forecast quality which is analysed in this thesis is forecast resolution [142, 217]. Resolution is, like forecast reliability, a key attribute of

---

[7]the notation for the bi-normal model is reproduced *ad verbatim* from Atger [8]

forecast performance [196, 8]. It refers to the differences between the conditional expectation of the outcomes given a forecast probability $E[y|p]$ and the marginal (unconditional) expectation of the event $E(y)$ (i.e. the unconditional climatology) [142]. Qualitatively, it can be described as the degree to which a forecast system is able to discriminate between observed events that are different from each other. In the case where a forecast is completely lacking in resolution, there is no difference between the conditional expectation of the outcomes and unconditional expectation of the event (i.e. $E[y|p] = E(y)$). Hence, larger differences indicate better forecast resolution. For example, if the outcomes following from two successive average daily temperature forecasts over London of, say, 10°C and 20°C are very different, the forecasts have resolved the two outcomes and demonstrate high resolution. On the other hand if those two outcomes were very similar, low forecast resolution would be indicated. Another attribute which is sometimes referred to in this thesis and closely relates to resolution is called *sharpness*. Sharpness is an attribute of the forecasts alone, and is a measure of the concentration of the forecast PDF [59, 142].

1.6.7 Forecast value

Recall from Section 1.6.1 that the forecast evaluation plays a role in communicating the value of a forecast to a forecast user. Forecast value is considered another aspect of forecast performance. Measuring the value, or *utility*, of a forecast is inherently a multi-disciplinary task (e.g. economic, societal, or otherwise [94]), and is thus not restricted directly to monetary worth. Nevertheless, most studies of value have focused on economic value since it is relatively straightforward way to communicate forecast value [171, 67]. Forecast skill has been considered to be intrinsically linked to forecast value [217], but the relationship has been shown to be complex [171, 193], and even inversely related [139]. The problem is normally addressed with a simple *decision-analytic* model called a *cost-loss* problem [171, 169, 193]. Forecast value is examined with respect to

monetary profit in a betting scenario [67] in Chapter 5, and its relationship with forecast skill is also investigated.

1.7 Forecast recalibration

Since models produce predictions of the future state of model variables rather than actual state of the real-world system, resulting in systematic model error, probability forecasts need to be calibrated as an integral part of the post-processing stage (see Section 1.4.5) before the final forecast PDF is produced [191]. A simple statistical approach for improving forecast skill is to recalibrate forecasts.

Recalibration is the process of making statistical corrections to a probabilistic forecast system using information about the joint distribution of forecasts and outcomes. This information could be sourced from previous forecast PDFs or from the observed climatological distribution for example. The reliability of forecasts can be improved through recalibration, although generally resolution cannot be improved. A method such as combining forecasts with other forecasts that have better resolution can lead to improved resolution however [191]. A common technique used in ensemble post-processing is called *Model Output Statistics* (MOS) [215, 218] which employs statistical methods such as linear regression.

For clarity, a more precise definition applicable to this thesis now follows: *Forecast recalibration is defined as the process of calibrating binary forecast probabilities $p \in [0,1]$ using a sample of independent binary forecast $p \in [0,1]$ and binary outcome $Y \in \{0,1\}$ pairs.* Figure 1.5 shows a schematic of the typical recalibration process followed in this thesis.
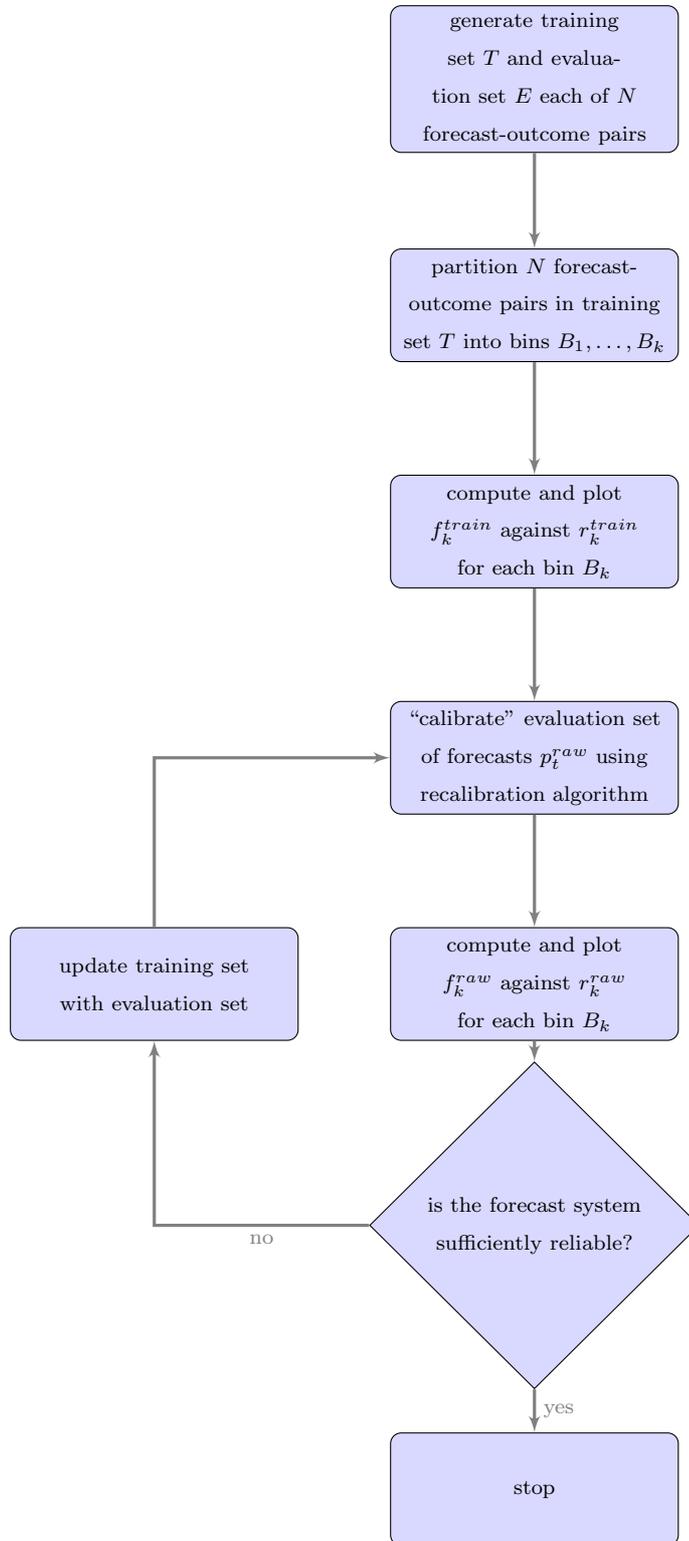
Figure 1.5: Schematic flowchart of forecast recalibration procedure

1.8 Forecast density construction methods

To convert a raw model ensemble forecast into a probabilistic forecast, some kind of density construction method is required. Four different methods have been employed in this thesis to construct forecast PDFs from raw model ensemble forecasts for *binary events*. A binary event is an event with two outcomes, e.g. yes/no, rain/no rain etc. From this point onwards, all outcomes, binary or otherwise, are denoted by the variable y. An example of one of the methods is described below.

Kernel Dressing and Blending (KDB) Method

Kernel dressing is a flexible, nonparametric method for translating an ensemble of model integrations into a forecast PDF by replacing the ensemble members with kernel functions. The approach is similar to kernel density estimation (KDE) [179] where each ensemble member is "dressed" with its own statistical error distribution belonging to some continuous class of distributions [179, 173, 26]. A density forecast is constructed by dressing the ensemble members with Gaussian "bumps" called kernels to obtain a continuous PDF. A standard kernel dressing approach is to transform the ensemble $\mathbf{x} = x_1, \ldots, x_m$ into a PDF $(y|\mathbf{x}, \sigma)$ by assigning a linear combination of kernels centred on each ensemble member x_j. The kernel dressed PDF is defined as

$$\hat{p}_\sigma(y|\mathbf{x}, \sigma) = \frac{1}{N_{ens}\sigma} \sum_{j=1}^{N_{ens}} K\left(\frac{y - x_j}{\sigma}\right), \tag{1.31}$$

where σ is the strictly positive bandwidth or smoothing parameter, and the kernel K is represented by a standard Gaussian density

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}. \tag{1.32}$$

Ideally, the optimal bandwidth is selected so that the divergence of the estimate $\hat{p}$ from the true PDF q, assuming it exists, is minimised, that is

$d(\hat{p}, q) = ||\hat{p} - q||$ where $d(\hat{p}, q)$ is some measure of the divergence. Obviously, measuring the divergence is not possible since q is unknown. The best alternative is to deploy an automated selection method such as K-fold cross-validation or "plug-in" selection [69]. K-fold cross-validation is useful method for fitting and validating a model where datasets are limited in size [155, 73]. The data is split into K roughly equal sized parts which are, in turn, used to validate the model which has been fitted with the other $K - 1$ parts. Leave-one-out cross-validation (i.e. K-fold cross-validation (CV) with $K = N$) is particularly preferred where datasets are limited in size which is case with hurricane data in Chapters 6 and 8. Where larger synthetic datasets are available, such as for those used in Chapters 5, 2-fold cross-validation is performed.
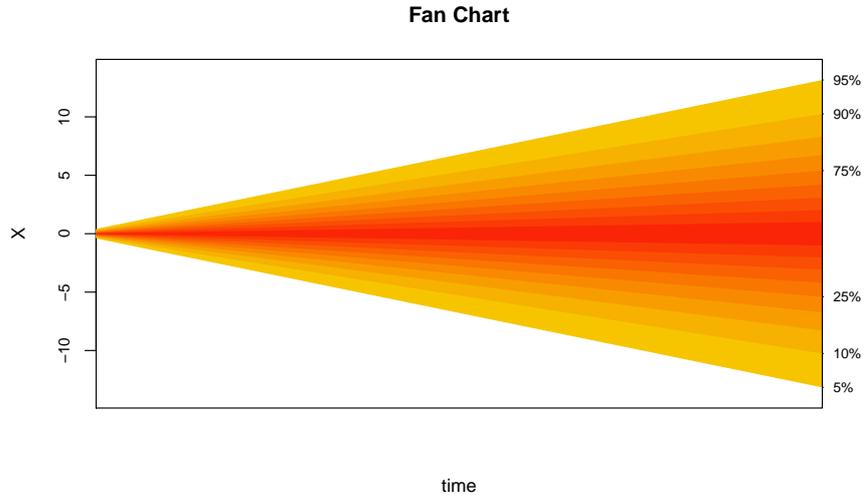
Figure 1.6: Evolution of a forecast PDF: a schematic of a fan chart for a forecast PDF evolved in time. The right-hand axis labels the percentiles of the PDF. Darker shades represent more probable system states. The increase in spread is evident with time reflecting the increase in uncertainty. This type of plot is used in several sections of thesis.

The optimised kernel width $\hat{\sigma}$ of the forecast PDF is obtained by minimising some cost function, ideally a proper probabilistic forecast scoring rule (formally

defined in Section 1.6) according to

$$(\hat{\sigma}) := \arg\min_{\sigma} -\frac{1}{N}\sum_{t=1}^{N} S(\hat{p}, Y_i; \sigma), \qquad (1.33)$$

where the score is evaluated over a sufficiently large number N of outcomes Y_i. Binter [12] demonstrates numerically that kernel dressing is an unbiased and robust evaluation method under PMS so that for a sufficient ensemble member size N_{ens}

$$\lim_{N_{ens}\to\infty} \hat{p}_{N_{ens}}(y) = p_{N_{ens}}(y), \qquad (1.34)$$

where $\hat{p}_{N_{ens}}$ is the forecast density estimate, and $p_{N_{ens}}$ is the model's forecast distribution. Hence, the forecast PDF derived from a kernel dressed perfect ensemble (PE) (see Section 1.5.2) is also expected to be unbiased.

A time series of continuous forecast PDFs can be graphically depicted using fan charts. Figure 1.6 shows a fan chart schematic where the percentiles of a forecast PDF at each time step, shown as different coloured bands, are connected, and appear as one continual plume from initialisation time $t = 0$ until lead time $t = \tau$. The plume typically spreads out with time, reflecting an increasing uncertainty of the true system state with time.

The dressed forecast PDF is blended with the climatological PDF, constructed by kernel dressing the sample climatological distribution to find σ_{clim}, the optimal kernel width of the climatological PDF. The sample climatology is the distribution of historical observations which is considered an estimate of the invariant measure of the system [40]. The optimal blending parameter α, along with $\hat{\sigma}$, can be determined by minimising the mean ignorance score over a training set of forecast ensemble-outcome pairs. The forecast PDF is finally produced by kernel dressing the ensembles in the outcome set and blending with the dressed climatological distribution using the optimised parameters $\hat{\sigma}$ and α. Hence, the forecast is given by

$$p(y) = \alpha p_{mod}(y) + (1 - \alpha)p_{clim}(y), \qquad (1.35)$$

where $\alpha \in [0, 1]$ is the blending parameter.
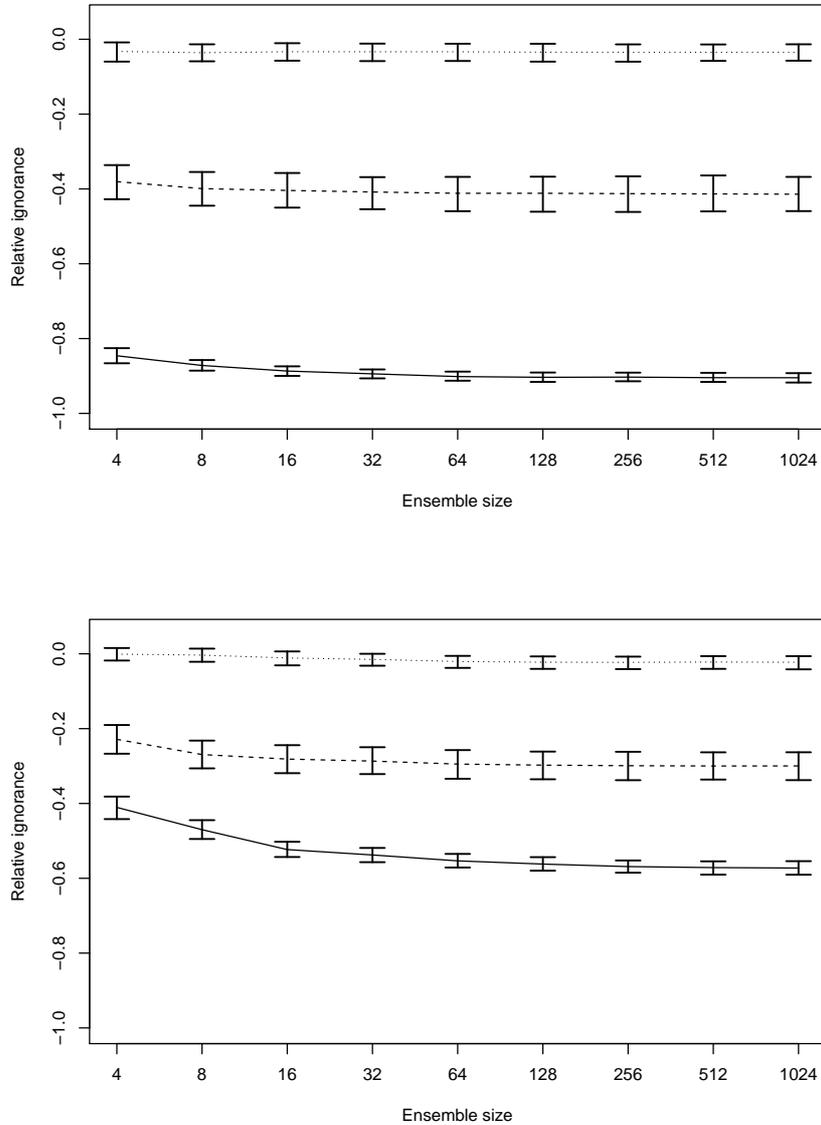
Figure 1.7: Skill of KDB forecasts: examples of empirical ignorance of perfect KDB forecasts at lead times 3.2 seconds (top) and 9.2 seconds (bottom). Lines denote climatological event probabilities as follows: $\theta = 0.5$ (solid), $\theta = 0.9$ (dashed), and $\theta = 0.99$ (dotted). The degree of forecast skill is dependent on forecast lead time and less so on ensemble size.

An example of the ignorance of forecasts produced with the KDB method over a set of n outcomes is plotted against member sizes for lead times of 3.2 seconds and 9.2 seconds, and shown in Figs. 1.7.

1.9 Atlantic basin hurricanes

Atlantic basin hurricanes have attracted much attention from the scientific and commercial sectors as well as from the general public. They are a powerful and awe-inspiring meteorological phenomenon, but also pose a serious threat to lives and livelihood. Hurricanes are typically accompanied by high magnitude winds, heavy precipitation, and storm surges to coastal areas which can inflict severe damage. Tropical cyclone activity in the North Atlantic basin accounts for only 11% of worldwide tropical cyclone activity, yet hurricanes have caused some of the largest losses of life and property caused by natural hazards, surpassed only by major earthquakes [43, 168].

1.9.1 Hurricane characteristics and data

A hurricane begins its lifetime as a cyclonic weather system with a centre of anomalously low surface level air pressure usually forming in the equatorial North Atlantic Ocean. If a tropical cyclone develops into a larger and more powerful storm, and attains 1-minute maximum sustained winds of at least $33ms^{-1}$ or $74mph$, it is classified as a hurricane. A wind speed-based classification index called the Saffir-Simpson scale categorises hurricanes by wind strength (CAT1-5: $\geq 33ms^{-1}$ or $74mph$) and major hurricanes (CAT3-5: $\geq 50ms^{-1}$ or $111mph$). Hurricanes range in diameter between 200km and 1300km, can have depths of up to 18km in altitude, and have lifespans of between 1 and 30 days. The annual Atlantic basin hurricane season runs from 1st June to 30th November with the most active period occurring around September. This peak period is concurrent with the annual extremes of two important ocean-atmospheric conditions for hurricane generation; sea surface temperatures (SSTs) are at their warmest and extend the furthest throughout the North Atlantic ocean, and vertical windshear is typically at a minimum over the tropical Atlantic [43].

The historical record of North Atlantic tropical cyclones extends back to over 500 years ago, with the first hurricane sighting in European history made

by Christopher Columbus near Cuba in June, 1494. Established trade routes between Europe, Africa and the Caribbean allowed observation of hurricanes but it was only those that caused human casualties or damage to ships and coastal communities that were documented [43]. Reported occurrences have increased steadily since that time with the increase in shipping traffic and observational capacity. The 1966-2012 climatological average of annual count is 6.2 a year for CAT1-5 hurricanes, and 2.3 a year for CAT3-5 hurricanes [143]. Detailed information on Atlantic hurricanes is currently provided by the National Oceanic and Atmospheric Administration (NOAA) through the National Hurricane Center (NHC). The most commonly used historical hurricane database (HURDAT)[8] contains data for 6-hourly wind speeds and locations during every tropical storm event since 1851.

There is scientific consensus that there has been a tropical cyclone undercount bias up until the mid-20th century due to the limited observational capabilities during a period of lower shipping lane density, lack of satellite technology and smaller populations in the Caribbean islands and American coastlines [80, 106, 123]. The Atlantic Hurricane Database Re-analysis Project [109, 107, 108, 68] led to a revision of the original data to "correct" for the undercount bias (and for other systematic biases and random errors) using a new methodology and new data sources. Nevertheless, due to ambiguities in the true counts due to these earlier observational limitations, this thesis considers data post-1966 using HURDAT database [106].

---

[8]http://www.aoml.noaa.gov/hrd/data_sub/re_anal.html

Chapter 2

Forecast Evaluation and Recalibration under PMS

The evaluation and recalibration of binary forecasts of the state of a nonlinear dynamical system under a *Perfect Model Scenario* (PMS) are investigated in this chapter to assess whether forecast recalibration is effective under PMS, and to identify the properties of a probabilistic forecast model which are key for forecast improvement. A perfect model allows a complete and exact description of the dynamics of a system, and producing perfectly accurate forecasts of future states of the system is limited only by imperfect knowledge of the current system state (i.e. initial condition uncertainty), and uncertainty in the model parameters. Forecast evaluation and recalibration are integral components of the operational forecasting framework (see 1.4.5) to monitor and improve the performance[1] of a forecast system.

The lessons learned in this chapter can aid in assessing whether it is more effective for forecasters to recalibrate to improve the quality of their forecast systems, or to concentrate efforts on advancing forecast techniques. Given its relative simplicity, forecast recalibration may provide a quicker and cheaper means to improve the performance of a forecast system than upgrading its

---

[1]Refers collectively to the attributes of forecast quality e.g. forecast skill, reliability, resolution, etc as defined in Section 1.6

various technical features (e.g. data assimilation scheme (DA), ensemble size, observation scheme, ...).

This chapter draws upon the discussion of Bröcker and Smith [24], and Bröcker [23] to inform the processes of forecast evaluation and recalibration with reliability diagrams, and for references to relevant terminology and notation. There are several new contributions to research on forecast evaluation and recalibration, however, which are included in the following chapter overview.

The perfect models used to construct binary forecasts of the state of a simple dynamical Lorenz63 system under PMS are described along with the forecast evaluation measures employed to assess their performance in Sections 2.1, 2.2, and 2.3. Next, a novel review and critique of various methods for binary probabilistic forecast recalibration is given in Section 2.4. Various challenges posed when recalibrating forecasts in principle and in practice are also highlighted.

Information-theoretical measures are appropriate diagnostic tools to evaluate the relative information content of forecasts before and after recalibration, and assess the effectiveness of forecast recalibration. Two such measures, relative entropy and ignorance [113, 172], are defined in Section 2.5, and employed to assess the performance of the various binary forecast construction methods described in Section 2.2. Reliability diagrams allow a quick visual evaluation of the effect of forecast recalibration. Comparison of forecast skill and forecast reliability is made in the context of recalibration using the decomposition of the ignorance score in Section 2.5. Decompositions of the ignorance score (see [199] and [194]) are shown to provide a novel assessment of the efficacy of recalibration; this is a new contribution.

Finally, a preliminary assessment of the efficacy of recalibration is made from the perspective of forecast lead time under PMS in Section 2.6 to prepare ground for a more complete investigation given in an *Imperfect Model Scenario* (IMS) in chapter 3. The results of this initial assessment indicate that recalibration is most effective where forecast skill is poorer (as at longer lead times), and where the climatological probability of a binary event is closer to 0.5.

2.1 Perfect model of the Lorenz63 system

The *Lorenz63* system [118] is a 3-dimensional dynamical system of deterministic nonlinear equations which is often employed as an illustrative system in weather modelling studies; it provides a tractable basis for the numerical demonstration of various aspects of forecast evaluation. For the parameter values specified in this work, it is chaotic (see Appendix A.1).

Consider a time series of observed states of a Lorenz63 system variable $s_0, \ldots, s_t, \ldots, s_N$ determined by the Lorenz63 equations plus some additive observational noise. An observation s_t at time t is defined as

$$s_t = \tilde{x}_t + \epsilon, \tag{2.1}$$

where $\tilde{x}_t$ is the true state of the system variable, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian additive observational noise term. The perfect model Ψ is initialised at time $t = 0$ with initial conditions $\mathbf{x}_0 = x_{0,1}, \ldots, x_{0,N_{ens}}$ which are generated by sampling from the inverse of a stochastic observational noise model. The process by which a core ensemble model is constructed is fully described in Section 1.5.2. Forecast error is accounted for entirely by IC uncertainty under PMS so, given a perfect data assimilation scheme, the effect of increasing model ensemble size on forecast performance is solely attributable to the model's ability to estimate the initial conditions. Probabilistic forecasts of the true state of the Lorenz63 system are generated here using four density construction methods. The best two methods are identified, and employed for the forecast evaluation and recalibration experiments performed under IMS which are discussed in chapter 3.

2.2 Binary forecasts

Probabilistic forecasts of binary events, or *binary forecasts* as they shall henceforth be referred to, represent the uncertainty in a prediction of a given binary

event occurring. Let x_θ denote a specified value of the observed climatological distribution of state variable x defined by the quantile θ, and τ denote the forecast lead time.

A binary forecast p_t is defined here and in Chapter 3 as the predictive probability that the true state $\tilde{x}_t$ at time $t = t_\tau$ lies above a given threshold x_θ on the 1-dimensional state space of the variable x. Let Y_t be the binary outcome variable representing the occurrence ($Y_t = 1$) or non-occurrence ($Y_t = 0$) of the event so that

$$Y_t = \begin{cases} 1, & \text{if } s_t > x_\theta. \\ 0, & \text{if } s_t \leq x_\theta. \end{cases} \tag{2.2}$$

A sample of independent binary forecast-outcome pairs (p_t, Y_t) of size N is used to evaluate a given forecast system (see Section 1.6).

2.2.1 Binary forecast construction

Given a model ensemble of N_{ens} members $\mathbf{x}_t = \{x_{t,1}, \ldots, x_{t,N_{ens}}\}$ which has been initialised at time $t = 0$, a binary forecast $p_t = P(Y_t = 1)$ is produced by translating the ensemble at time $t = t_\tau$ into a probability of the outcome lying above x_θ. The four forecast density construction methods employed here for constructing forecast PDFs are summarised below.

Counting methods

Forming binary forecasts with the *naïve* (NC) and *adjusted counting* (AC) methods involves the simple step of determining the relative frequency of ensemble members lying above x_θ.

The *naïve counting* (NC) method consists of construction of a density forecast by simply counting the number of ensemble members in each prescribed bin or category. In the case of binary forecasts, there are just two bins defined by the state variable threshold. A forecast constructed from raw ensemble relative frequencies is subject to sampling error associated with counting statistics at

smaller ensemble sizes [170]. 0% and 100% forecasts are unwise and prone to forecast busts. Different sets of forecast probability values are likely to make comparison and interpretation difficult [191].

The *adjusted counting* (AC) method is identical to the NC method except that an extra ensemble member is divided between the bins according to the climatological distribution about the threshold e.g. for 50% forecasts, if the counted probability above the threshold is $27/N_{ens}$, it would be adjusted to $(27 + 0.5)/(N_{ens} + 1)$.

Table 2.1: Configurations for PMS Lorenz63 binary forecast experiments

Experiment	System	State	Observational	PDF	Forecast-parameters		
No.		Variable	Noise	Method	θ	N_{ens}	τ^*
1	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.5	256	6.4
2	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.5	64	12.8
3	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.99	256	6.4
4	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.5	256	All
5	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.9	256	All
6	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.99	256	All

*in Lorenz63 seconds [118].

A Bayesian method

A sequential Bayesian updating approach to constructing binary forecasts can be employed to utilise forecast information from previous time steps (i.e. longer forecast lead times). The technique employed here is to update the latest forecast with the previous forecast. For example, a forecast with lead time τ_{i-1} is taken to represent a forecaster's prior belief of the outcome occurring at lead time τ_i where the lead times are ordered from longest lead time to shortest (i.e. $\tau_{i-1} > \tau_i$). The Bayesian interpretation of the probability of the binary outcome occurring at lead time τ_i is then obtained via

$$p_i^{Bayes}(y|p_i) \propto p_i(y) \times p_{i-1}(y). \tag{2.3}$$

where $p_i(y)$ is the ith forecast. The prior distribution $P(Y_t = 1)$ is constructed by using the information from the most recent forecast $p_{i-1}(y)$, and is updated at every time step. When τ_{i-1} exceeds the lead time of the forecast with the longest lead time, the climatological probability p^θ_{clim} is assigned to the prior probability value $p_{i-1}(y)$.

Kernel dressing and blending method (see Section 1.8)

A forecast density construction scheme such as kernel dressing can be deployed to construct a smoothed, continuous probability density function (PDF) and reduce the forecasting error associated with counting methods [214]. Kernel dressing has the added advantage over counting methods that the dressed ensemble is optimised with a proper scoring rule before it is evaluated out-of-sample. The kernel dressing and blending (KDB) method consists of dressing the raw ensemble to form a PDF, and then producing a binary forecast by taking the linear weighted average of the PDF and the climatological probability p^θ_{clim}. This latter process of pooling weighted probabilities is called *blending* [26].

The construction of binary forecasts from the perfect Lorenz63 model is now illustrated using the NC and KDB methods with a specified configuration of forecast-parameter values (i.e. $\{\theta = 0.5, N_{ens} = 256, \tau = 6.4\}$). This parameter value configuration (along with all configurations used in this chapter) is detailed in table 2.1, and labelled Expt. 1. The top plot in Fig. 2.1 shows the distribution of $N_{ens} = 256$ ensemble members $\hat{\mathbf{x}}_\tau$ after iterating the initial condition ensemble $\mathbf{x}_0$ forward in time to lead time $\tau = 6.4s$. Under the naïve counting method, a binary forecast of the x state variable is determined by the number of ensemble members lying above the specified threshold θ. There are 33 ensemble members above the threshold $x_{0.5}$ giving a binary forecast probability of $p_\tau(Y_\tau = 1) = 33/256 = 0.13$ under the NC method. To produce a binary forecast using the KDB method, the raw ensemble shown in the top plot is first kernel dressed, and then blended with the climatological PDF as

described above. The kernel dressed PDF is shown at every time step of 0.2s in the lower plot in Fig. 2.1. The true state of the variable $\tilde{x}_t$ is shown as the blue line. So, a binary forecast with lead time $\tau = 6.4s$ constructed from the kernel dressed ensemble in this case is evaluated with the outcome $Y_{6.4} = 0$ since the target lies below the specified threshold denoted by the horizontal dashed line (i.e. $\tilde{x}_{t_\tau} < x_\theta$). Figure 2.2 shows the forecast PDF, and the true state of the variable $\tilde{x}_{t_\tau}$ at lead time $\tau = 6.4s$.

The investigation is carried out in this chapter for a range of ensemble sizes and forecast lead times to identify values of these parameters for which forecasts are skilful under PMS. The parameter values for which recalibration can yield improvements in forecast performance are also investigated. The forecast evaluation and recalibration procedure using reliability diagrams is demonstrated in Sections 2.3 and 2.4.1 with a selected few experimental configurations within the PMS testbed.

2.3   Forecast evaluation

The performance of the binary forecasts generated in the experiments listed in table 2.1 is assessed using reliability diagrams, and the ignorance score IGN (see Section 1.6.2) in this chapter. Reliability diagrams (see Section 1.6.4) are a graphical representation of the full joint distribution of a sample of N binary forecast-outcome pairs in the form of the calibration function (see Eqn.(1.26)). The reliability diagrams in Figs. 2.3 and 2.4 correspond to Expt. 1 at $\tau = 6.4$. The forecast system which produced the forecasts in this case cannot be considered reliable because only two out of five bins fall within the consistency bars.
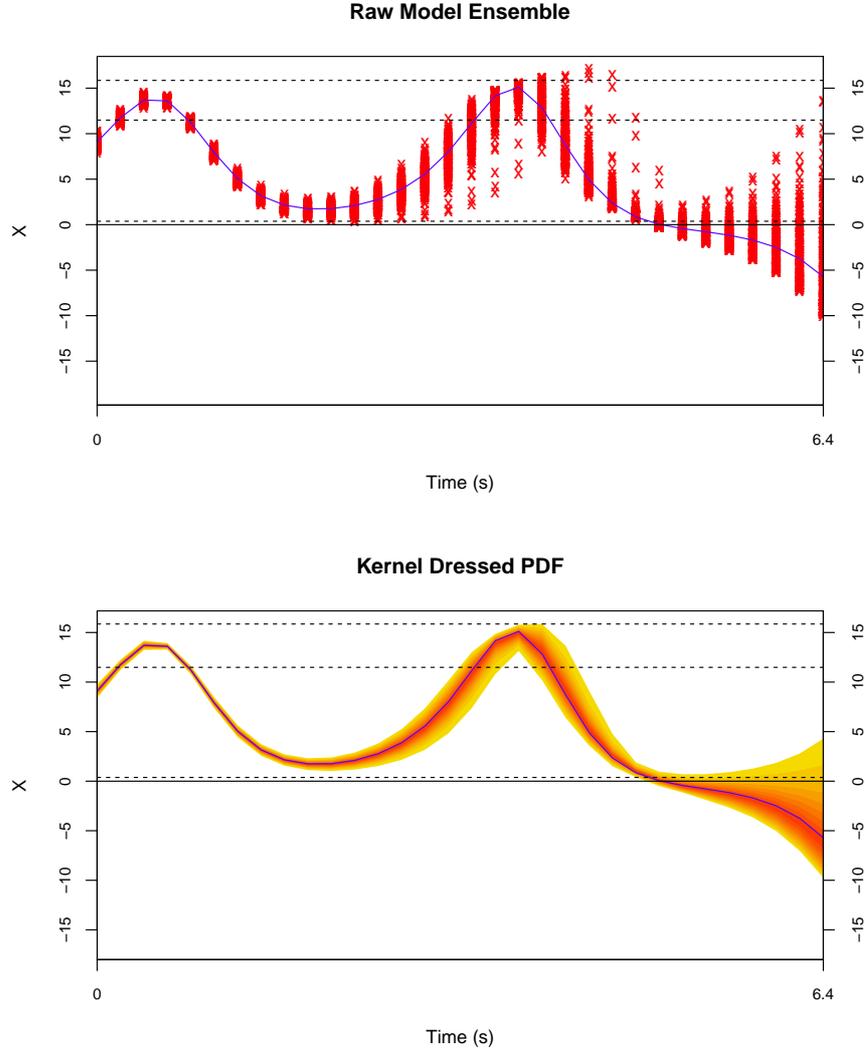
Figure 2.1: Ensemble forecasting under PMS: raw perfect Lorenz63 model ensemble generated in Expt. 1 (see table 2.1) (top), and fan chart showing the kernel dressed ensemble (PDF) constructed from the raw ensemble shown in the upper plot at every time step from $t = 0$ up to $t = t_\tau = 6.4s$ (bottom). Each individual colour band represents a 5% probability density percentile range of the PDF, from the 5th percentile to the 95th percentile (see Fig. 1.6 for the fan chart key). In each plot, the true state of the system variable is shown as a blue trajectory, and the dashed horizontal lines denote the 50th, 90th, and 99th percentiles of the climatological distribution representing the thresholds $\theta \in \{0.5, 0.9, 0.99\}$. The kernel dressed ensemble at time $t = t_\tau = 6.4s$ would be blended with the climatological distribution to produce the forecast PDF under the KDB method.
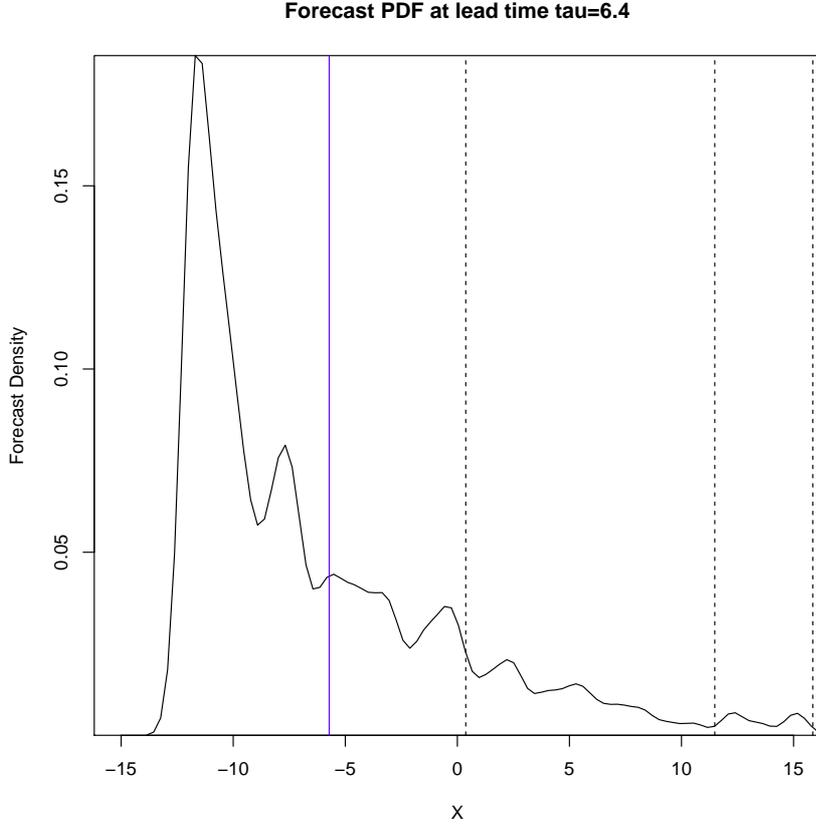
Figure 2.2: Ensemble forecasting under PMS: kernel dressed ensemble (PDF) corresponding to the PDF in Fig. 2.1 at $t = t_\tau = 6.4s$. The true state of the system variable is shown as a blue line at $\tilde{x}_{t_\tau} = -5.7$, and the dashed horizontal lines denote the 50th, 90th, and 99th percentiles of the climatological distribution representing the thresholds $\theta \in \{0.5, 0.9, 0.99\}$. Given that $\tilde{x}_{t_\tau} < x_\theta$, and that most of the probability density is below $x_\theta = 0.37$, the forecast appears more skilful than a climatological forecast $p^\theta_{clim} = 0.5$.

2.4 Forecast recalibration

Improving the performance of a probabilistic forecast system is an important step within the operational forecasting framework (see 1.5). A number of techniques which may or may not improve forecast skill have been proposed. These include technical upgrading of the forecast model (see Section 1.5), using a Bayesian approach to combine the output from several models [163], and a *recalibration* approach where forecast probabilities issued from a single model are

corrected based on historical performance. Most seasonal forecasting centres use a calibration scheme based on correcting *systematic biases* (i.e. persistent error trends) in the means and variances of past forecast statistics, which excludes the full information available from the joint distribution of past forecasts and outcomes [191]. Potentially, recalibrating forecasts with a technique which includes the joint distribution can prove superior. Forecast recalibration can
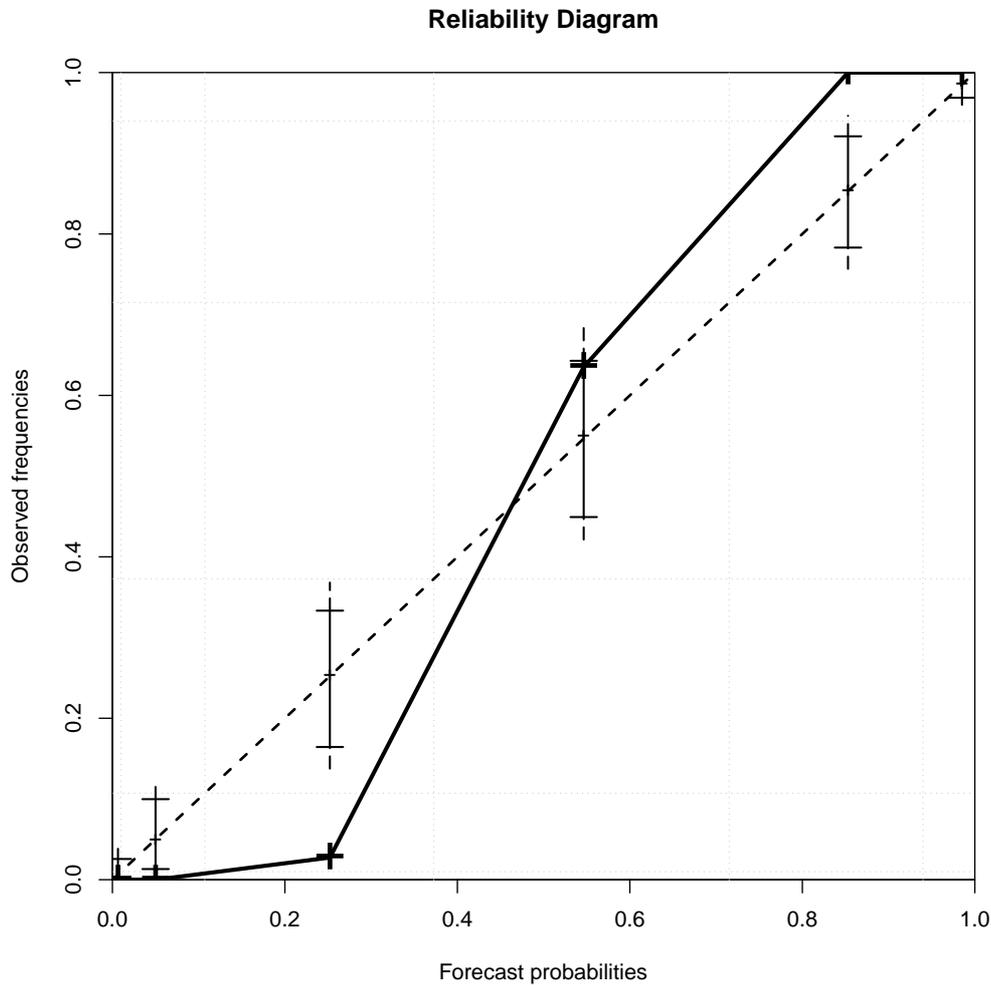
Figure 2.3: Forecast reliability: reliability diagram for Lorenz63 Expt. 1 at $\tau = 6.4$ (see table 2.1). Only two of the five observed frequencies at bins defined by $[0.373, 0.715]$ and $[0.940, 1.0]$ fall within the 5% - 95% (1% - 99% vertical dashed line) consistency bars indicating that the forecast system cannot be considered reliable.
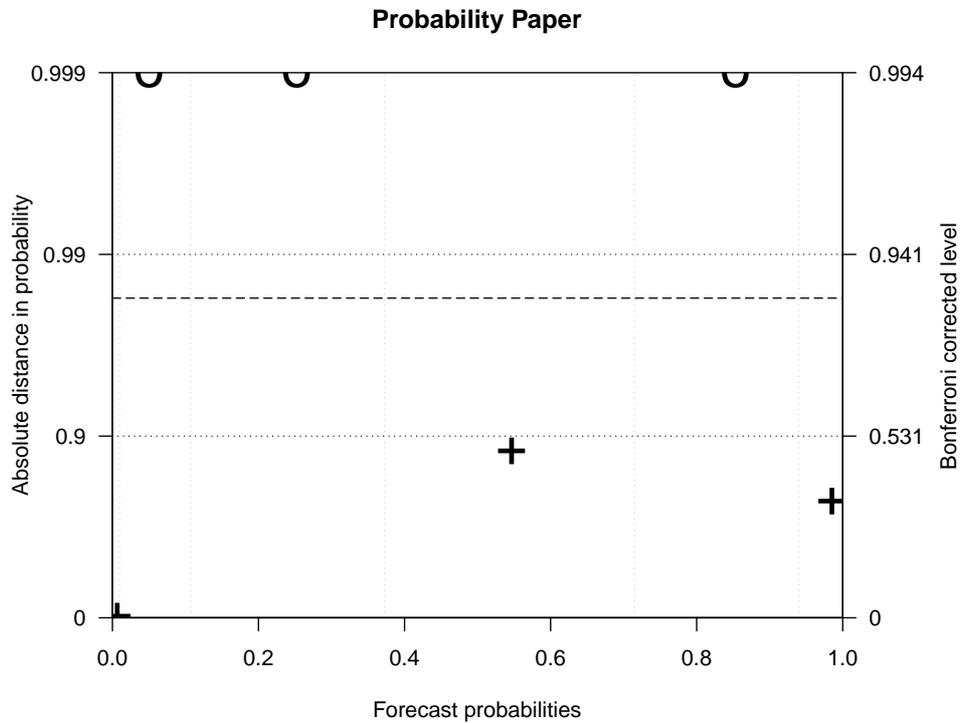
Figure 2.4: Forecast reliability on probability paper: reliability diagram on probability paper corresponding to Fig. 2.3. The two reliable forecast bins defined by [0.373, 0.715] and [0.940, 1.0] lie below the 0.9 probability distance dotted line. Circled symbols indicate an observed frequency outside the range of the y axis. The right-hand axis indicates the equivalent Bonferroni corrected levels for a reliable forecast so that the entire diagram (all 5 bins) would be expected to fall within the 0.99 probability distance band with an 95.1% chance. The dashed lines indicate where the entire diagram would be expected to fall within with a 90% chance if the forecast system was reliable.

be deployed to improve the reliability, and ideally the skill, of probability forecast systems [132]. Stephenson [191] argues that recalibration often leads to improved forecast performance because of the inadequacy of imperfect models.

A review and critique of the various recalibration methods proposed in the literature is presented in Section 2.4.1 followed by a discussion in Section 2.4.2 which contrasts the challenges of forecast recalibration in principle and practice. Both of these sections are, to the author's knowledge, new contributions to the

study of forecast recalibration. The effect of recalibration on forecast reliability is demonstrated *out of sample* using the set of experimental configurations defined in table 2.1.

2.4.1 Recalibration algorithms

Recalibration of binary forecasts can be implemented by computing the conditional distributions of the outcomes given a set of forecasts to estimate the calibration function [23]

$$\kappa(r) = P(Y = 1|p = r), \tag{2.4}$$

introduced in Section 1.6.4. Bröcker [23] outlines the forecast calibration problem in a Bayesian framework by defining the calibration function in terms of the compound distribution function

$$F(y, r) := P(Y = y|p < r), \tag{2.5}$$

where $y \in \{0, 1\}$, so that

$$\kappa(r) = \frac{dF(1, r)}{d[F(0, r) + F(1, r)]}, \tag{2.6}$$

where $F(0, r) + F(1, r)$ denotes the marginal distribution of r.

From Eqn. (2.4) it is clear to see that assigning a forecast with $\kappa(p)$ rather than p would result in a perfectly calibrated and skilful forecast [23]. Hence, a recalibrated forecast would ideally be assigned the probability

$$p^{re} = \kappa(r = p^{raw}), \tag{2.7}$$

where p^{raw} is the pre-recalibrated, or *raw*, forecast value. Unfortunately, the calibration function is unknown since usually the "true" PDF is unknown (and indeed may not even exist), so the task of forecast recalibration becomes an estimation problem. Estimates of the calibration function, henceforth denoted $\hat{\kappa}(p)$, need to be performed with samples of random data, and are thus considered random variables which are subject to residual errors (i.e. bias and

variance). Ideally, a balance is found between the bias and the variance of the calibration function, but typically the trade-off is non-trivial [23]. In the limit of an infinitely large sample of data, the estimate of the calibration function ideally converges onto the true calibration function (i.e. the diagonal line on the reliability diagram), so that

$$
\lim_{N \to \infty} \hat{\kappa}(r) = \kappa(r)
$$
$$
= r. \tag{2.8}
$$

In reality, the estimate of the calibration function is made using a finite training set T of forecast-outcome pairs denoted by

$$
T := \{(p_i^{train}, Y_i); i = 1, \ldots, N\}, \tag{2.9}
$$

where (p_i^{train}, Y_i) denotes the ith forecast– outcome pair. Obtaining accurate estimates of the calibration function is based on the assumption that all of the forecast-outcome pairs are independent and identically distributed (i.i.d.) according to the underlying distributions of (p, Y) (in chapter 4, the assumption of independence in real-world forecasting scenarios is considered). A common method for finding $\hat{\kappa}(p)$ is to categorise or "bin" the training set T into a number of partitions [196, 8, 24] in the same way that reliability diagrams are constructed. Binning proceeds as follows: let B_k, $k = 1, ..., K$ be the bins defined by partitioning the unit interval into K exhaustive and non-overlapping subintervals which are ideally of equal width if the forecasts are uniformly distributed over $[0, 1]$, or are at least equally populated if not, with the data from T if the forecast values [8]. Let I_k^{train} denote the sample of all indices i in B_k so that

$$
I_k^{train} := \{i; p_i^{train} \in B_k\}. \tag{2.10}
$$

Following the partitioning of the forecasts, a discretised estimate of the calibration function is evaluated at each bin average r_k^{train} by finding the *observed relative frequency* f_k^{train} i.e. the conditional frequency of occurrence of the binary event where $p_i^{train} \in B_k$. The observed frequencies and forecast averages

at each bin B_k are defined by

$$f_k^{train} = \frac{\sum\limits_{i \in I_k} Y_i}{\#I_k},$$
(2.11)

and

$$r_k^{train} = \frac{\sum\limits_{i \in I_k} p_i^{train}}{\#I_k},$$
(2.12)

respectively. $\#I_k$ is the number of indices in bin B_k. A conventional reliability diagram of the training set of forecast-outcome pairs consists of a plot of f_k^{train} against r_k^{train}, and thus, the estimate of the calibration function (2.4) is given by

$$\hat{\kappa}_T(r_k) = P(Y_i = 1 | r_k^{train})$$
$$= f_k^{train}.$$
(2.13)

As explained in Section 1.6.4, a forecast system at bin B_k is reliable if f_k^{train} falls within the consistency bars computed according to r_k^{train}. Note that evaluation of the calibration function for values $p \neq r_k^{train}$ can be performed using linear interpolation [23].

Recall from Eqn. (2.7) above that recalibration ideally consists of re-assigning a raw forecast value p^{raw} with the calibration function evaluated at that value (i.e. $\kappa(p^{raw})$). $\kappa(p^{raw})$ is a perfectly reliable and more skilful forecast than p^{raw} [23]. Without knowing the true calibration function, however, the best alternative is to estimate it. Estimation of κ can be performed with a number of algorithms, including Eqn. (2.13), after binning the forecasts as if constructing a reliability diagram. Unfortunately, the binning method is likely to introduce further residual errors into the forecast recalibration process where the probabilities are not fixed in each bin, particularly if a bin population, $\#I_k$, is small [23]. The additional error is unavoidable since reliability diagrams are constructed on the basis that all the forecast values X_i are binned and averaged so as to yield non-trivial observed relative frequencies [24]. The best that one can do

is to reach a balanced trade-off between bias and variance by specifying the bins so that they are equally populated (this is discussed in more depth in Section 3.4.1). There are, however, recalibration algorithms that circumvent the bin averaging of forecast values before recalibration, such as kernel-smoothing estimation of the calibration function [23]. All of the recalibration algorithms reviewed in this thesis are defined with respect to calibration function estimate $\hat{\kappa}(p)$.

Simple Translation

A simple algorithm for recalibrating binary forecasts is to find $\hat{\kappa}_T(p)$, and then reassign the forecast probabilities p^{raw} of an evaluation set [70, 7, 196], as described in the previous paragraph. In short, for each probability bin B_k in the evaluation set, the calibrated probability p_i^{re}, $i \in I_k$ is equal to the observed frequency f_k^{train} corresponding to bin B_k in the training set T, so

$$
\begin{aligned}
p_i^{re} &= \hat{\kappa}_T(r_k^{train}) \\
&= P(Y_i = 1 | r_k^{train}) \\
&= f_k^{train}.
\end{aligned}
\tag{2.14}
$$

Equation 2.14 implies that the probability bins in the training and evaluation sets are identical (i.e. $B_k^{train} = B_k^{raw}$). This constraint is not ideal where the categorisation of the forecast probabilities defined for the training set leads to small populations, and hence, under-representation of the evaluation forecasts at certain bins. Estimates of the calibration function after recalibration are then prone to larger sampling uncertainty. If so, linear interpolation can be utilised, as in Atger [7], to estimate the calibration function $\hat{\kappa}_T(r_k^{raw})$ at each r_k^{raw}, and hence, the calibrated probability p_i^{re} in such a case. Increasing the size of the evaluation set to increase bin populations might also alleviate this problem under the assumption that the training forecasts and evaluation forecasts share the same distribution.
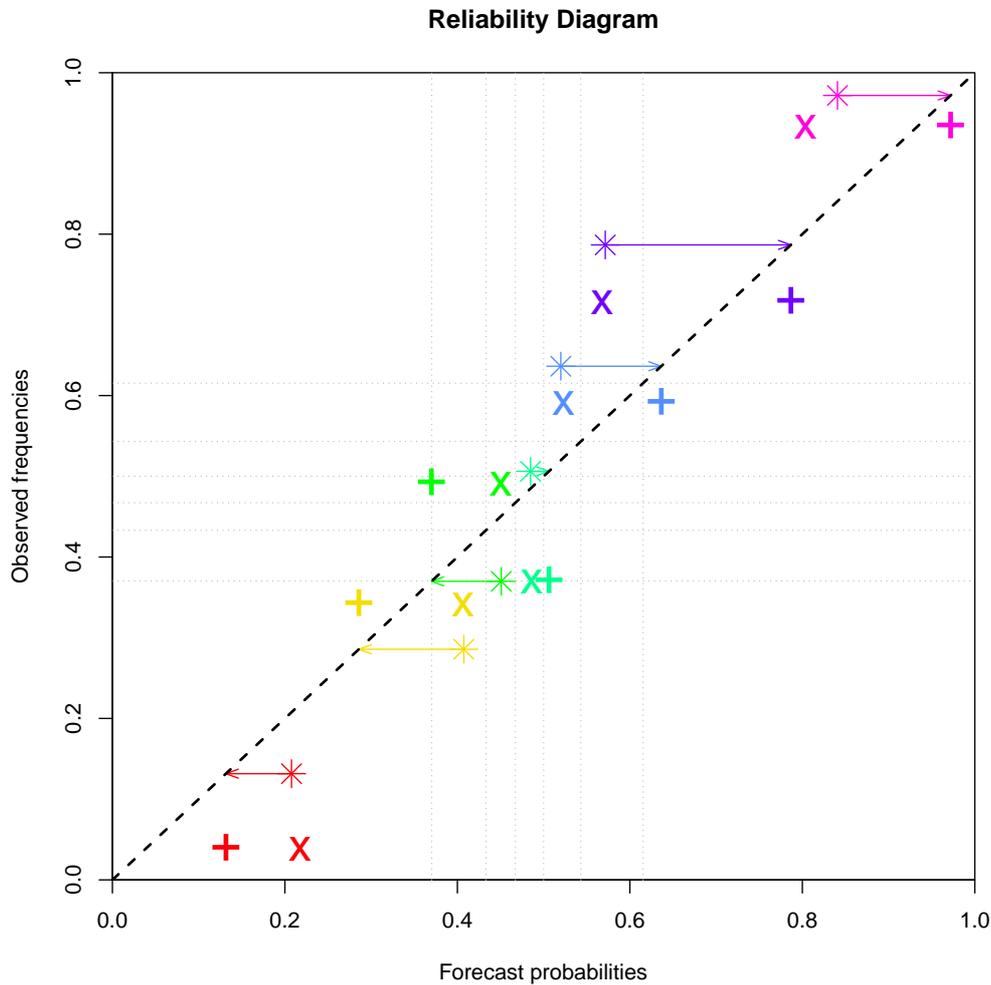
Figure 2.5: Simple translation recalibration: reliability diagram schematic of the *simple translation* recalibration algorithm using a training set of Lorenz63 binary forecast (asterisks) to recalibrate the evaluation set of forecasts (pluses → crosses) both generated in Expt. 2 (see table 2.1). Most bins are translated closer to the diagonal suggesting improved forecast reliability. Each bin is coloured differently for clarity.

Consider a set of Lorenz63 binary forecasts issued by a forecast system which produces forecasts using the AC forecast density construction method defined in Section 1.8. The *simple translation* algorithm defined by Eqn. 2.14 is deployed to recalibrate the forecast system. The schematic plot shown in Fig. 2.5 demonstrates the recalibration procedure. For each probability bin

B_k, the average of the forecast values r_k^{train} is translated horizontally to the diagonal line, determining the magnitude and direction of the translation of each forecast $\{p_i^{raw} \in B_k\}$. Fig. 2.5 is simplified to demonstrate the process through the translation of the bin average r_k^{train} rather than every single forecast p_i^{raw}.

The schematic highlights a previously unreported limitation of the simple translation algorithm. Increasing the reliability of a forecast system at a given bin B_k through simple translation can be difficult where there are large deviations between the forecast values p_i^{raw} in bin B_k and the bin average r_k. Those forecast values p_i^{raw} which have larger deviations from f_k^{train} will be subject to relatively larger adjustments, rendering them less reliable after recalibration if, in fact, they were reliable before recalibration, that is $\hat{\kappa}_T(p_i^{raw}) \simeq \kappa(r)$. Of course, the reliability of individual forecasts cannot be evaluated, but, by collectively recalibrating forecasts through simple translation, the ability to increase the overall reliability of the forecast system at B_k is reduced. Hence, the simple translation algorithm is only likely to be effective for increasing the reliability of forecast values $\{p_i^{raw} \in B_k\}$ which are near to the bin average value. The limitation could be addressed by reducing the bin interval widths but this reduces the bin population $\#I_k$, and can result in under-sampling.

The limitation of the simple translation algorithm is firstly illustrated with the following hypothetical example: consider a bin B_K defined by the interval $(0.7, 1.0]$ with bin population $\#\{p_i^{raw} \in B_K\} = 999$ that has a high proportion of large probability values such that $p_i^{train} = 0.995$, $i = 6, \ldots, 999$, but also has five forecast values $p_i^{train} = 0.75$, $i = 1, \ldots, 5$, so that $r_K = (5 \times 0.75 + 994 \times 0.955)/999 = 0.994$. Let the forecast bin be perfectly reliable to within $5\%-95\%$ consistency so that, after recalibration, all of the forecast values are translated to the observed relative frequency $f_K = r_K = 0.994$. The recalibrated forecast system will only issue forecasts $p_i^{re} = 0.994$ for all i even though the forecast $p_1^{train} = 0.75$ might already be reliable.

Recall the Lorenz63 forecasts from Expt. 1 which are plotted on reliability

Bin Probabilities
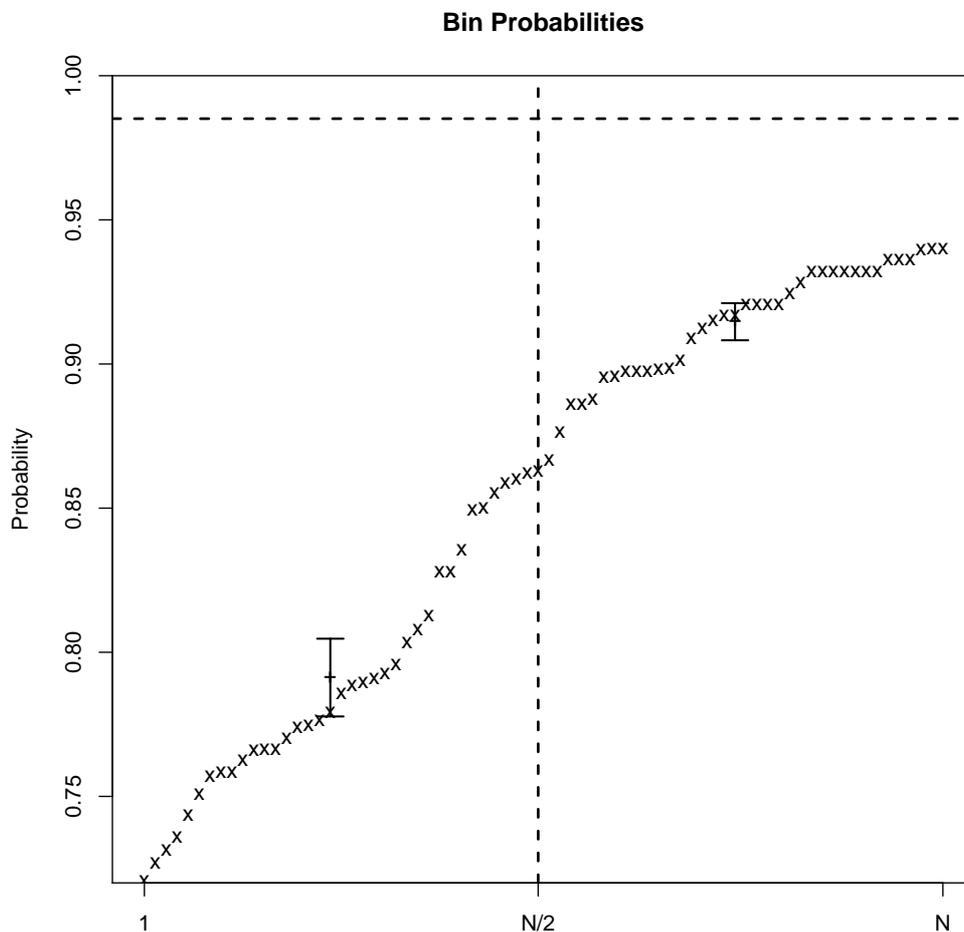
Figure 2.6: Limitation of simple translation: distribution of forecasts p_i^{re} in the fifth bin B_5 ($\#\{p_i^{re} \in B_5\} = 73$) sorted in ascending order (forecasts are generated in Expt. 1; see table 2.1). 5% - 95% Wald confidence intervals, plotted for the two sub-bin averages at both sub-bin mid-points show that the difference between the sub-bin average is highly significant (p-value $< 2.2 \times 10^{-16}$).

diagrams in Figs. 2.3 and 2.4. Figure 2.6 shows the distribution of forecast probability values in the highest bin B_5, where $r_5 = 0.828$, which has a large variance over the interval $(0.715, 0.940]$, and is also skewed towards higher probability values. Recalibration through simple translation will assign all values $p_i^{re} = f_5 = 0.853$, a large increase to lower forecasts in the bin. Ideally, the effec-

tiveness of this recalibration algorithm would be checked before implementing it. For instance, by dividing a bin into two equally populated sub-bins, and applying a simple t-test [217] to determine if the difference between the averages of the two sub-bins r_{k_1} and r_{k_2} is consistent with the independence assumption.

The test statistic is given as

$$z = \frac{r_{k_2} - r_{k_1}}{(s_{k_1}^2/(\#I_k/2) + s_{k_2}^2/(\#I_k/2))^{1/2}}, \tag{2.15}$$

where $s_{k,\cdot}$ is the standard deviation of the sub-bins. The test indicates whether the difference between the averages is significant at a given significance level α, that is, for p-values $< \alpha$, it is unlikely that the distributions sampled in each of the two sub-bins have the same mean (i.e. $r_{k_1} = r_{k_2}$). In that case, the simple translation algorithm can be considered inappropriate for recalibration. The test for the difference between the sub-bin averages of bin B_5 is highly significant (p-value $< 2.2 \times 10^{-16}$), indicating that the averages are different under the assumption of independence. This result may reduce the effectiveness of recalibration for increasing forecast reliability at that bin.

Linear Regression

Palmer et al [152] propose a recalibration algorithm employing a linear regression to estimate the calibration function. The algorithm is implemented as follows: a weighted least-squares regression line is fitted to the plots of f_k^{train} versus r_k^{train} representing the calibration function estimate $\hat{\kappa}$ on a reliability diagram for all bins B_k, $k = 1, \ldots, K$. The regression line is expressed in terms of the calibration function as

$$\hat{\kappa}_T(r_k^{train}) = \hat{\beta}_0 + \hat{\beta}_1 r_k^{train}, \tag{2.16}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and slope of the regression line, respectively. The weights w_k are determined according to the sizes of the bins so that

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{K} w_k \big| y_k - \beta_0 - \beta_1 r_k^{train} \big|^2. \tag{2.17}$$

Reliability Diagram
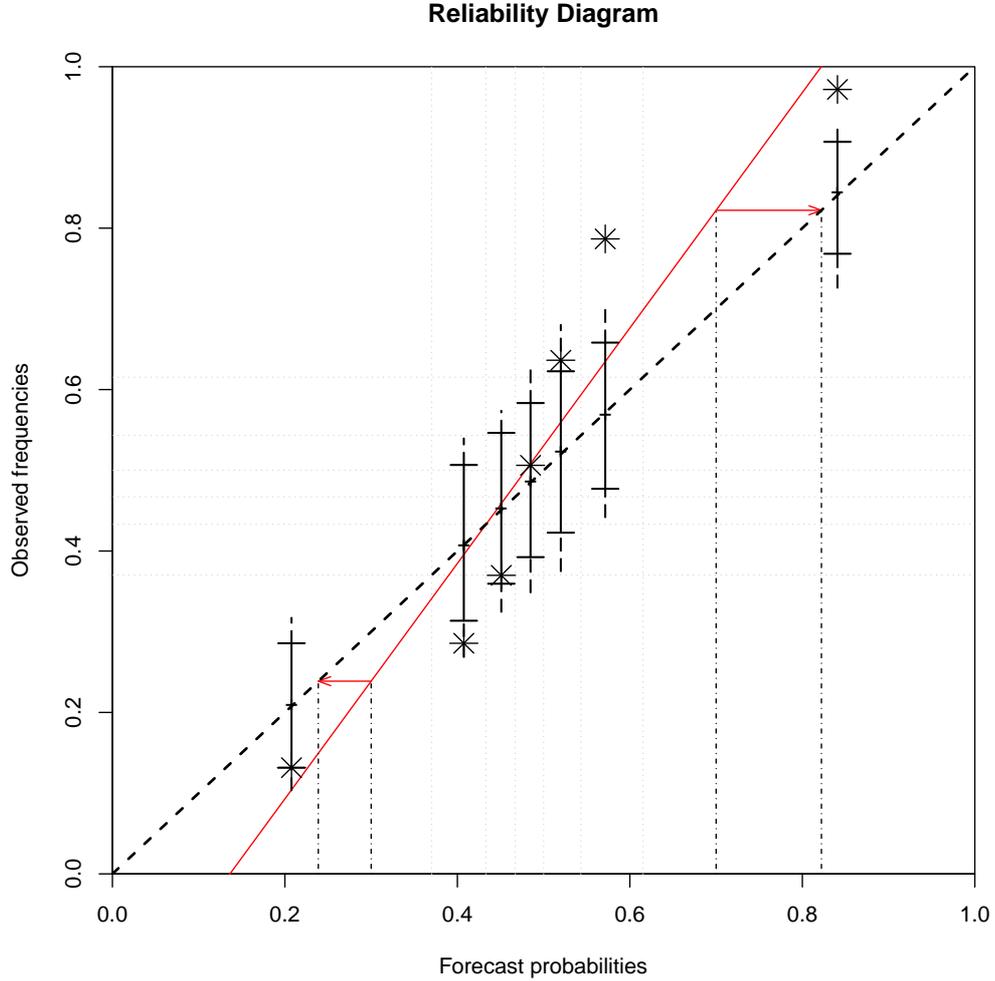
Figure 2.7: Forecast recalibration using linear regression: Reliability diagram schematic demonstrating the linear regression recalibration algorithm using a training set of binned and averaged Lorenz63 binary forecasts (asterisks) to recalibrate the evaluation set of forecasts both generated in Expt. 2 (see table 2.1). A linear regression line is fitted to the plotted points, from which the horizontal distance to the diagonal line determines the magnitude by which a raw forecast needs to be translated to be recalibrated.

The forecast probabilities p_i^{raw} in the evaluation set are calibrated by finding the point on the regression line whose abscissa value corresponds to p_i^{raw} (i.e. $(p_i^{raw}, \hat{\kappa}_T(p_i^{raw}))$). The calibrated probability value p_i^{re} is then given by the abscissa value corresponding to point on the diagonal line of the reliability

diagram at the same ordinate value $\hat{\kappa}_T(p_i^{raw})$. This is expressed as

$$p_i^{re} = \hat{\kappa}_T(p_i^{raw}) \tag{2.18}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 p_i^{raw}. \tag{2.19}$$

Figure 2.7 shows a schematic example of the calibration procedure using a linear regression line fitted to the training set of forecast-outcome pairs generated under Expt. 2 (see table 2.1). The regression coefficients of the fitted red line in Fig. 2.7 are $\beta_0 = 0.199$ and $\beta_1 = 1.459$. The forecast value $p_i^{raw} = 0.7$ is translated to the diagonal line in the same fashion as with the simple translation algorithm described in the previous section, giving a calibrated forecast value $p_i^{re} = 0.822$. In this case, the original forecast system is underconfident, that is, it over-forecasts at lower probabilities, and under-forecasts at higher probabilities (see Wilks [217]), hence the calibrated forecasts all tend to be decreased.

One advantage of the linear regression algorithm over the simple translation algorithm is that there is no requirement to bin the raw probabilities before recalibration, and calibrate them according to a single translated forecast bin average r_k^{train}. This difference implies that information is not lost when recalibrating each individual forecast value with linear regression. Applying a linear regression to a reliability diagram to estimate the calibration function $\kappa_T(p)$ is problematic, however, if the line has a slope $\beta_1 > 1$, and does not span the unit interval (i.e. the uncalibrated forecast system is underconfident). If one attempts to calibrate a forecast value where

$$p_i^{raw} > \frac{1 - \hat{\beta}_0}{\hat{\beta}_1}, \tag{2.20}$$

or

$$p_i^{raw} < \frac{-\hat{\beta}_0}{\hat{\beta}_1}, \tag{2.21}$$

then the calibrated forecast value p_i^{re} will lie outside the range $[0, 1]$. Obviously, this is a nonsense [217], as it implies that forecast probabilities can take values

less than 0% and greater than 100%. Assuming a 0% or 100% probability violates Cromwell's law[2].

Logistic Regression

The problem of recalibrating forecasts to values outside of the range $[0, 1]$, highlighted in the previous section, can be circumvented by employing a generalised linear model (GLM) such as a logistic regression. Although logistic regressions have been used to recalibrate ensemble forecasts [161, 215, 218], or forecast model predictors based on forecast data [120], the algorithm described here, based on the linear regression algorithm of Palmer et al. [152] above, is a new technique. The logistic regression model links a single predictor variable x to the mean of the dependent variable γ, which is assumed to be binomial, via a *logit* link function, given by

$$ln\left(\frac{\gamma(x)}{1 - \gamma(x)}\right) = \beta_0 + \beta_1 x. \tag{2.22}$$

To fit a logistic regression curve to a reliability diagram, the calibrated probabilities p_i^{re} are modelled as the mean parameter (i.e. $\gamma(x) = p_i^{re}$) by regressing the observed frequencies f_k^{train} on the forecast averages r_k^{train}, in a similar fashion to the linear regression algorithm in Section 2.4.1. The fitted model can then be used to determine the calibrated probabilities via the logit link function, expressed as

$$ln\left(\frac{p_i^{re}}{1 - p_i^{re}}\right) = ln\left(\frac{\hat{\kappa}_T(p_i^{raw})}{1 - \hat{\kappa}_T(p_i^{raw})}\right) \tag{2.23}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 p_i^{raw}, \tag{2.24}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the fitted regression coefficients. The relationship between $\hat{\kappa}_T$ and p_i^{raw} need not be linear, which is the case for linear regressions. Instead, the coefficients represent a change in the logit for a unit change in r_k^{train}. So,

---

[2]is named by Lindley [115], based on the quote "I beseech you, in the bowels of Christ, think it possible that you may be mistaken." (in a letter Cromwell wrote to the Church of Scotland on August 5th, 1650.)

applying a logistic regression allows for non-linearity in the calibration function, and restricts the calibrated forecast values p_i^{re} to the range $[0, 1]$. Primo et al [161] highlight that, although nonlinear techniques may be more flexible and better for correcting biases in forecast systems than linear techniques, both techniques tend to reduce the range of forecast probability values after recalibration, especially if the forecast system was poor before recalibration. This is discussed in more depth in Section 3.4.3. Figure 2.8 shows a schematic example of forecast calibration using a fitted logistic regression curve. The forecast-parameter values are the same as those in Fig. 2.7 for comparison. So, for example, a forecast value $p_i^{raw} = 0.7$ is translated to the diagonal line, giving a calibrated forecast value $p_i^{re} = 0.893$.

There are three other recalibration algorithms which, unlike the preceding three, do not require binning and averaging of the *training* set of forecasts p_i^{train}, and linear interpolation between the bin averages r_k^{train}[3]. Instead, the calibration function is estimated directly from the forecasts p_i^{train} and, as a result, is not subject to the residual errors associated with mean estimates (i.e. r_k^{train}). These other algorithms - kernel density estimation, beta-transformed linear pooling, and relative operating characteristic (ROC) curve fitting - are described in the following sections. A comparison of the performances of all the algorithms is deferred to Section 3.4.2 where the results of forecast recalibration under IMS are presented. In short, the kernel density estimation algorithm and beta-transformed linear pool algorithm perform better than the simple translation and regression algorithms, while the ROC curve algorithm tends to perform rather poorly.

---

[3]Note that binning the forecasts is not a general requirement of linear and logistic regressions applied to recalibration, see, for example, Wilks [215] and Hamill and Whitaker [120]

Reliability Diagram
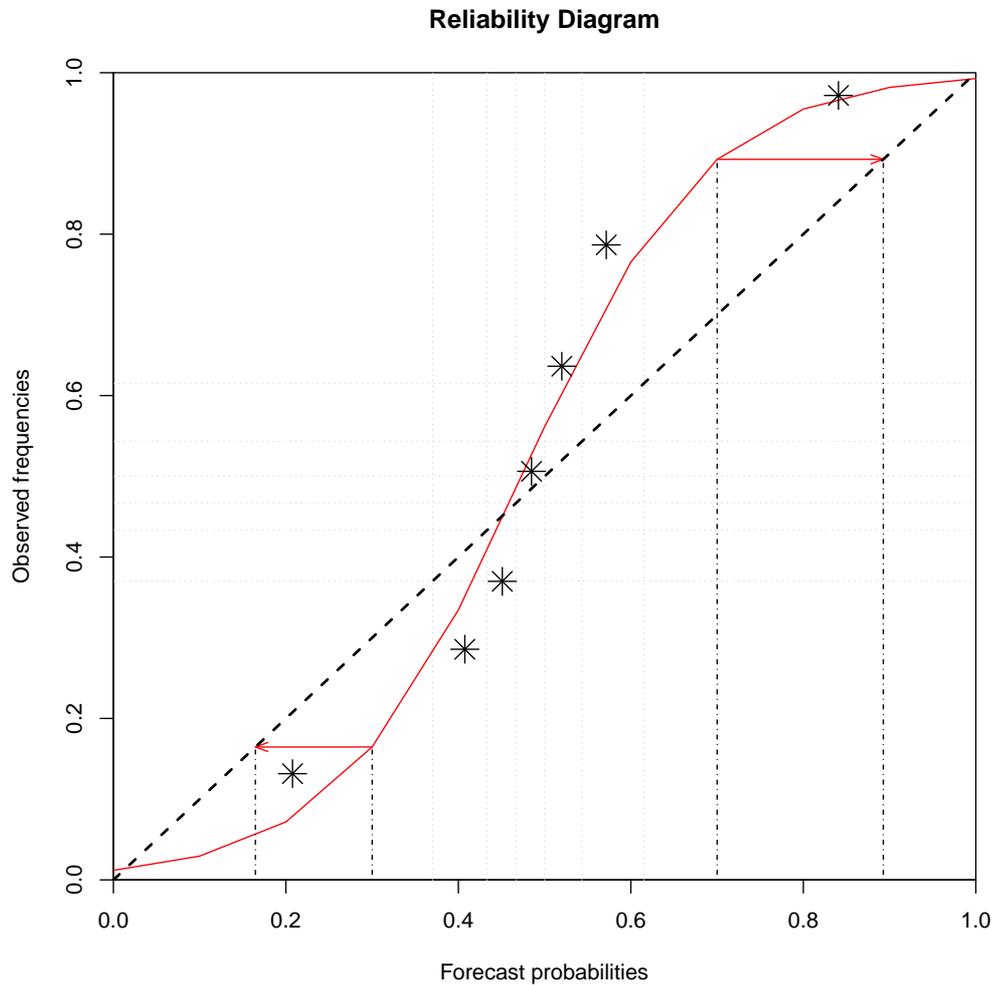
Figure 2.8: Forecast recalibration using logistic regression: Reliability diagram schematic demonstrating the logistic regression recalibration algorithm using a training set of binned Lorenz63 binary forecasts (asterisks) to recalibrate the evaluation set of forecasts both generated in Expt. 2 (see table 2.1). A logistic regression line is fitted to the plotted points, from which the horizontal distance to the diagonal line determines the magnitude by which a raw forecast needs to be translated. For example, the two red lines show evaluation forecast probability values of 0.3 and 0.7 are calibrated to values of 0.16 and 0.89, respectively. Note that the fitted curve is a better fit than the linear regression line plotted in Fig. 2.7.

Relative Operating Characteristic (ROC) curve fitting

Atger [8] proposes a statistical framework to estimate the calibration function for small forecast-outcome pair sample sizes. A *Relative Operating Character-*

istic (ROC) curve is fitted with a bi-normal model, and used to estimate the relative bin populations $\#I_k/N$ and observed relative frequencies f_k of each bin B_k in the training set. ROC curve fitting with bi-normal models was outlined in Section 1.6.5. Under the bi-normal assumption, a ROC curve is a straight line after transformation of its x and y coordinates (i.e. FAR and HR) into their corresponding standardised normal deviates. HR and FAR are approximations of the compound distribution functions $F(0, p_k)$ and $F(1, p_k)$ respectively defined by Eqn. (2.5) [23].

The relative bin populations $\#I_k^{binorm}/N$ and observed relative frequencies f_k^{binorm} under the bi-normal assumption can be estimated as follows: the ROC curve is fitted by transforming HR and FAR into z_{HR} and z_{FAR} and finding the best-estimate linear fit. Next, the (FAR, HR) points of the original ROC curve are orthogonally projected onto the fitted curve to estimate HR and FAR according to the bi-normal model, which are in turn used to recursively compute $\#I_k^{binorm}/N$ and f_k^{binorm}. The raw forecasts are finally recalibrated using the simple translation algorithm outlined in Section 2.4.1, that is

$$p_i^{re} = \hat{\kappa}_T(p_i^{raw}) \tag{2.25}$$

$$= f_k^{binorm}, \tag{2.26}$$

where $i \in I_k^{raw} := \{i; p_i^{raw} \in B_k\}$. Bröcker [23] points out that this algorithm leads to low variance of estimates of the calibration function because they are restricted to very few *degrees of freedom*, and therefore subject to a possible trade-off towards larger bias of calibration function estimates. A more detailed discussion of controlling the bias-variance trade-off and degrees of freedom of calibration function estimates is given in Section 3.4.1 in a context including imperfect models.

Kernel Density Estimation

Bröcker [23] proposes a forecast recalibration algorithm which, like the ROC curve approach above, includes estimating the compound distribution func-

tions $F(0, r)$ and $F(1, r)$ to derive the calibration function (see Eqn. (2.6)). Unlike the ROC curve approach, however, kernel density estimators are instead employed to estimate $F(0, r)$ and $F(1, r)$, that is

$$\hat{F}(y, r) = p_y^{train} \frac{1}{\#I_y^{train}} \sum_{i;Y_i=y} \Phi\left(\frac{r - p_i^{train}}{\delta}\right), \tag{2.27}$$

where $p_y^{train} = P(Y = y)$ and $\#I_y^{train} := \#\{i; Y_i = y\}$ are estimated from the training set, and Φ is the standard Gaussian CDF. Hence, $F(0, r)$ and $F(1, r)$ are estimated by a sum of Gaussian kernels with bandwidth δ which is optimised by applying a proper scoring rule to the calibration function estimate $\hat{\kappa}_T$. The optimised estimate of δ using ignorance can be expressed as

$$\hat{\delta} := \arg\ \min_{\delta} \sum_{i=1}^{N} -log_2 \hat{\kappa}_T(p_i^{train}(Y_i); \delta). \tag{2.28}$$

Note that kernel density estimation differs somewhat to kernel dressing which was explained in Section 1.8 (see Binter [12] for a discussion on the differences between kernel density estimation and kernel dressing).

The raw forecasts are finally recalibrated using the calibration function estimate with optimised kernel bandwidth parameter δ, that is

$$p_i^{re} = \hat{\kappa}_T(p_i^{raw}; \delta) \tag{2.29}$$

$$= \frac{d\hat{F}(1, p_i^{raw})}{d[\hat{F}(0, p_i^{raw}) + \hat{F}(1, p_i^{raw})]}, \tag{2.30}$$

where $i \in I_k^{raw} := \{i; p_i^{raw} \in B_k\}$.

Beta-transform linear pool

Empirical evidence abounds for the improvement in predictive performance when combining two or more available probabilistic forecasts of some event compared to individual probabilistic forecasts of that same event [165]. Despite this evidence, Ranjan and Gneiting [165], and Hora [77] prove that achieving perfect forecast reliability via the recalibration of probability forecasts using a

non-trivial weighted average of two or more distinct, calibrated probability forecasts is not possible. Ranjan and Gneiting [165] suggest, however, that linear combined probability forecasts perform better than individual forecasts which is supported by a wealth of empirical evidence across various fields including meteorology, economics, and medical science [165].

Ranjan and Gneiting [165] propose a parametric beta-transformed linear pooling (BLP) technique for nonlinear recalibration of linear combinations of probabilistic forecasts which is highly effective for increasing forecast reliability and forecast skill. In its general form, the BLP technique consists of aggregating m probabilistic forecasts $p_1, \ldots, p_m$ into a weighted linear combination, and then applying a *beta transform*. This process is formulated as

$$p = H_{\alpha,\beta} \left(\sum_{j=1}^{m} \omega_j p_j \right), \tag{2.31}$$

where $\omega_1, \ldots, \omega_m \geq 0$ and $\omega_1 + \ldots + \omega_m = 1$, and

$$H_{\alpha,\beta}(x) = \mathcal{B}(\alpha, \beta)^{-1} \int_0^x t^{\alpha-1} (1-t)^{\beta-1} \mathrm{d}t, \tag{2.32}$$

for $x \in [0, 1]$, is the cumulative distribution function of the beta density with parameters $\alpha > 0$ and $\beta > 0$.

A forecast p_i^{raw} is recalibrated by compositing a beta transform and the linear combination of each forecast p_i^{train} in the training set with the climatological probability of the observed state variable lying above the specified threshold θ. In effect, the recalibrated forecast is given by

$$p_i^{re} = \hat{\kappa}_T(p_i^{raw}) \tag{2.33}$$

$$= H_{\alpha,\beta} \left(\omega p_i^{train} + (1-\omega) p_{clim}^{\theta} \right). \tag{2.34}$$

Maximum likelihood estimates of the weights $\omega_1, \ldots, \omega_m$, and the parameters α and β of the beta transform are found by numerically optimising the log-likelihood function of the BLP model

$$l(\omega_1, \ldots, \omega_m; \alpha, \beta) = \sum_{i=1}^{N} [Y_i log(p_i^{blp}) + (1 - Y_i)log(1 - p_i^{blp})]$$

$$= \sum_{i=1}^{N} [Y_i log(\omega p_i^{train} + (1 - \omega)p_{clim}^{\theta})$$

$$+ (1 - Y_i)log(1 - \omega p_i^{train} + (1 - \omega)p_{clim}^{\theta})], \qquad (2.35)$$

under the constraints $\omega_1, \ldots, \omega_m \geq 0$, $\omega_1 + \ldots + \omega_m = 1$, $\alpha > 0$, and $\beta > 0$. Note that the model is not constrained to having non-trivial weights so linear pooling of p_i^{train} and p_{clim}^{θ} is not enforced.

2.4.2 Contrasting the challenges of forecast recalibration in principle and in practice

The process of forecast recalibration has been outlined in this section, and framed by the problem of estimating the calibration function κ to correct forecast probabilities. An overview of proposed algorithms for estimating the calibration function has also been presented along with discussion of the various challenges in principle and in practice for performing forecast recalibration. These are summarised below.

In principle:

- forecast recalibration requires collection of forecast-outcome pairs to make statistical corrections to forecast probabilities

- the calibration function κ must be assumed to exist when there may be no reason for it to exist and, if it does, it is generally unknown

In practice:

- the calibration function κ must be estimated from a training set T of forecast-outcome pair data (p_i^{train}, Y_i)

- accurate estimation of the calibration function is based on the assumption that all of the forecast-outcome pairs (p_i^{train}, Y_i) are i.i.d.

- estimates of the calibration function $\hat{\kappa}$ with finite sample sizes of data are always subject to residual errors (i.e. bias and variance). Ideally, a balance between the bias and variance of $\hat{\kappa}$ is reached, but in practice, identifying the balance is difficult.

- binning and averaging steps introduce further residual errors unless forecast probabilities are taken as fixed values and bin populations are similar

- the simple translation recalibration algorithm $p_i^{re} = \hat{\kappa}_T(r_k^{train})$ is susceptible to erroneous recalibrated forecast probabilities; for example, at a given bin where training set forecast probabilities deviate substantially from the bin average r_k^{train}. Specification of bin widths is suggested using, for example, t-tests for differences in sub-bin averages.

- interpretation of linear regression recalibration algorithms is complicated by fitted lines leading to recalibrated forecast values outside the range $[0, 1]$

- the ROC curve fitting recalibration algorithm is restricted to very few degrees of freedom, leading to low variance, but possibly large bias in $\hat{\kappa}$

2.5 Forecast Information Content

Effective forecast recalibration of any forecast system requires additional information about a target system if it is to improve forecast performance. Information content can be quantified using a number of information-theoretical measures. One such measure is *relative entropy* (or Kullback-Liebler divergence) which evaluates the difference in uncertainty about a set of outcomes described by two probability distributions [113, 172, 194, 154]. Although relative entropy

can be interpreted as the "information divergence" between two probability distributions, it is not a measure of true distance because it lacks symmetry and does not satisfy the triangle inequality [35, 172]. Consider two PDFs, denoted by the vectors $\mathbf{p}$ and $\mathbf{q}$, and K possible mutually exclusive outcomes. If the jth component of these vectors represent the probability of the jth outcome occurring, then the relative entropy is given by

$$D(\mathbf{p}|\mathbf{q}) = \sum_{j=1}^{K} p_j log \frac{p_j}{q_j}. \tag{2.36}$$

where $D > 0$ if $\mathbf{p} \neq \mathbf{q}$. Relative entropy reflects the additional information required to reduce the uncertainty of $\mathbf{q}$ so that it exactly describes $\mathbf{p}$ (i.e. $\mathbf{p} = \mathbf{q}$).

Another information-theoretical measure closely related to relative entropy is called *Ignorance*, which quantifies the information content (in bits) of a single, observed outcome (see Good [63] and Section 1.6.2 for a full description). Ignorance is a scoring rule which provides a summary measure of forecast skill for a given sample of forecast-outcome pairs. The score is expressed as

$$IGN = -\frac{1}{N} \sum_{i=1}^{N} log_2 p_i(Y_i), \tag{2.37}$$

where p_i is the forecast probability assigned to the verifying outcome Y_i. The ignorance of the climatological PDF defines zero skill so that the skill of the forecast can be expressed as

$$IGN = -\frac{1}{N} \sum_{i=1}^{N} log_2 p_i(Y_i) + \frac{1}{N} \sum_{i=1}^{N} log_2 p_{clim}^{\theta}(Y_i) \tag{2.38}$$

$$= -\frac{1}{N} \sum_{i=1}^{N} log_2 \frac{p_i(Y_i)}{p_{clim}^{\theta}(Y_i)}. \tag{2.39}$$

Ignorance is a useful measure of forecast skill because it provides an additive quantification of the difference between two forecasts in bits of information [172]. The skill of the binary forecasts of the position of the Lorenz63 system x state variable described in Section 2.2 is now evaluated using ignorance. Figure 2.9 provides a sample of the relative scores of the binary forecasts for the parameter
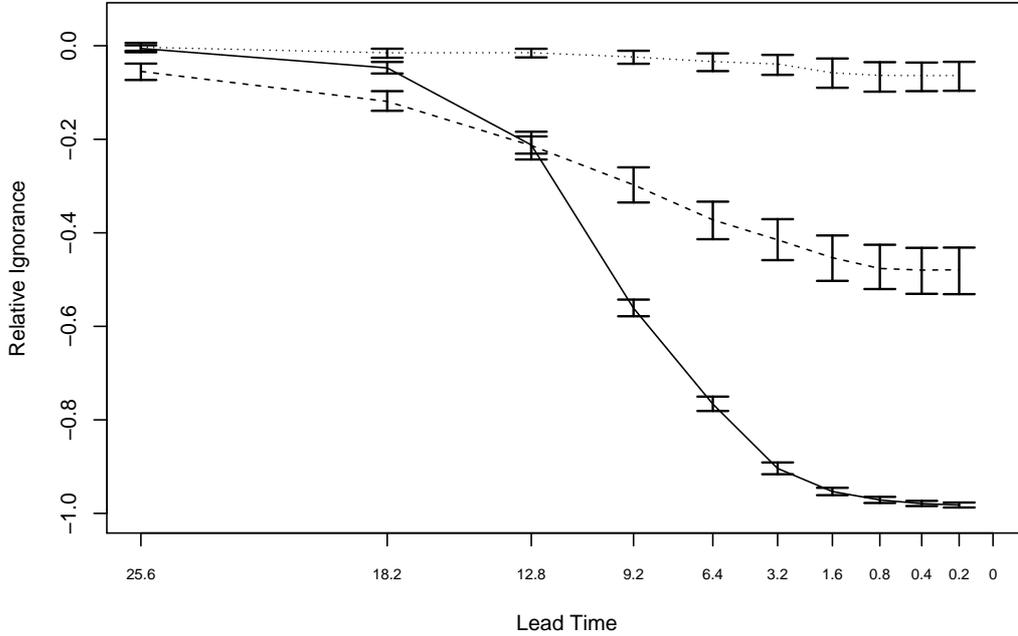
Figure 2.9: Forecast ignorance under PMS: Ignorance (with 5% - 95% uncertainty intervals) for binary forecasts at all lead times generated in Expts. 4, 5, and 6 (see table 2.1) of the observed state of the Lorenz63 x-variable lying above the climatological thresholds of $\theta = 0.5$ (solid line), $\theta = 0.9$ (dashed line), and $\theta = 0.99$ (dotted line) with increasing lead time τ. The curves slope downwards with decrease in lead time reflecting the increased predictability of the outcome and increased skill of the forecasts. Greater forecast skill is also generally achieved by binary forecasts of lower climatological threshold events.

configurations labelled as Expts. 4-6 in table 2.1. The complete set of ignorance scores of the binary forecasts constructed with each of the density construction methods for all parameter configurations under PMS are tabulated in B.3 in appendix B. From this set of results the best performing density construction method can be determined.

Firstly, however, the general effects of varying climatological event frequency, ensemble size, and forecast lead time on forecast skill are now examined. Forecast skill generally improves for lower values of θ reflecting the larger observational uncertainty (i.e. lack of sharpness) of the climatological PDF. The highest

skill at each threshold value is $IGN = -0.99$ ($\theta = 0.5$), $IGN = -0.48$ ($\theta = 0.9$), and $IGN = -0.06$ ($\theta = 0.99$) occurring at shortest forecast lead times. There is a pattern of increasing skill with increasing ensemble size although the increases in skill are relatively marginal. The differences in ignorance between the lowest ensemble size $N_{ens} = 4$ and largest ensemble size $N_{ens} = 1024$ are on the order of 10^{-2} bits for forecasts produced with all four density construction methods, and for all three thresholds. The small margin is not surprising under PMS, however, given that the level of observational noise is relatively low. Clearly, an ensemble of four members is sufficient in this case to estimate the uncertainty about the current state of the system (i.e. the initial condition uncertainty), and produce accurate forecasts. More substantial is the margin of skill between the shortest and longest lead times which is, at its greatest, of the order of 10^{-1} bits for the lowest climatological threshold value $\theta = 0.5$. In fact, the forecasts are often less skilful than the climatological forecast at longer lead times, suggesting that an ensemble forecast might be improved by deploying forecast recalibration at those lead times.

The results are tabulated in appendix B. In short, the KDB method performs the best overall, particularly for smaller ensemble sizes and longer forecast lead times. The superiority is generally marginal, however, particularly at $\theta = 0.5$ where the margin is, at greatest, 0.1 bits. The Bayesian method is not competitive at the higher thresholds, and worse than the climatological forecast (i.e. > 0 bits), but appears to perform consistently the best at $\theta = 0.5$ for larger ensemble sizes. Both the counting methods are competitive under PMS at the higher thresholds and generally where ensemble sizes tend to be larger. On the other hand, both counting methods sometimes exhibit superior forecast skill to the Bayesian and KDB methods at lower ensemble sizes. This atypical skill is attributable to less smoothness in the forecast PDFs, resulting in "lucky strikes" [185]. Of the four forecast density construction methods, only the NC method is susceptible to "forecast busts" [26] (i.e. $IGN = \infty$), since zero forecast density is placed on the same side of the threshold as the outcome. The

probability of this occurring from a perfect ensemble is $\sim \frac{2}{N_{ens}}$ [40]. Assigning a zero probability to any outcome is highly inadvisable and in violation of Cromwell's law. These forecast busts tend to occur for smaller ensemble sizes and longer lead times, and are symptomatic of the coarsely defined forecast PDFs produced from the NC method.

2.5.1 Forecast skill and forecast reliability

The relationship between forecast skill and forecast reliability in the context of forecast recalibration is now examined. There is an intrinsic link between the two forecast attributes, and forecast skill is often considered a combination of forecast reliability and a third attribute, sharpness [142, 59] (see Section 1.6.6). Murphy [136] stresses the importance of understanding the contributions of the individual components to overall skill to comprehensively assess forecast quality. Jolliffe and Stephenson [86] state that forecast reliability is not a necessary condition for forecast skill, but it is shown in Section 3.4.2 that increases in forecast skill are generally not possible without increasing forecast reliability.

Forecast evaluation with a proper scoring rule (e.g. ignorance) provides a measure of the skill of a forecast system, but does not give an indication of whether a better score is attributable to a larger relative gain of reliability or sharpness, if indeed, such a separation is sensible. Only by assessing forecast reliability can systematic biases be removed from a forecast system through recalibration [135, 159]. Such *unconditional* biases are revealed in a reliability diagram if the calibration function plot is consistently above (underforecasting) or below (overforecasting) the diagonal line (see [217]).

An alternative approach to assessing forecast reliability using reliability diagrams is to quantify it by means of an algebraic decomposition of the scoring rule into its components of reliability and sharpness. Weijs et al. [199] derive the decomposition of the ignorance score in terms of forecast reliability, forecast resolution, and uncertainty. If these additive components (defined below) are

denoted by IGN_{REL}, IGN_{RES}, and IGN_{UNC} respectively then the decomposition of the ignorance score is expressed as

$$IGN = IGN_{REL} - IGN_{RES} + IGN_{UNC}, \tag{2.40}$$

where the right hand side of the equation represents the sharpness term (i.e. $IGN_{UNC} - IGN_{RES}$). IGN_{REL} is, like IGN, negatively oriented (i.e. lower values indicate better reliability) while IGN_{RES} is positively oriented (i.e. higher values indicate better resolution). Equation (2.40) provides a useful formulation of the score whereby changes to forecast skill and reliability before and after recalibration can be quantified. A novel examination of the impact of recalibration on the performance of the binary forecasts with respect to IGN and IGN_{REL} is given in Section 2.6. Totder and Ahrens [194] explain that the decomposition is based on the conditional frequency of an event occurrence on all occasions where p_j is forecast, which is equal to the observed frequency f_j (i.e. $P(Y|p_j) = (f_j, 1 - f_j)$). Forecast reliability is described as the average relative entropy between each unique forecast distribution $(p_j, 1 - p_j)$ and the conditional observed distribution $P(Y|p_j)$ so that

$$IGN_{REL} = \sum_j P(p_j) \left[f_j log \frac{f_j}{p_j} + (1 - f_j) log_2 \frac{1 - f_j}{1 - p_j} \right]. \tag{2.41}$$

Comparison of forecast resolution before and after recalibration of the binary forecasts under IMS is presented in chapter 3 by quantifying IGN_{RES}. Authors of previous studies have noted that recalibration often leads to a decrease in forecast resolution, but Section 3.4.3 provides the first numerical evaluation of the changes in resolution.

2.6 Recalibration under PMS

A preliminary analysis of the results of recalibration on forecast skill and forecast reliability under PMS is presented in this section by comparing the performance of the Lorenz63 binary forecasts before and after recalibration. Focus is given

to how the impact of recalibration on forecast performance varies with lead time. The aim is to determine the upper bound of τ beyond which no substantive information can be gained from sampling the initial conditions to improve forecasts produced from perfect models. In short, at what lead time does the forecast PDF become no more informative than the climatological distribution? At that point, the forecast is no longer "useful" [182], and recalibration might be more beneficial for improving forecast skill.

Table 2.2: Forecast skill before and after recalibration under PMS

Recalibration (& PDF method) algorithm	θ	N_{ens}	τ*	Before recalibration		After recalibration		Difference	
				IGN	IGN_{REL}	IGN	IGN_{REL}	ΔIGN	ΔIGN_{REL}
Simple	0.5	4	0.2	-0.832	0.093	-0.930	0	-0.098	-0.093
translation			6.4	-0.625	0.122	-0.649	0.006	-0.024	-0.116
(AC)			25.6	0.136	0.121	0.001	0.003	-0.135	-0.118
	0.5	1024	0.2	-0.982	0.001	-0.846	0.014	0.136	0.013
			6.4	-0.774	0.036	-0.827	0.006	-0.053	-0.03
			25.6	-0.004	0.013	0.004	0.011	0.008	-0.002
	0.99	4	0.2	-0.058	0	-0.055	0	0.003	0
			6.4	-0.008	0.003	-0.012	0	-0.004	-0.003
			25.6	0.022	0.007	0.003	0	-0.019	-0.007
	0.99	1024	0.2	-0.065	0	-0.049	0.007	0.016	0.007
			6.4	-0.033	0	-0.024	0	0.009	0
			25.6	-0.004	0.004	0.029	0	0.033	-0.004
Logistic	0.5	4	0.2	-0.832	0.053	-0.847	0.050	-0.015	-0.003
regression			6.4	-0.625	0.076	-0.701	0.006	-0.076	-0.07
(AC)			25.6	0.136	0.121	0.001	0.002	-0.135	-0.119
	0.5	1024	0.2	-0.982	0	-0.875	0.050	0.107	0.050
			6.4	-0.774	0.036	-0.885	0.004	-0.111	-0.032
			25.6	-0.004	0.013	0	0.012	0.004	-0.001
	0.99	4	0.2	-0.058	0	-0.066	0	-0.008	0
			6.4	-0.008	0.003	0.001	0	0.009	-0.003
			25.6	0.022	0.007	0.004	0	-0.018	-0.007
	0.99	1024	0.2	-0.065	0	-0.066	0	-0.001	0
			6.4	-0.033	0	0.414	0.085	0.447	0.085
			25.6	-0.004	0.004	-0.001	0.001	0.003	-0.003
KDE	0.5	4	0.2	-0.832	0.093	-0.941	0.031	-0.109	-0.062
(AC)			6.4	-0.625	0.122	-0.670	0.009	-0.045	-0.113
			25.6	0.136	0.121	0.001	0.003	-0.118	-0.135
	0.5	1024	0.2	-0.982	0.001	-0.997	0	-0.015	-0.001
			6.4	-0.774	0.036	-0.887	0.004	-0.113	-0.032
			25.6	-0.004	0.013	-0.005	0.004	-0.001	0.009
	0.99	4	0.2	-0.058	0	-0.065	0	-0.007	0
			6.4	-0.008	0.003	-0.013	0	-0.005	-0.003
			25.6	0.022	0.007	0.003	0	-0.019	-0.007
	0.99	1024	0.2	-0.065	0.002	-0.066	0	-0.001	-0.002
			6.4	-0.033	0	-0.034	0.007	-0.001	0.007
			25.6	-0.004	0.004	0	0.002	0.004	-0.002

*in Lorenz63 seconds [118]. All values are rounded to 3 decimal places.

Samples of the numerical results of forecast recalibration using the simple translation, logistic regression, and kernel density estimation (KDE) algorithm are shown in table 2.2. These show that forecast recalibration is evidently more effective at increasing forecast skill and reliability at all lead times where ensemble sizes are smaller, achieving improvements on the order of 0.14 bits of information. The biggest increases in skill and reliability tend to occur at the longest lead times where raw forecasts are less skilful and less reliable. The KDE algorithm performs the best overall but improvements are generally achieved for the same parameter values by all recalibration algorithms. Only forecasts constructed with the AC method are shown in table 2.2, but the results for the other density construction methods (not shown) indicate that recalibration of more skilful forecasts is not as effective.

Forecast recalibration is performed and assessed in more realistic circumstances under IMS in the next chapter. The KDB method for producing binary forecasts demonstrates the best skill overall, and provides the benchmark for assessing whether forecast recalibration can be beneficial for forecast performance at longer lead times and for smaller ensemble sizes. The AC and KDB methods only are deployed in the next chapter to illustrate the impacts of forecast recalibration.

2.7 Conclusions

The evaluation and recalibration of binary forecasts under a perfect model scenario has been reviewed and examined in this chapter. A perfect model has been used to produce forecasts of the state of the Lorenz63 dynamical system using four density construction methods. The performance of these forecasts is then compared before and after forecast recalibration. A framework has been proposed for best-practice binary forecast evaluation and recalibration from the perspective of forecast skill and forecast reliability; two different but related attributes of forecast quality. Quantitative evaluation of the relative effects of

recalibration on these forecast attributes can be useful. Such an evaluation can be achieved using a decomposition, where available, of a proper scoring rule such as the ignorance score into components of reliability and resolution, along with reliability diagrams.

The task of forecast recalibration has been framed by the calibration function $\kappa(p)$ which measures the conditional probability of a binary outcome occurrence given a forecast probability p. The calibration function $\kappa(p)$ is generally unknown [23] and must be estimated from a finite sample of forecast-outcome pairs (p_i^{train}, Y_i) which are ideally i.i.d.. In general, estimation of $\kappa(p)$, and hence the efficacy of recalibration, is limited by imperfect observations of the true state of the dynamical system at hand. A comprehensive range of algorithms for estimating the calibration function and performing recalibration have been reviewed and critiqued, providing a unique perspective of the challenges of recalibration in both principle and practice. Like all estimation problems, the calibration function is subject to residual errors, which can be described with respect to bias and variance. These errors originate from several sources such as non-independent forecast-outcome pairs, and under-sampling and specification of probability bins where binning and averaging the forecasts. A balance between bias and variance of the calibration function is ideally identified, but in practice the trade-off is typically non-trivial.

Information theoretical measures of forecast performance employed to assess the effect of recalibration have also been introduced and discussed. Measures such as relative entropy and ignorance are appropriate for evaluating recalibration, both because they have ideal properties, and because each can be decomposed into attributes of forecast quality such as reliability and resolution. Hence, the effect of recalibration on forecast performance can be assessed with respect to these attributes as well as forecast skill. These measures contribute to the novel evaluation framework introduced in this thesis for investigating the impact of recalibration.

Finally, in Section 2.6, forecast recalibration has been demonstrated on fore-

casts constructed with the AC density construction method. One of the key aims of the examination was to identify for which conditions and forecast-parameters is forecast recalibration effective. It has also been enquired whether, given that a perfect model is structurally correct and only suffers from initial condition uncertainty, improving forecast skill is effective through recalibration, or by increasing ensemble size. It has been determined that recalibration is most effective at longer lead times particularly for smaller ensemble sizes and where the climatological probability of the binary event is closer to 0.5 (i.e. $\theta \to 0.5$), but can improve both forecast skill and forecast reliability at lead times as short as 0.2 Lorenz63 seconds for the smallest ensemble sizes at $\theta = 0.99$. Improvements in forecast skill are generally accompanied by increases in forecast reliability as measured by the decomposition of the ignorance score, but not exclusively so. Recalibration has also been found to be less effective where raw forecast skill is high, as demonstrated by the KDB density construction method.

The following novel contributions or innovations in this chapter include:

- critique of existing recalibration algorithms for binary probabilistic forecast recalibration

- identification of the challenges of forecast recalibration in principle and in practice

- deployment of a range of recalibration algorithms to assess their respective effectiveness for improving forecast performance

- examination of the relationship between forecast skill and forecast reliability in the context of recalibration using the decomposition of the ignorance score

- investigation of the conditions where recalibration is effective for increasing forecast reliability and forecast skill in a perfect model scenario

Chapter 3

Forecast Evaluation and Recalibration under IMS

In a world of perfect models with known parameters, forecast model error originates purely from uncertainty in the true state of a system at the point at which a model is initialised. Initial condition (IC) uncertainty inhibits both correctly determining the exact current state of the system, and making accurate predictions of its future state. In the real world, all models are imperfect and are subject to both observational uncertainty and structural imperfections. The latter source of model inadequacy is an unavoidable consequence of an incomplete understanding of the dynamics of the modelled system [90]. The differences between a *perfect model scenario* (PMS) and an *imperfect model scenario* (IMS) are explicated in Section 1.4.

Chapter 2 presented a novel investigation into the evaluation of binary forecast skill and reliability, and the effectiveness of recalibration of binary forecasts given a perfect model. The investigation demonstrated that recalibration leads to improved forecast skill and reliability at longer forecast lead times and climatological probabilities are closer to 0.5 where predictability is limited by uncertainty in the state of a dynamical system, and IC uncertainty which is larger for smaller model ensemble size. In other words, it was shown that there is a

greater potential for forecast improvement after recalibration where forecasts have less skill before recalibration.

In this chapter, the same line of inquiry into forecast evaluation and recalibration is taken, but under IMS. The conditions are surveyed where forecast recalibration is effective, and where resources should be instead dedicated to improving forecast techniques. Following the results of Chapter 2, the expectation is that, since the performance of imperfect model forecasts is worse than that of a perfect model, recalibration would be of greater value for improving performance.

This chapter is structured as follows: imperfect model inadequacy is described in Section 3.1 along with the design of the imperfect forecast model employed to produce binary forecasts of the Lorenz63 system state. The binary forecasts are constructed from raw imperfect model ensemble output using the same four different density construction method as in Chapter 2. This novel comparison of the skill of these forecast systems is presented in Section 3.2. The investigation of binary forecast performance under PMS and IMS has also led to the discovery of an interesting property of the ignorance score which is discussed in Section 3.2. Surprisingly, the relative forecast skill of binary forecasts produced from perfect models compared to those produced from imperfect models can be marginal, even for perfect forecasts (i.e. $p(Y) = 1$); it is shown that limit is dependent on the climatological probability.

An overview of forecast performance before and after recalibration under both PMS and IMS is first provided in Section 3.4. Comparisons of forecast performance under the two scenarios reveal that forecast recalibration is more beneficial for imperfect model forecasts with smaller ensemble sizes and longer lead times, and where climatological probability closer to 0.5. While not surprising, this is the first quantitative demonstration of the effect.

Finally, in Section 3.4, forecast recalibration is demonstrated on binary forecasts; the two best performing forecast models of Section 3.2 are used. All of the recalibration algorithms outlined in Chapter 2 are employed. As already

explained in Chapter 2, forecast recalibration is performed by computing the conditional distributions of the outcomes given a set of forecasts to estimate the calibration function κ, and then making corrections to forecast probability values p according to the calibration function estimate $\hat{\kappa}(p)$. In practice, categorisation or binning of the forecasts inherently leads to errors in estimates resulting from both bias and variance in $\hat{\kappa}(p)$. Various approaches to forecast categorisation are described, with those achieving a balanced bias-variance trade-off being most ideal. Again, a novel approach evaluation of forecast performance before and after recalibration with respect to both forecast skill and forecast reliability is performed. The results confirm that forecast recalibration is most effective where raw forecast skill is poorer, that is, generally for longer lead times, smaller ensemble size, and higher climatological event uncertainty. It is concluded that forecast recalibration provides a useful technique for improving poorly performing forecast systems, and should be considered as a simple and cost-effective first option. This recalibration experiment is, to the author's knowledge, the first of its kind in the published literature.

3.1 Challenges of model inadequacy

There is arguably no such thing as a perfect model for any physical dynamical system in the real world [90]. There are always imperfections in the mathematical structure of forecast models, not merely in their estimation of initial conditions. These imperfections are attributable to different forms of model inadequacy which are now described. Model inadequacy refers to a model's inability to simulate the trajectory of a system's state, even given precise initial conditions [96]. *Model error* [88] arises as a result of the model formulation containing an incomplete mathematical description of the system dynamics (i.e. *structural error*), perhaps due to an absence of sub-space where a component of the system's dynamics is not included in the model [90]. Ignored sub-space inadequacy features in numerical weather models, for example, where the model

Raw Model Ensemble
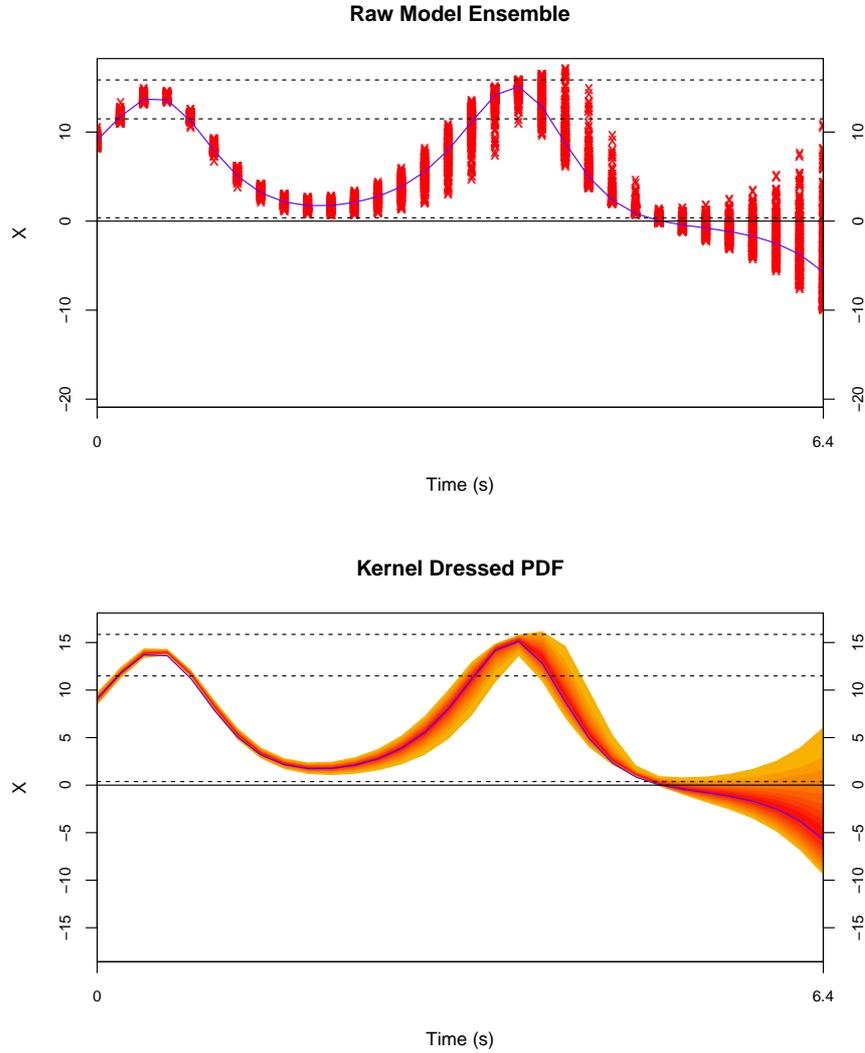
Kernel Dressed PDF

Figure 3.1: Ensemble forecasting under IMS: raw imperfect Lorenz63 model ensemble generated in Expt. 1 (see table 3.1) (top), and fan chart showing the kernel dressed ensembles (PDFs) constructed from the raw ensembles shown in the upper plot at every time step from $t = 0$ up to $t = t_\tau = 6.4s$ (bottom). The PDF represents the probabilities of the system's state and the blue trajectory shows the actual true state at a given time t. See Fig. 2.1 for further details.

variables represent the physical variables of the weather system on a *grid-box* discretisation of model space [153]. These types of computer models are unable to resolve sub-grid processes [150]. Model inadequacy leads inevitably to *forecast error* (see Section 1.6) in which the projected state of the model is different

Forecast PDF at lead time tau=6.4
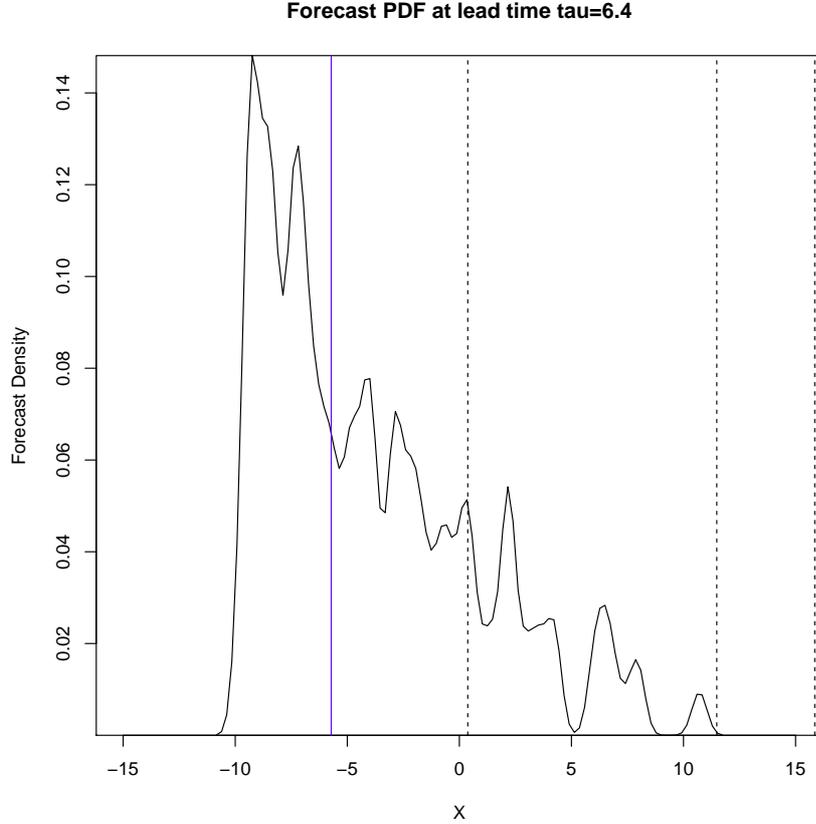
Figure 3.2: Ensemble forecasting under IMS: kernel dressed ensemble (PDF) with $N_{ens} = 256$ corresponding to that in Fig. 3.1 at $t_\tau = 6.4s$. The true state of the system variable is shown as a blue line at $\tilde{x}_{t_\tau} = -5.7$, and the dashed horizontal lines denote the 50th, 90th, and 99th percentiles of the climatological distribution representing the climatological event thresholds $\theta \in \{0.5, 0.9, 0.99\}$. Given that $\tilde{x}_{t_\tau} < x_\theta$, and most of the probability density is below $x_\theta = 0.37$, the forecast is more skilful than a climatological forecast $p^\theta_{clim} = 0.5$, although not as skilful as the perfect model forecast in the equivalent plot in Fig. 2.2.

to the actual system state at a given lead time (see also section 1.5). Unlike PMS, it may not be possible to isolate the effect of modification of a forecast system on its performance to a single property of that system (see 2.1) because forecast error stems from both IC uncertainty and model inadequacy under IMS. It should be possible, however, to at least identify the properties of a forecast system where it is more effective to recalibrate rather than to improve forecast techniques in a real world scenario. The Lorenz63 ensemble model described

Table 3.1: Configurations for IMS Lorenz63 binary forecast experiments

Experiment No.	Dynamical Equations	State Variable	Observational Noise	PDF Method	Forecast Parameters		
					θ	N_{ens}	τ^*
1	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.5	256	6.4
2	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	All	0.9	All	25.6
3	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	AC	0.5	4	25.6
4	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.99	4	0.2
5	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	KDB	0.5	1024	18.2
5	Lorenz63	x	$\mathcal{N}(0, 0.37^2)$	AC	0.5	4	18.2

*in Lorenz63 seconds [118]; see Appendix A.1.

in Chapter 2 is again employed along with the naive counting (NC), adjusted counting (AC), Bayesian, and kernel dressing and blending (KDB) forecast density construction methods (see Section 2.2) to examine forecast evaluation and forecast recalibration under IMS. Structural imperfection is introduced into the Lorenz63 ensemble model by substituting the x state variable in the system's differential equations (see Appendix A.1) so that

$$x' = csin\left(\frac{x}{c}\right), \tag{3.1}$$

where x' is the imperfect model variable. In these examples $c = 16$. Figure 3.1 shows example plots of the model ensemble and kernel dressed ensemble iterated forward in time up until lead time τ corresponding to Fig. 2.1 in Chapter 2. Figure 3.2 shows the forecast PDF, and the true state of the variable $\tilde{x}_{t_\tau}$ at lead time $\tau = 6.4s$.

3.2 Which forecast system is best?

The performance of forecasts of the Lorenz63 system state under IMS are assessed and compared in this section. The assessment and comparison of binary forecasts constructed each of the four density construction methods is a new contribution. The forecasts are evaluated against climatological forecast ($p_{clim}^\theta = \theta$). Figure 3.3 provides a sample of the ignorance scores of forecasts con-

structed from all four methods in Lorenz63 Expt. 2 (see table 3.1). The curves show that the more skilful forecasts are produced from the KDB, AC, and NC density construction methods with forecast-parameters $\theta = 0.9$ and $\tau = 25.6s$. In general, the KDB and AC methods are also found to be generally better than the NC and Bayesian methods over the whole range of forecast-parameters (the full set of forecast skill results are tabulated in table B.4). Superior forecast skill is predominantly demonstrated by the KDB method at longer lead times and smaller ensemble sizes. It should be noted that the degree of relative skill
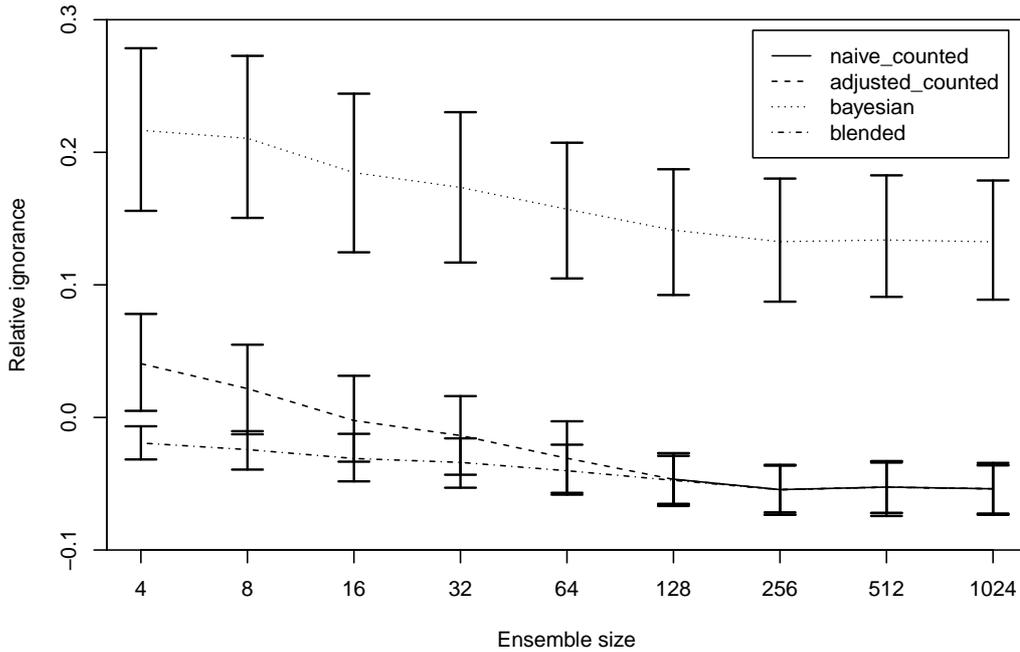
Figure 3.3: Forecast ignorance under IMS: ignorance (with 5% - 95% uncertainty intervals) for binary forecasts produced from the NC (solid line), AC (dashed line), Bayesian (dotted line), and KDB (dash-dotted line) density construction methods under Expt. 2 (see table 3.1) with $\theta = 0.9$, $\tau = 25.6s$ and all ensemble sizes. The KDB method performs best at smaller ensemble sizes, and is equalled in skill for ensemble sizes $N_{ens} \geq 128$ by the NC and AC methods. Note: there is no curve for the NC method where $N_{ens} < 128$ because it produces forecast busts at these ensemble sizes.

of the KDB imperfect model compared to the other methods is quite marginal, however, and that the AC and NC methods are actually competitive with the KDB method for larger ensemble sizes, particularly at the highest climatological event frequency $\theta = 0.5$. The margin of relative skill is 0.15 bits of information at most between the KDB and the counting methods, and typically close to zero at higher ensemble sizes. The Bayesian method produces marginally the most skilful forecasts at shorter lead times and larger ensemble sizes at the highest climatological event frequency with relative skill of up to 0.05 bits.

The overall superior performance of the KDB method is attributable to the unique forecast *post-processing stage* (see Section 1.5) which employs the following additional steps after constructing the raw model ensemble to improve the quality of the forecast:

1. the kernel dressed ensemble is translated into a smoothed and continuous PDF providing a more precise estimate of the underlying distribution [214]

2. the kernel dressed ensemble is blended with the climatological PDF

3. the blending parameter α and kernel width parameter σ are optimised using a training set of observations before a binary forecast is issued

The optimisation of the KDB parameters α and σ serves to correct systematic forecast error in the raw ensemble. Blending the dressed ensemble with the climatological PDF improves forecast performance at longer lead times where the imperfect model reaches its own limit of predictability, that is, the lead time beyond which the forecast is no longer useful. After this limit, an imperfect model is systematically unable to simulate the system's trajectories, and the ensemble converges onto the imperfect model's own climatology. All of the three reasons given above explain how forecast error can be reduced and forecast skill increased when employing the KDB method. The NC method produces forecasts which, like their perfect model counterparts, tend to bust (i.e. $IGN = \infty$) for smaller ensemble sizes and longer lead times, but over an

increased range of these two forecast-parameters. In fact, even where the model ensemble is perfect (see Section 1.5.2), the probability of forecast busts with the NC method is $\sim \frac{2}{N_{ens}}$ [40]. The AC method is designed to avoid busts by artificially adding probability density to both binary outcomes. Conversely, an NC binary forecast might achieve a lucky strike if the ensemble members are positioned all the same side of the threshold as the outcome at time t_τ. This would appear plausible for smaller values of τ given that an insufficient amount of time has elapsed for the trajectory of the system to differ significantly from the trajectories of the ensemble. As discussed in in Chapter 2, forecast busts and lucky strikes risk violation of Cromwell's law, and should otherwise be avoided. Examination of the Lorenz63 forecast evaluation results in Chapters 2 and 3 has revealed a previously unreported property of the ignorance score for binary forecasts. Even in cases where a density construction method is superior, the difference in ignorance is always marginal. In fact, the relative loss of skill of the imperfect model forecasts compared to their perfect model counterparts is also minimal (compare tables in appendices B.3 and B.4 in appendix B) for all construction methods, particularly at shorter lead times. There is evidently a limit to the maximum skill of a binary forecast relative to the climatological forecast which is dependent on the climatological event frequency θ. This property of limited maximum relative skill can be explained by examining the ignorance score for a forecast system where the perfect forecast is always issued for each binary outcome, that is

$$IGN_{opt} = \sum_{j=1}^{2} -p_{clim}^{\theta} log_2 \left(\frac{p_j}{p_{clim}^{\theta}} \right) \tag{3.2}$$

$$= -p_{clim}^{\theta} log_2 \left(\frac{1}{p_{clim}^{\theta}} \right) - (1 - p_{clim}^{\theta}) log_2 \left(\frac{1}{1 - p_{clim}^{\theta}} \right) \tag{3.3}$$

$$= p_{clim}^{\theta} log_2(p_{clim}^{\theta}) + (1 - p_{clim}^{\theta}) log_2(1 - p_{clim}^{\theta}), \tag{3.4}$$

where p_j is the perfect forecast, and p_{clim}^{θ} is the climatological forecast. If the climatological probabilities are plugged into Eqn. (3.4) the maximum expected gain in relative skill of a forecast over a climatological forecast is $IGN_{opt} = -1.0$

for $\theta = 0.5$, $IGN_{opt} = -0.47$ for $\theta = 0.9$, and $IGN_{opt} = -0.08$ for $\theta = 0.99$. So, as the climatological probability of the binary event decreases so does the skill of the forecast, and in the limit of the climatological probability approaching zero, the optimal ignorance approaches 0, that is

$$\lim_{\theta \to 1} IGN_{opt} = 0. \tag{3.5}$$

Eqn. (3.5) implies that a perfect forecast can only demonstrate a very small degree of skill relative to a climatological forecast where a binary event has low probability of occurrence. This margin of skill is much smaller than what can be achieved by a forecast in a continuous outcome scenario, where the margin can approach infinity. The margin of difference between values of IGN_{opt} and empirical ignorance can at least quantify how close a forecast is to being perfect, however.

In the examples of Chapter 2, recalibration is shown to be most effective at longer lead times, smaller ensemble sizes, and higher climatological event frequencies. These ranges of forecast-parameter values are targeted in Section 3.4 whilst also employing the KDB and AC density construction methods, which produce the better binary forecasts overall under IMS, to assess the effectiveness of forecast recalibration. With poorer raw forecast skill exhibited by the imperfect models compared to their perfect model counterparts, the expectation is for recalibration to lead to larger improvements in forecast skill. Also of interest are shorter lead times and larger ensemble sizes at $\theta = 0.5$ where the Bayesian method, usually the poorest at higher values of θ, produced the most skilful forecasts. The failure of the recalibrated forecasts constructed using the KDB and AC methods to outperform those raw Bayesian forecasts would suggest that improving the density construction method appears more beneficial than recalibration for those forecast-parameter values.

3.3 Comparison of recalibration under PMS and IMS

In Section 2.6, forecast recalibration was found to be most effective at improving the performance of perfect model forecasts at longer lead times, with smaller ensemble sizes, and where $\theta \to 0.5$. The expectation is for recalibration to be even more effective under IMS. Table 3.2 summarises the results of the recalibration exercise under PMS and IMS for comparison. The raw and recalibrated forecast scores, and their differences are shown for both PMS and IMS. The imperfect model exhibits inferior forecast skill to the counterpart perfect model over the entire range of forecast-parameter values. The inferiority in performance is indicative of the impaired ability of the model to simulate the trajectory of the system's state over time under IMS. The increases in forecast skill are predominantly larger, although marginal, under IMS suggesting that the performance of forecasts produced from imperfect models can benefit more from forecast recalibration.

One of the forecast-parameter configurations of interest for investigating the effectiveness of recalibration is $\theta = 0.5$, $N_{ens} = 4$, and $\tau = 25.6s$ since forecast skill is generally the poorest under both PMS and IMS for this configuration. There is an increase in skill of $IGN = 0.007 - 0.195 = -0.188$ bits after recalibration of the AC forecasts employing the simple translation method. A similar gain in skill is attained with the KDE recalibration algorithm. These are the largest increases in forecast skill of all the parameter configurations (NB: not all are shown in table 3.2) indicating that recalibration is indeed more beneficial where raw forecast skill is poorer. Recalibration is less effective where raw forecast skill is already high, as is the case for forecasts constructed from the KDB method. For example, for parameters $\theta = 0.5$, $N_{ens} = 4$, and $\tau = 0.2s$ the difference is $IGN = -0.846 - -0.979 = 0.133$ bits, a decrease in skill.

Table 3.2: Forecast skill before and after recalibration under IMS

Recalibration (& PDF) method	θ	N_{ens}	τ^*	PMS raw IGN	PMS recalib. IGN	PMS diff ΔIGN	IMS raw IGN	IMS recalib. IGN	IMS diff ΔIGN
Simple translation (AC)	0.5	4	0.2	-0.832	-0.930	-0.098	-0.832	-0.930	-0.098
			25.6	0.136	0.001	-0.135	0.195	0.007	-0.188
	0.5	1024	0.2	-0.982	-0.846	0.136	-0.982	-0.846	0.136
			25.6	-0.004	0.004	0.008	-0.003	0.008	0.011
	0.99	4	0.2	-0.058	-0.055	0.003	-0.057	-0.054	0.003
			25.6	0.022	0.003	-0.019	0.012	0.018	0.006
	0.99	1024	0.2	-0.065	-0.049	0.016	-0.062	-0.049	0.013
			25.6	-0.004	0.029	0.033	-0.006	0.010	0.016
Simple translation (KDB)	0.5	4	0.2	-0.980	-0.846	0.134	-0.979	-0.846	0.133
			25.6	0.016	0.001	-0.015	0.031	0.005	-0.026
	0.5	1024	0.2	-0.983	-0.839	0.144	-0.983	-0.846	0.137
			25.6	-0.004	0.006	0.010	-0.003	0.011	0.014
	0.99	4	0.2	-0.064	-0.044	0.020	-0.061	-0.041	0.020
			25.6	0.007	0.002	-0.005	0	0.018	0.018
	0.99	1024	0.2	-0.065	-0.049	0.016	-0.062	-0.047	0.015
			25.6	-0.004	0.051	0.055	-0.006	0.009	0.015
Kernel density estimation (AC)	0.5	4	0.2	-0.832	-0.941	-0.109	-0.832	-0.941	-0.109
			25.6	0.136	0.001	-0.135	0.195	0.008	-0.187
	0.5	1024	0.2	-0.982	-0.997	-0.015	-0.982	-0.997	-0.015
			25.6	-0.004	-0.005	-0.001	-0.003	-0.005	-0.002
	0.99	4	0.2	-0.058	-0.065	-0.007	-0.057	-0.060	-0.003
			25.6	0.022	0.003	-0.019	0.012	0.014	0.002
	0.99	1024	0.2	-0.065	-0.066	-0.001	-0.062	-0.066	-0.004
			25.6	-0.004	0	0.004	-0.006	0	0.006
Kernel density estimation (KDB)	0.5	4	0.2	-0.980	-0.977	0.003	-0.979	-0.977	0.002
			25.6	0.016	0.002	-0.014	0.031	0.005	-0.026
	0.5	1024	0.2	-0.983	-0.997	-0.014	-0.983	-0.997	-0.014
			25.6	-0.004	-0.005	-0.001	-0.003	-0.005	-0.002
	0.99	4	0.2	-0.064	-0.066	-0.002	-0.061	-0.065	-0.004
			25.6	0.007	0.001	-0.006	0	0	0
	0.99	1024	0.2	-0.065	-0.066	-0.001	-0.062	-0.066	-0.004
			25.6	-0.004	0	0.004	-0.006	0	0.006

*in Lorenz63 seconds [118]. All values are rounded to 3 decimal places.

3.4 Recalibration under IMS

A more complete illustration of forecast recalibration under IMS is presented throughout the remainder of this section from the perspective of both forecast skill and forecast reliability. Comparisons of the relative performances of the recalibration algorithms (see Section 2.4.1) are provided. The key forecast-parameter values, identified in Section 3.2 above (i.e. longer lead times, smaller ensemble sizes, and higher climatological event frequencies), for which recalibration is effective are targeted to glean important insights into the effectiveness of forecast recalibration. Each forecast-parameter configuration is listed in table 3.1.

Recall from Chapter 2 that forecast reliability can be expressed both graphically on a reliability diagram, and mathematically as a component of the algebraic decomposition of the ignorance score (i.e. IGN_{REL}), as in Eqn. (2.41). While reliability diagrams plotted on probability paper [25] provide some quantitative evaluation of forecast reliability, IGN_{REL} provides a numerical measure of forecast reliability, or loss of information due to miscalibration in bits of binary information [194]. Still, both reliability diagrams and IGN_{REL} should be employed for evaluating forecast reliability.

An example of changes to forecast reliability after recalibration is given in Figs. 3.4 and 3.5 where the simple translation algorithm has been used to recalibrate forecasts constructed with the AC method in Expt. 3 (see table 3.1). The change in reliability of the forecasts is evident with the two recalibrated forecast bins lying within the 1% - 99% consistency bars whereas only one out of the four raw forecast bins did so prior to recalibration. The numerical values of the reliability component of ignorance before and after recalibration, $IGN_{REL} = 0.178$ and $IGN_{REL} = 0.007$, support the visual evidence in the reliability diagram. Clearly, recalibration has been effective in this particular recalibration experiment. Figures 3.4 and 3.5 also highlight a challenge of forecast recalibration which was raised in Section 2.4.1. This challenge arises when

partitioning forecast values into bins so that recalibration can be performed, and reliability diagrams can be plotted. An algorithm which partitions the training set of forecast-outcome pairs denoted by

$$T := \{(p_i^{train}, Y_i); i = 1, \ldots, N\}, \tag{3.6}$$

into equally populated bins has been employed throughout this thesis. The raw forecasts p^{raw} and recalibrated evaluation forecasts p_i^{re} are then partitioned according to the same bins. Forecast recalibration has clearly resulted in all forecast values being adjusted to within the range of the two central bins defined by $[0.3, 0.5]$ and $[0.5, 0.7]$. The result is larger bin populations at those two central bins but zero populations at the other bins, and a calibration function estimate which has less degrees of freedom, and may be biased [23].

The *bias-variance* trade-off of the calibration function estimate may vary between the training, raw, and recalibrated forecasts (see Section 2.4.1), making equitable comparisons of forecast reliability difficult. A numerical investigation of the effects of recalibration on the bias-variance of the calibration function is beyond the scope of this thesis (see Bröcker [23] for more information). Before proceeding further with the assessment of forecast recalibration under IMS, however, a novel discussion of the limitations of forecast binning/categorisation, and review and critique of binning/categorisation methods in the literature is now presented.

3.4.1   Binning methodology

Several recalibration algorithms employed in this thesis require that forecasts are partitioned into exhaustive and mutually exclusive bins B_k. Several authors [8, 23] have studied the effects of forecast categorisation on forecast reliability when either using reliability diagrams as an evaluation tool, or estimating the calibration function to measure forecast reliability. The categorisation or binning problem where forecast values exhibit large deviations from the bin average r_k is briefly discussed in Section 2.4.1. An example was presented demonstrating

Reliability Diagram
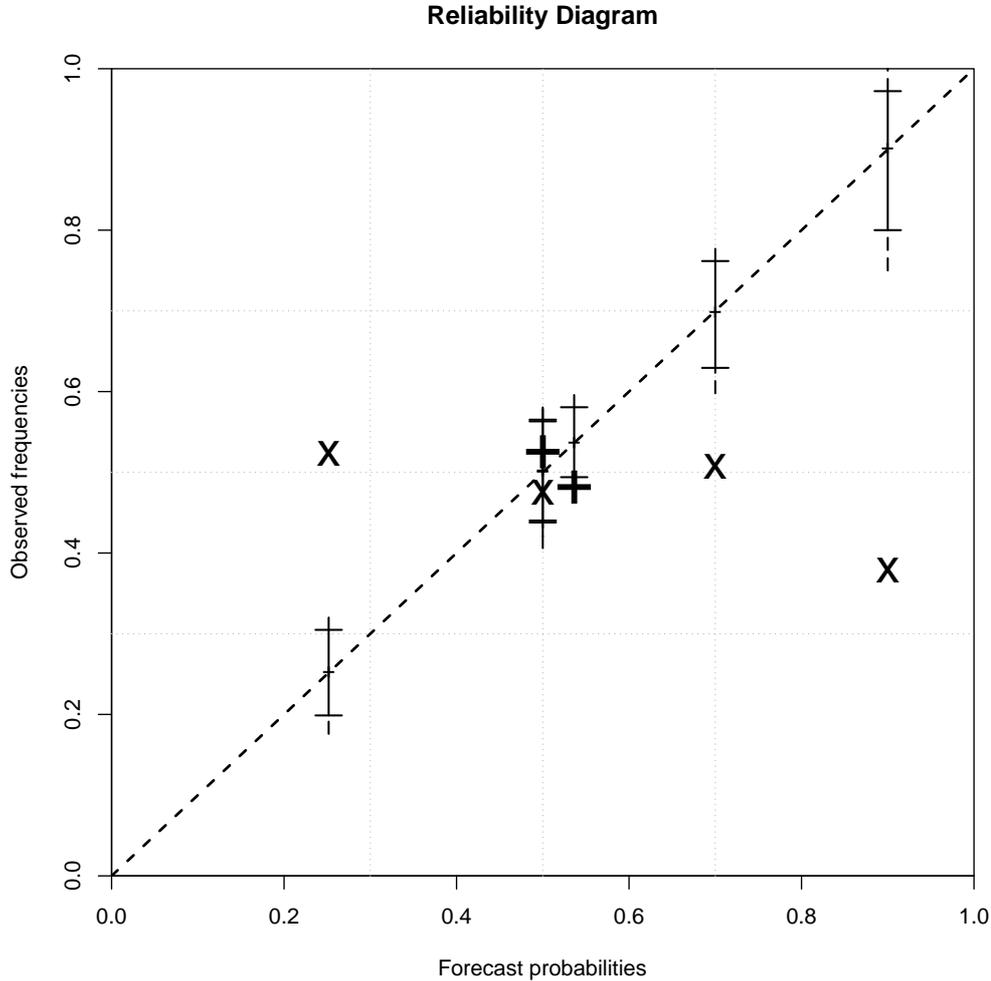
Figure 3.4: Forecast reliability after recalibration: An example reliability diagram showing the changes in reliability of the raw set (crosses) and recalibrated evaluation set (pluses) of AC forecasts using the simple translation algorithm. The recalibrated forecasts appear to be more reliable than the raw forecasts; this is supported by the numerical values of the reliability component of ignorance before and after recalibration are $IGN_{REL} = 0.178$ and $IGN_{REL} = 0.007$. All sets of forecast-outcome pairs are generated under Expt. 3 (see table 3.1)

how these deviations lead to uncalibrated out-of-sample forecast judgements. In fact, partitioning forecast values into bins that are too wide, so that the bin populations $\#I_k$ are sufficiently large, may reduce sampling error, but can result in "under-utilisation" of the forecast information. As already explained in Section
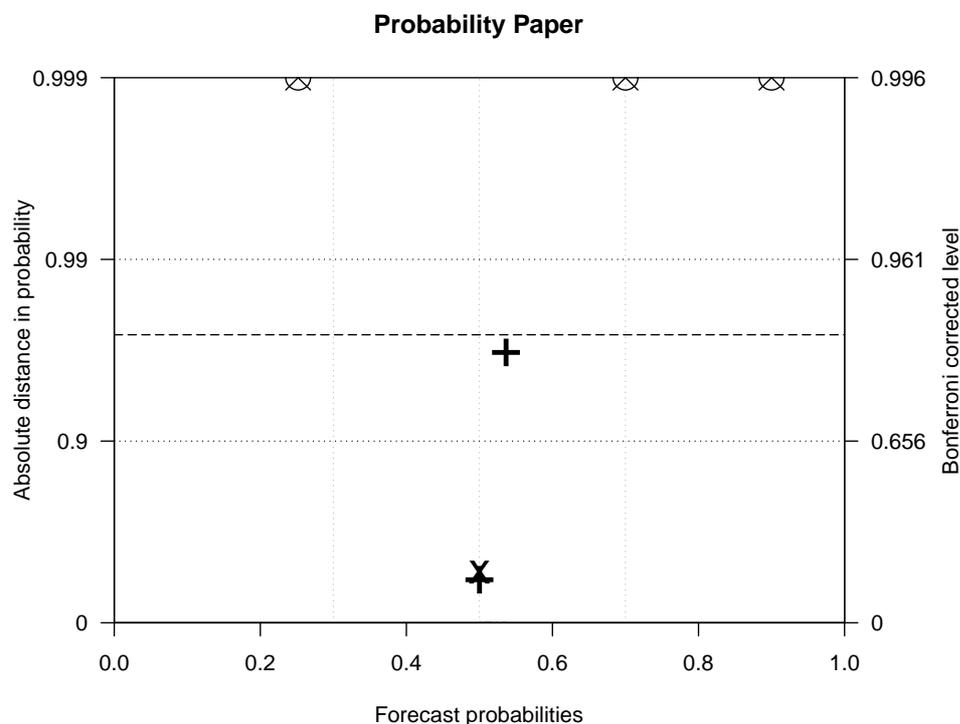
Probability Paper

Figure 3.5: Forecast reliability after recalibration: Reliability diagram on probability paper showing the changes in reliability of the raw set (crosses) and recalibrated evaluation set (pluses) of AC forecasts using the simple translation a with 5% - 95% (1% - 99% vertical dashed line) consistency bars. The recalibrated forecasts are clearly more reliable than the raw forecasts. Only one out of four raw forecast bins falls within the Bonferroni corrected 0.99 probability distance (upper dotted) band, indicating an unreliable forecast before recalibration. All sets of forecast-outcome pairs are generated under Expt. 3 (see table 3.1).

2.4.1, under-utilisation of the information contained in the joint distribution of forecasts and outcomes will impede robust forecast evaluation, and hence, forecast recalibration. On the other hand, partitioning forecast values so that there are too few in each bin may result in an excessive influence of each sample on the calibration function estimate $\hat{\kappa}$. Estimation of the calibration function is likely to contain error, reflected in a poorer ignorance score. Put simply, smaller bin populations (under-sampling) generally lead to increased variance whereas larger bin populations (over-sampling) generally lead to increased bias.

This result is not without exception, however, as a forecast system that always issues the climatological probability p_{clim}^{θ} will correspond to a perfectly reliable forecast (i.e. zero bias and variance). In that case, however, forecast reliability comes at the sacrifice of forecast resolution (see Section 3.4.3), since the forecast system is not able to discriminate between scenarios where the event occurs at other than the climatological frequency. In general, however, specifying the bins to strike the right balance between the two undesirable scenarios (i.e. under-sampling and over-sampling) represents achieving a balanced bias-variance trade-off of the calibration function estimate $\hat{\kappa}$. Fortunately, a good balance can be checked for *ex post* by quantifying the effect on a scoring rule of varying bin specifications (see Bröcker [23]). There are several possible methods for specifying the bins on a reliability diagram which are now discussed (see table 3.3 for a listing of the binning methods).

Method No. 1: Fixed bin width

The most straightforward bin specification method is to pre-determine the number of bins so that the unit interval is divided into fixed, equal intervals [152, 131]. The specification of the number of bins is somewhat arbitrary, however, which can lead to an imbalanced bias-variance trade-off of $\hat{\kappa}$. In general, forecast values are not distributed uniformly over the unit interval where forecast PDFs are sharp (i.e. the distribution of forecasts is likely to be heavily skewed towards lower and/or higher probability values), in which case, some bins are more likely to suffer from under-sampling [8].

Method No. 2: Equi-probable bins

An ideal bin specification method is to partition the forecast values into equi-probable bins so that the data are equally represented in all bins across the unit interval. This method may yield widely varying bin interval widths, but is generally more robust to under-sampling, and achieves a better bias-variance

trade-off than the fixed, equal bin width method. The equi-probable bin method can be limited by non-uniformity of the distribution of forecast values over the unit interval, however. Such a limitation often arises when forecasting the state of a nonlinear dynamical system such as Lorenz63. Sensitivity to initial conditions can lead an imperfect yet reliable model to produce non-uniform distributions of binary forecasts, depending on location on the Lorenz63 *attractor* [118]. Figure 3.6 shows an example of a reliability diagram containing binary forecasts produced from the imperfect Lorenz63 model partitioned into 2 bins (the lowest bin is very narrow $[0, 5.8 \times 10^{-13}]$) with the forecast values being concentrated at a single, very low probability value $p_1 = 5.8 \times 10^{-13}$ so that $\#I_1 = 496$. The much smaller population (i.e. $\#I_2 = 16$) of the second bin defined by the boundaries $[5.8 \times 10^{-13}, 1.0]$ means that the consistency bar interval is wide, and the estimate of the calibration function at that bin $\hat{\kappa}(r_2)$ is subject to large variance. Recalibration of the forecasts in this bin using the calibration function estimate may lead to decrease in forecast reliability and skill. The equi-probable binning algorithm is a straightforward approach that generally minimises any bias-variance trade-off, however, compared to, for example, the fixed bin width method (see Palmer et al. [152]), and is employed in this thesis.

Method No. 3: Binomial distribution sampling

A relatively simple binning method is described by Atger [8] whereby the observed frequency f_k, corresponding to a given forecast probability value p_k, is assumed to follow a binomial distribution with parameters $\#I_k$ and p_k. The binomial assumption is the same basis for determining the consistency bars on a reliability diagram as outlined by Bröcker and Smith [24], and reported in Section 1.6.4. The expected sampling variance of the observed frequency is a function of the bin population $\#I_k$. Hence, $\#I_k$ is determined so that f_k falls within a specified consistency bar interval. Depending on the reliability of the forecasts p_k, however, $\#I_k$ may have to be reduced to the point where the

Table 3.3: Reliability diagram bin specification methods

Method No.	Method Name	Method Description
1	Fixed bin width	Pre-specified number of bins
2	Equi-probable bins	Equal bin populations
3	Binomial sampling	Determine $\#I_k$ so that f_k falls within consistency bars
4	Bin merging	Two bins are merged if resulting f_k is not significantly different (see Atger [8])
5	Score optimisation	Specify bins after score optimisation

variance of $\hat{\kappa}$ becomes large, potentially leading to poorly recalibrated forecasts.

Method No. 4: Bin merging

A second, more complicated method proposed by Atger [8] involves optimising the forecast bin specification by merging two forecast bins, say B_k and B_{k+1}, to make a bin $B_{k'}$ if the observed frequency $f_{k'}$ is not significantly different to f_k. A resampling procedure is used to test the significance of the difference.

Method No. 5: Score optimisation

Bröcker [23] considers the estimation of the calibration function $\kappa(p)$ as an ill-posed problem due to the dependence of the calibration function on the size $\#I_k$ of each bin B_k. To address this problem, Bröcker [23] suggests that $\#I_k$ can be determined by way of a *regularisation* parameter δ which controls the degrees of freedom of the calibration function estimator $\hat{\kappa}(p; T, \delta)$ (i.e. the bin diameters) to balance the bias and variance of the calibration function estimate. The bandwidth parameter of the kernel dressing estimation recalibration method outlined in Section 2.4.1 is an example of the regularisation parameter.

Reliability Diagram
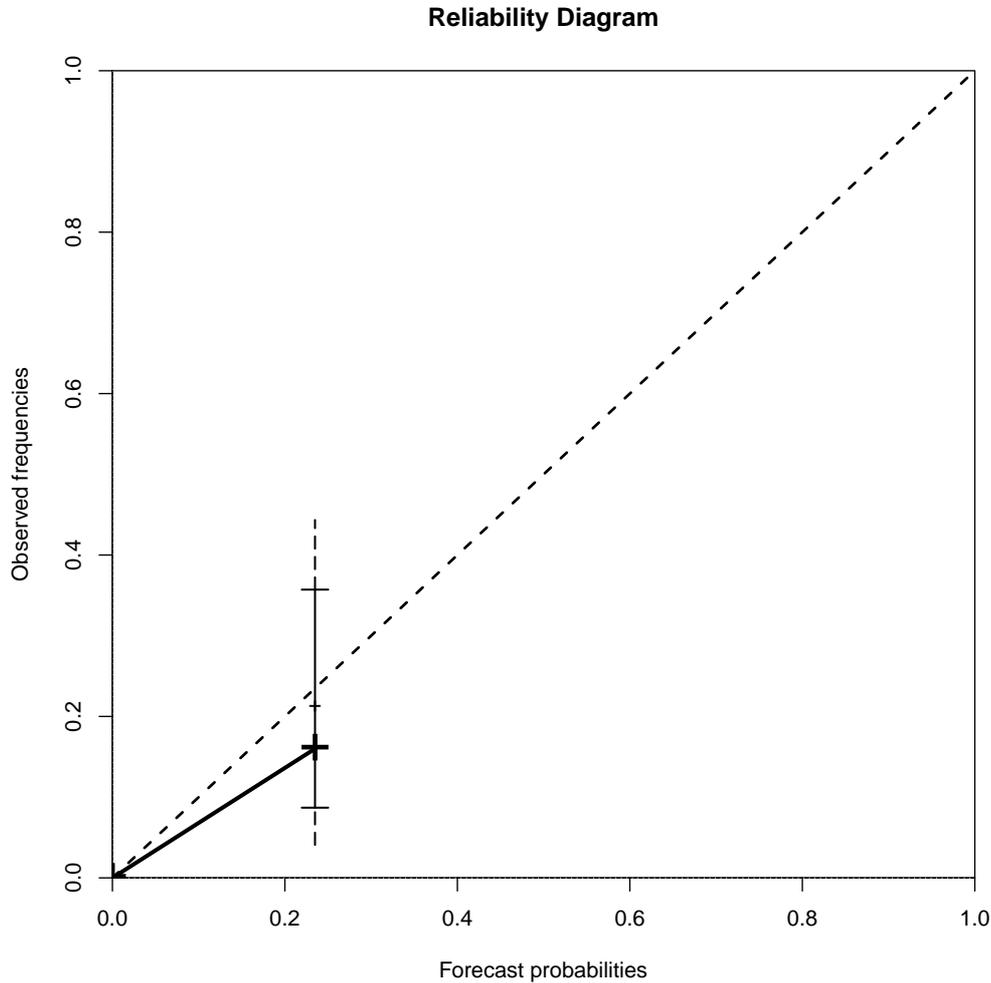
Figure 3.6: Forecast binning: Reliability diagram showing the variation in sampling error where the forecast bins are not equi-probable. The bin with boundaries $[5.8 \times 10^{-13}, 1.0]$ has a bin population of 16 whereas the bin with boundaries $[0, 5.8 \times 10^{-13}]$ has a bin population of 496. The calibration function estimate $\hat{\kappa}(r_2)$ has a considerably large variance potentially rendering recalibration ineffective for the higher probability values. The forecasts are generated under Expt. 4 (see table 3.1).

Regularisation normally involves the use of algorithms to determine the value of such a parameter under asymptotic conditions [73], but, in the case of binning a limited number of forecasts, it is more difficult to implement. Alternatively, δ could be selected by optimising a proper score such as ignorance as a function of δ using a training set of forecast-outcome pairs. Like the estimation of the

calibration function, however, this approach to determining the regularisation parameter is likely to be subject to mis-estimation since the training set differs from the evaluation set.

3.4.2 Binary forecast recalibration results

The effectiveness of forecast recalibration is assessed in this section through examination of the numerical results of recalibration under IMS. Improvement in forecast performance after recalibration is measured with respect to both the ignorance score, and forecast reliability (see Eqns. (2.37) and (2.41)). The uncertainty component of ignorance IGN_{UNC} remains the same before and after recalibration since it is dependent only on the climatological probability, that is

$$IGN_{UNC} = -p_{clim}^{\theta} log p_{clim}^{\theta} - (1 - p_{clim}^{\theta}) log(1 - p_{clim}^{\theta}). \qquad (3.7)$$

Changes in forecast skill can be regarded purely as the difference between the change in forecast reliability and forecast resolution, that is

$$\Delta IGN = \Delta IGN_{REL} - \Delta IGN_{RES}. \qquad (3.8)$$

Primo et al. [152] find that binary forecast skill is increased through the improvement of forecast reliability. The same conclusion is reached here, and in fact, it has been found that increase in forecast reliability is generally required to achieve increased forecast skill. The results vary for different calibration methods so consideration is given to the relative benefits of each of the methods. A sample of IGN and IGN_{REL} of forecasts before and after recalibration with the kernel density estimation (KDE) algorithm is given in table 3.4. The KDE algorithm has proved to be one of the more effective recalibration approaches for improving forecast skill.

Adjusted Counted (AC) Method

The skill of forecasts produced with the AC density construction method was found to be competitive with the best performing KDB method in section 3.2

Table 3.4: Forecast skill before and after recalibration

Recalibration (& PDF) method	θ	N_{ens}	τ^*	Before recalibration		After recalibration		Difference	
				IGN	IGN_{REL}	IGN	IGN_{REL}	ΔIGN	ΔIGN_{REL}
Kernel density estimation (AC)	0.5	4	0.2	-0.832	0.053	-0.941	0.031	-0.109	-0.022
			6.4	-0.634	0.084	-0.649	0.012	-0.024	-0.072
			25.6	0.195	0.178	0.008	0.006	-0.187	-0.172
	0.5	1024	0.2	-0.982	0	-0.997	0	-0.015	0
			6.4	-0.775	0.047	-0.827	0.004	-0.052	-0.043
			25.6	-0.003	0.012	-0.005	0.003	-0.002	-0.009
	0.99	4	0.2	-0.057	0	-0.060	0	-0.003	0
			6.4	-0.008	0.003	-0.012	0	-0.004	-0.003
			25.6	0.012	0.014	0.014	0.016	0.002	-0.002
	0.99	1024	0.2	-0.062	0	-0.066	0	-0.004	0
			6.4	-0.033	0	-0.024	0.006	0.009	0.006
			25.6	-0.006	0.001	0	0	0.006	-0.001
Kernel density estimation (KDB)	0.5	4	0.2	-0.979	0.053	-0.977	0.031	0.002	-0.022
			6.4	-0.683	0.084	-0.688	0.012	-0.005	-0.072
			25.6	0.031	0.178	0.005	0.006	-0.026	-0.172
	0.5	1024	0.2	-0.983	0	-0.997	0	-0.014	0
			6.4	-0.774	0.047	-0.881	0.004	-0.107	-0.043
			25.6	-0.003	0.012	-0.005	0.003	-0.002	-0.009
	0.99	4	0.2	-0.061	0	-0.065	0	-0.004	0
			6.4	-0.015	0.003	-0.021	0	-0.016	-0.003
			25.6	0	0.014	0	0.016	0	0.002
	0.99	1024	0.2	-0.062	0	-0.066	0	-0.004	0
			6.4	-0.030	0	-0.031	0.006	-0.001	0.006
			25.6	-0.006	0.001	0	0	0.006	-0.001

*in Lorenz63 seconds [118]. All values are rounded to 3 decimal places.

for most forecast-parameter values. For forecast-parameter values where the raw AC forecasts lack skill and reliability, specifically at longer lead times, smaller ensemble sizes, and higher climatological event frequencies, forecast recalibration is shown below to be more effective and robust. Improvement in forecast performance after recalibration is relatively minimal, however, with increases in IGN ranging from ~ 0 to ~ 0.19 bits, and increases in IGN_{REL} ranging from ~ 0 to ~ 0.17 bits. Increases in recalibrated AC forecast skill tend to arise mostly from increases in reliability, because they are proportionately higher than changes in forecast resolution at larger ensemble sizes. In fact, resolution is decreased after recalibration for larger ensemble sizes which implies that the reliability increases relatively more.

The KDE algorithm 2.4.1 is one of the more effective recalibration algorithms and achieves an increase of forecast skill of $\Delta IGN = 0.008 - 0.195 = -0.187$,

and an increase of forecast reliability of $\Delta IGN_{REL} = 0.006 - 0.178 = -0.172$ for forecast-parameter values $\tau = 25.6s$, $N_{ens} = 4$, and $\theta = 0.5$. Figs. 3.7 and 3.8 show reliability diagrams which demonstrate the effect of recalibration on reliability for those forecast-parameter values. The improvement of skill corresponds to a percentage increase of 16% in terms of bits of information. Such an increase in skill may justify forecast recalibration rather than improving forecast technique (e.g. improving a data assimilation (DA) scheme) to improve forecast performance. An investigation into the relative benefits of forecast recalibration and resource cost of forecast technique improvement would prove useful but is beyond the scope of this thesis.

The beta transform algorithm produces similarly effective recalibration results to the KDE algorithm, and achieves the largest improvements in forecast skill and reliability for forecast-parameter values $\tau = 25.6s$, $N_{ens} = 4$, and $\theta = 0.5$. The increases in skill and reliability are $\Delta IGN = 0.008 - 0.195 = -0.18$ and $\Delta IGN_{REL} = 0.006 - 0.178 = -0.172$, respectively. The improvement in forecast performance is quite marginal, however, for most forecast-parameter values, and the algorithm does not perform as well as the KDE algorithm at lower climatological event frequencies. Recalibration is evidently more effective with the algorithms which do not impose binning of forecasts before estimating the calibration function κ. The exception is the ROC curve fitting algorithm which generally leads to a decrease of forecast skill and reliability. Bröcker [23] points out that estimating κ with a *bi-normal model* reduces its variance, and the degrees of freedom, and as a result the ROC curve fitting method may be prone to bias. Furthermore, the degrees of freedom cannot be controlled so that data cannot be used effectively for estimating κ. These limitations may restrict the degree of improvement of forecast skill and reliability.

The linear and logistic regression, and simple translation recalibration algorithms are generally effective in improving forecast performance where raw AC forecast skill is poorer (i.e. longer lead times, smaller ensemble sizes, and higher climatological event frequencies). The linear regression algorithm ac-
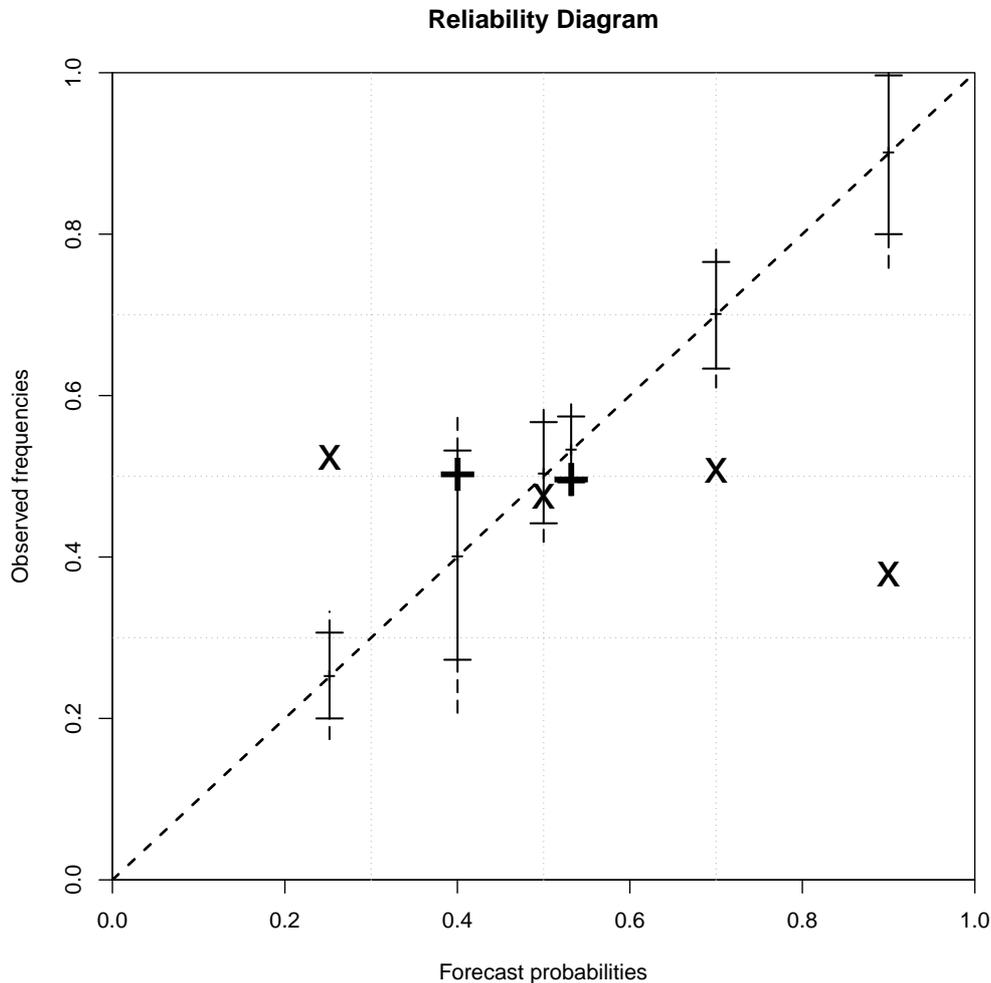
Figure 3.7: Forecast reliability after recalibration: reliability diagram showing the forecast reliability of the raw set (crosses) and recalibrated evaluation set (pluses) using the KDE algorithm. The position of the raw and recalibrated forecast bins suggests that the recalibrated forecasts are more reliable than the raw forecasts and the changes to forecast skill and reliability (i.e. $\Delta IGN = -0.187$ and $\Delta IGN_{REL} = -0.172$, respectively) confirm the improvement. All sets are generated under Expt. 3 (see table 3.1).

tually achieving the best improvement in forecast performance out of all six algorithms at the smallest ensemble sizes and shortest lead times. The linear and logistic regression, and simple translation algorithms are less effective as ensemble size increases, however, and tend to result in deterioration of forecast performance particularly at the lowest climatological event frequency ($\theta = 0.5$).
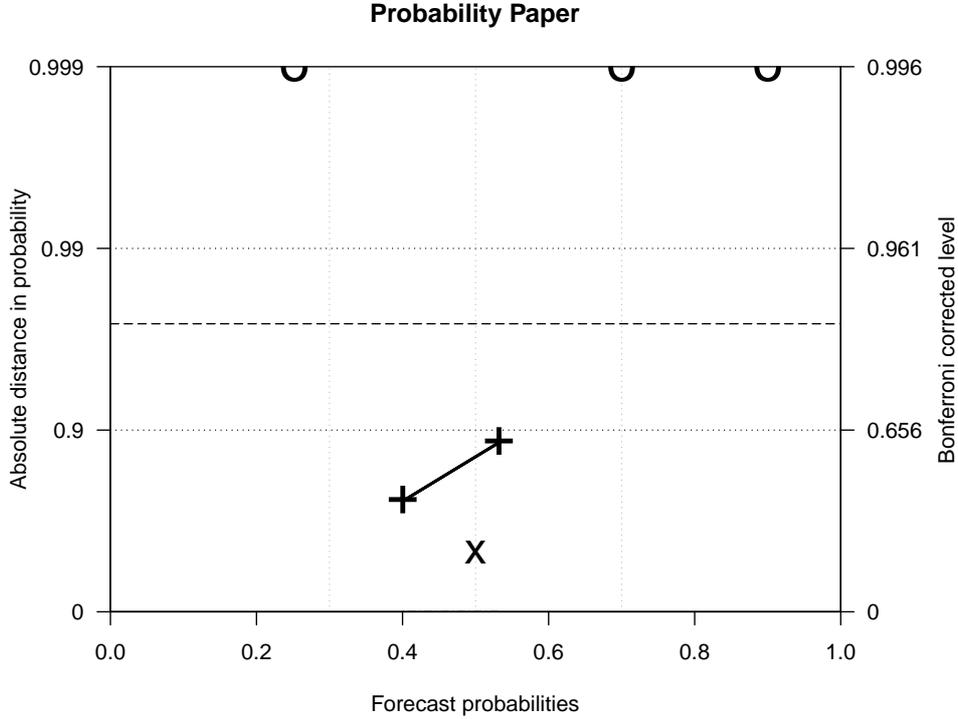
Probability Paper

Figure 3.8: Forecast reliability after recalibration: reliability diagram on probability paper showing the forecast reliability of the raw set (crosses) and recalibrated evaluation set (pluses) using the KDE algorithm with 5% - 95% (1% - 99% vertical dashed line) consistency bars.The improvement in reliability is more evident since both the recalibrated forecast bins lie within the Bonferroni corrected 0.99 probability distance (upper dotted) band whereas only one raw forecast bins does so. All sets are generated under Expt. 3 (see table 3.1). All other details are identical to Fig. 3.5.

The decrease in forecast skill after recalibration with the simple translation algorithm is $\Delta IGN = -0.846 - -0.982 = 0.136$ for forecast-parameter values $\tau = 0.2s$, $N_{ens} = 1024$, and $\theta = 0.5$. In this case, the decrease in forecast reliability is relatively minimal (i.e. $\Delta IGN_{REL} = 0.007 - 0 = 0.007$), however, indicating that the decrease in forecast skill is chiefly caused by a loss of forecast resolution in accordance with Eqn. (3.8) above. This proportionately larger loss of resolution is actually a common effect of recalibration with the simple translation algorithm where ensemble sizes are larger. A discussion of

the effects of recalibration on forecast resolution is given in section 3.4.3.

Blending Method
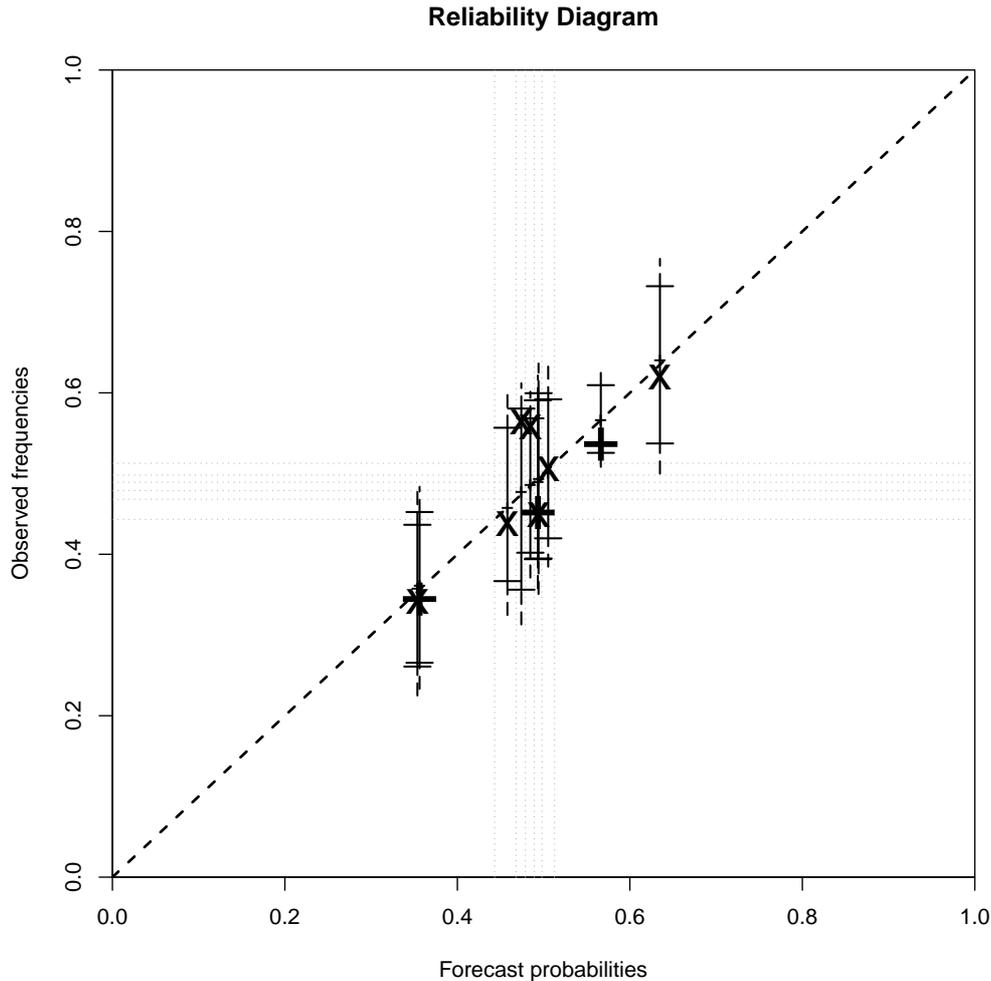
Figure 3.9: Forecast reliability after recalibration: Reliability diagram showing KDB forecast reliability of the raw set (crosses) and recalibrated evaluation set (pluses) using the simple translation method. Recalibration is ineffective here since the forecasts are already well-calibrated. All sets are generated under Expt. 5 (see table 3.1).

Raw probabilistic forecasts which have been constructed with the KDB method have more skill than those constructed with the AC method, yielding smaller improvements in skill and reliability after recalibration. Most recali-
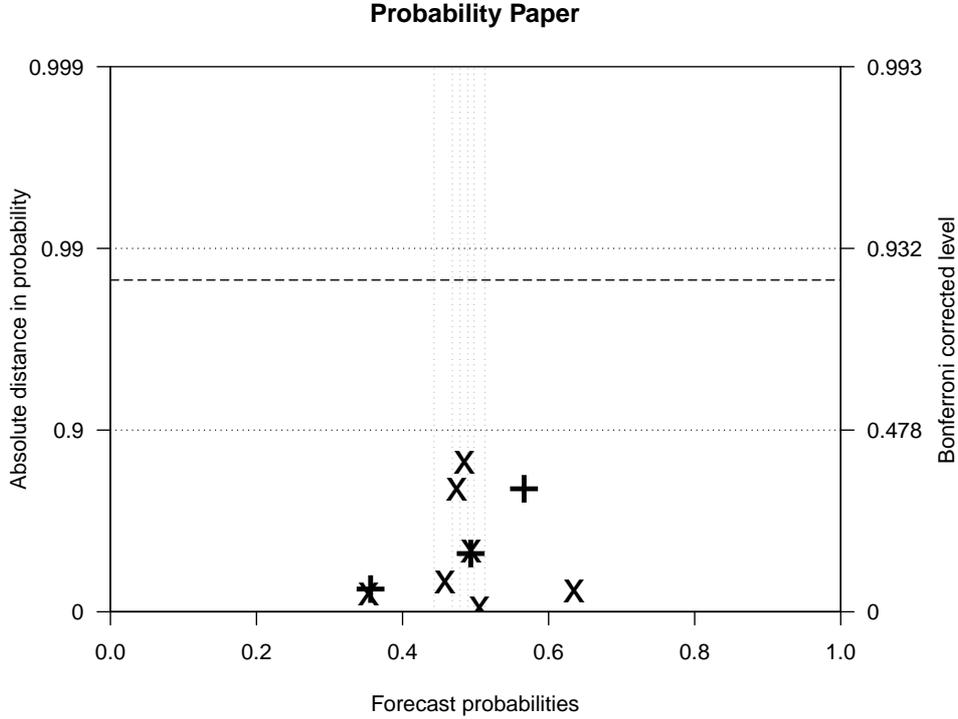
Probability Paper

Figure 3.10: Forecast reliability after recalibration: Reliability diagram on probability paper showing KDB forecast reliability of the raw set (crosses) and recalibrated evaluation set (pluses) using the simple translation method with 5% - 95% (1% - 99% vertical dashed line) consistency bars. The right-hand axis indicates the equivalent Bonferroni corrected levels e.g. for a reliable forecast, all of the points (7 bins) would be expected to fall within the 0.99 probability distance band with a 93.2% chance. In addition, the dashed lines indicate where the entire diagram would be expected to fall within with a 90% chance. Recalibration is ineffective here since the forecasts are already well-calibrated. All sets are generated under Expt. 5 (see table 3.1).

bration algorithms are ineffective, and often lead to a degradation of forecast performance, particularly at lower climatological event frequencies. A non-degradation of skill after recalibration would seem to be a minimum requirement of any recalibration algorithm. Only the KDE algorithm demonstrates relatively consistent efficacy, however, with the largest increases in skill ($\Delta IGN = -0.25$ bits) and reliability ($\Delta IGN_{REL} = -0.16$ bits) occurring for larger ensemble sizes and the highest climatological event frequency $\theta = 0.5$, but increases in

skill and reliability are mostly less than 0.1 bits of information. These more marginal increases in skill suggest that recalibration should not be opted for ahead of advancing forecast technique to improve binary forecast performance where it is already strong. Note that any marginal increases and decreases in the skill and reliability of the KDB forecasts are difficult to confidently attribute to the effectiveness of the recalibration process, and may be explained to some extent by variation in sampling uncertainty. Increasing the sample size of evaluated forecasts can overcome this, of course, if potential changes in forecast performance are deemed important.

Unlike the AC forecasts, any increases in recalibrated KDB forecast skill originate from more equal increases in reliability and resolution, and losses of resolution are not so severe at larger ensemble sizes. Furthermore, the largest increases in skill and reliability do not occur at the longest lead time $\tau = 25.6s$ with the KDB forecasts because poor raw, or pre-recalibration, skill is avoided by blending the forecast with the climatological forecast. The relationship between recalibration efficacy and forecast-parameters is less definitive for the KDB forecasts with improvements occurring more randomly for given forecast-parameter configurations. The lack of trend in the relationship and overall improvement in forecast performance can be attributed to the higher degree of raw forecast skill, and the upper bounds of skill imposed by optimal ignorance.

The overall relative performance of the recalibration algorithms with the AC forecasts is replicated with the KDB forecasts. The KDE algorithm is again the most effective. The beta transform algorithm is not as effective, however, and results in degradation rather than improvement of forecast performance for a larger range of forecast-parameters. In fact, recalibration is sometimes more effective with the linear regression algorithm than the beta transform algorithm. The ROC curve fitting method is again ineffective, and leads to substantial reductions in forecast skill, reliability and resolution.

The linear and logistic regression, and simple translation recalibration algorithms are also generally ineffective at improving forecast performance where

raw KDB forecast skill is higher (i.e. shorter lead times, larger ensemble sizes, and lower climatological event frequencies). Recalibration often leads to degradation for various forecast-parameters with the largest decrease in skill ($\Delta IGN = 0.41$ bits) occurring after execution of the logistic regression algorithm at the lowest climatological event frequency $\theta = 0.99$. Figs. 3.9 and 3.10 illustrate how recalibrating perfectly reliable raw forecasts can result in a decrease of the reliability of the recalibrated forecasts. The bins corresponding to the raw forecasts (crosses) all lie within the 5% - 95% consistency bars, and are therefore reliable. Performing recalibration with the simple translation algorithm has minimal effect on increasing the reliability of the forecasts, and in fact reduces the forecast resolution slightly ($\Delta IGN_{RES} = -0.047$ bits). The overall effect is to reduce the skill of the forecasts ($\Delta IGN = 0.043$ bits). The overall result of recalibration being more effective where forecast performance is worse before recalibration is expected to generalise beyond these particular examples of binary forecasting.

3.4.3 Forecast Resolution after recalibration

Forecast resolution (see Section 1.6.4) is also a key attribute of forecast performance [196, 8]. It pertains to the differences between the conditional expectation of a binary event given a forecast probability (i.e. $E(y|p)$) and the marginal (unconditional) expectation of the event (i.e. $E(y) = p_{clim}^{\theta}$) [142]. In short, it measures the ability of a forecast system to discriminate between scenarios where the event occurs more or less frequently than the climatological frequency p_{clim}^{θ}. Hence, the poorest possible forecast resolution occurs where p_{clim}^{θ} is always forecast.

As explained in Section 3.4.2, the ignorance score can be decomposed into components of reliability, resolution and uncertainty (see Eqn. (2.40)). The difference between the second two components (i.e. $IGN_{UNC} - IGN_{RES}$) is referred to as the *sharpness* of the forecast, and is a measure of the concentration

Reliability Diagram
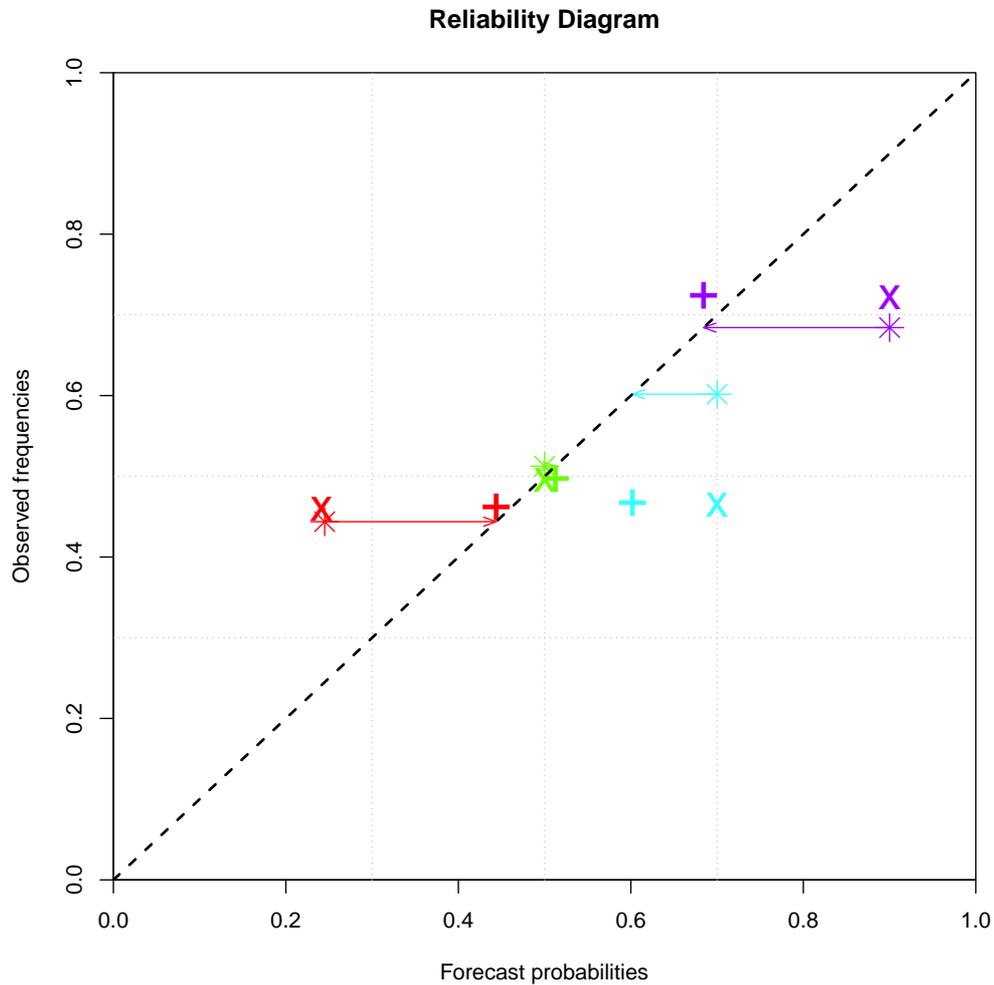
Figure 3.11: Simple translation recalibration: reliability diagram schematic of the *simple translation* recalibration method using a training set of Lorenz63 binary forecasts (asterisks) to recalibrate the evaluation set of forecasts (pluses → crosses) both generated in Expt. 6 (see table 3.1). Recalibration has resulted in most bins being translated closer to the diagonal so that forecast resolution is decreased. Raw resolution was already low in this case so the decrease is relatively small $\Delta IGN_{REL} = -0.003$, but this example has merely been selected to demonstrate the effect. Each bin is coloured differently for clarity.

of the forecast PDF [59]. In the case of binary forecasts, the sharpest forecast corresponds to $p \simeq 0$ or $p \simeq 1$. To maximise forecast skill, the aim is to issue a forecast which is as sharp as possible subject to it being perfectly reliable (i.e. zero bias and variance of the calibration function κ) [142, 59]. The resolution

component of ignorance can be expressed as

$$IGN_{RES} = \sum_j P(p_j) \left[f_j log \frac{f_j}{p_{clim}^\theta} + (1 - f_j) log_2 \frac{1 - f_j}{1 - p_{clim}^\theta} \right], \qquad (3.9)$$

where f_j is the conditional frequency of an event occurrence on all occasions where p_j is forecast. As can be inferred from Eqn. (2.40), IGN_{RES} is positively oriented, and, as referred to above, a climatological forecast yields perfect reliability but the poorest resolution (i.e. $IGN_{RES} = IGN_{REL} = 0$).

While recalibration often improves the reliability of binary forecasts, it often results in a decrease of resolution of the KDB forecasts, particularly at the highest climatological event frequency ($\theta = 0.5$) and shortest lead times. This effect has been noted previously in the literature [165, 196, 161]. Jolliffe and Stephenson [86, 191] point out that, in principle, recalibration can only be used to improve reliability, but not resolution. They conclude that resolution is a necessary condition for forecast skill, whereas reliability is not. The results of recalibration under IMS indicate that KDB forecasts recalibrated with the simple translation and logistic regression algorithms are particularly susceptible to decreases in resolution. These decreases are relatively smaller than increases in reliability where recalibration leads to improvement in forecast skill (see Eqn. (3.8)). Recalibration of the KDB forecasts with the shortest lead times where the climatological event frequency is higher (i.e. $\theta = 0.5$) with these two algorithms can result in some of the largest decreases in skill. Typically, the raw resolution is higher for these forecast-parameters, indicating that there is a larger potential for loss of resolution after recalibration. Figure 3.11 shows the effect of recalibration on resolution using the simple translation algorithm. The KDE and beta transform algorithms are more resistant to decreases in resolution, with any decreases in IGN_{RES} being of the order of 10^{-2} bits.

3.5   Forward view and conclusions

In this chapter, the forecast evaluation and recalibration framework proposed in Chapter 2 has been deployed under an imperfect model scenario to survey the conditions under which recalibration proves an effective tool for improving forecast performance (i.e. forecast skill and reliability). Model inadequacy inhibits the ability of imperfect models to accurately simulate the trajectories of a nonlinear dynamical system's state in the real world. It was concluded in Chapter 2 that forecast systems that demonstrate less skill may benefit more from forecast recalibration given the larger margin for improvement in forecast performance. This conclusion has been assessed using an imperfect Lorenz63 model in this chapter.

An imperfect model has been used to produce forecasts of the state of the Lorenz63 dynamical system using the adjusted counted (AC) and kernel dressing and blended (KDB) density construction methods. Forecast recalibration has indeed been found to be more effective in terms of ignorance where pre-recalibration forecast performance is most poor. The relative improvement of the performance of the AC forecasts, which had predominantly lower skill and reliability before recalibration, was substantially greater than the KDB forecasts after recalibration. With respect to the forecast-parameters, poorer pre-recalibration forecast performance usually occurs where model ensemble size is smaller, at longer forecast lead times, and where the probability of the climatological event occurring is more uncertain (i.e. values of θ closer to 0.5). The results of the recalibration experiments under both PMS and IMS suggest that recalibrating forecasts with poorer raw skill may be preferable to improving various technical features of a forecast system since it is straightforward and quick to implement. Quantifying this suggestion, while intriguing, is beyond the scope of this thesis.

A previously unreported property of the ignorance score has also been presented in Section 3.2. It has been discovered that the degree of skill of binary

forecasts is limited by a lower bound on the skill of climatological reference forecasts. This imposes bounds on the score of a perfect forecast, defining the optimal ignorance IGN_{opt}, and, hence, a limit on the potential improvement of forecast performance after recalibration. Indeed, the same limitation on the best possible skill applies to other scoring rules (e.g. the Brier score) where there is a lower bound on the skill of reference forecasts. Nevertheless, the margin of difference between the score of a perfect forecast IGN_{opt} and the actual score for a given climatological event frequency θ indicates just how skilful a forecast system is. IGN_{opt} has values: $IGN_{opt} = -1.0$ for $\theta = 0.5$, $IGN_{opt} = -0.47$ for $\theta = 0.9$, and $IGN_{opt} = -0.08$ for $\theta = 0.99$.

Finally, the effect of recalibration on forecast resolution has been investigated following observations made in the literature that resolution is decreased after recalibration [165, 196, 161], and in this thesis, that forecasts sometimes have less skill after recalibration despite little or no decrease in forecast reliability. A more thorough investigation of this undesirable result of recalibration has been conducted here, revealing that recalibration can indeed result in decreases in forecast resolution when performed with the simple translation and logistic regression algorithms. The largest decreases in resolution occur at the highest climatological event frequency $\theta = 0.5$, and shorter lead times where the pre-recalibration forecast resolution is already high.

The novel contributions or innovations in this chapter include:

- quantification of optimal skill of binary forecasts

- new insights regarding the limitations of forecast binning/categorisation for forecast recalibration, and review and critique of binning/categorisation methods in the literature

- novel investigation of the efficacy of forecast recalibration under IMS using all recalibration methods reviewed in Chapter 2, including determination of the forecast-parameters, e.g. ensemble sizes and lead times, where forecast recalibration performs the best

- investigation of the changes to forecast skill and forecast reliability after recalibration using the decomposition of the ignorance score

- exploration and analysis of the conditions where recalibration has a detrimental effect on forecast resolution

Chapter 4

The effect of serial dependence on estimates of forecast skill

Establishing statistical confidence in forecast skill can be complicated by serial dependence in the time series of evaluation outcomes. Wilks [216] demonstrates how serial correlation of forecasts and outcomes can be transmitted to forecast evaluation statistics, so that their sampling variances are inflated relative to uncorrelated forecast evaluation statistics. This is an important result because it leads to estimates of forecast skill that are overconfident. Moreover, the magnitude of the effect increases with forecast skill as the forecasts correspond more closely to the serially correlated outcomes, and become themselves more serially correlated. The effect of serial dependence on forecast evaluation has also been noted in several other studies such as Hamill [119], Ferro [55], and Pinson [159]. Such effects on the sampling properties of evaluation statistics have important implications for proving forecast skill because increased sample sizes are required to obtain reliable skill estimates. As demonstrated below, serial dependence in a forecast-outcome time series need not always be transmitted to the evaluation statistics, however; forecasting scenarios where estimates of forecast skill are not misleading have been identified. Examples of each of three possible scenarios are described for the first time in this chapter. Firstly, in cases where

linear serial correlation in an outcome time series is transmitted to the forecast evaluation statistics, secondly in cases where linear serial correlation in an outcome time series is not transmitted to the forecast evaluation statistics, and thirdly in cases where there is nonlinear serial correlation[1] in an outcome time series resulting in linear serial correlation in the forecast evaluation statistics.

The effect of serial dependence on the sampling distributions of statistics, although often overlooked in forecast evaluation studies [216], is commonly encountered in the statistical analysis of geophysical variables, and well covered in the literature [111, 197, 192]. Consider a random variable which has a population distribution with mean μ and standard deviation σ. An intuitive result of the Central Limit Theorem is that the finite-time average of a sample of N independent and identically distributed (i.i.d.) observations of the random variable is a normal random variable with mean μ and standard error $\sigma/\sqrt{N}$. The scaling of the standard error as $1/\sqrt{N}$ need not hold, however, if the observations of the random variable are not i.i.d.. As sample size increases, the rate of convergence of the sample averages onto the true mean μ can be significantly slower (or faster) than those which are serially independent. Importantly, this means that the duration of time required to obtain realistic estimates of μ is prolonged (or shortened) under serial dependence.

Typically, geophysical phenomena are *red* processes, meaning that positive linear serial correlation is present in observational data. In short, a time series of observations does not satisfy the assumption of independence. Given that geophysical phenomena can exhibit cycles of variability on timescales of, for example, up to at least 10^6 years [111], there are cases where samples of data are collected at time intervals which are too short for the assumption of independence to hold. The sampling variance of a time average computed from serially correlated geophysical data need not scale as $1/\sqrt{N}$, as do i.i.d. data (i.e. a *white-noise* process). Making the assumption of independence leads to a

---

[1] while serial correlation can be either linear or nonlinear, serial dependence is the term generally used here to refer to either definition
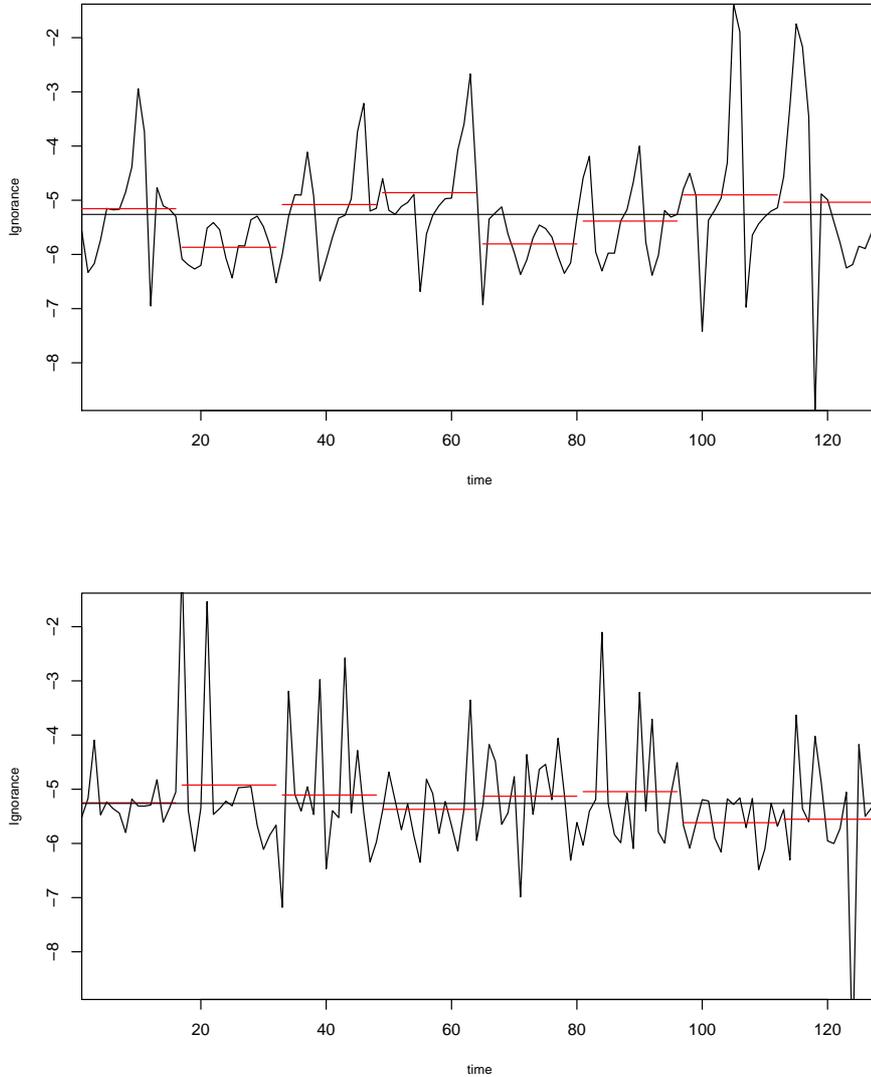
Figure 4.1: Serial correlation in forecast skill statistics: time series of 2^7 IGN scores of forecasts of Lorenz63 system states (top) and bootstrap resamples of the same time series (bottom). The time series is serially correlated while the bootstrap resamples are serially independent. Averages over sequential samples of size $N = 16$ (red lines) tend to deviate from the IGN estimate over the entire time series ($IGN = -5.05$; horizontal line) in the top plot compared to the bottom plot, resulting in a sampling distribution of the averages which is larger. The sampling variances of the 8 subsamples are $s^2_{IGN} = 0.15$ and $s^2_{IGN_{boot}} = 0.06$.

discrepancy in the sampling variance, which if not accounted for, will result in excessively precise estimates of the "true" value of a given statistic. Critically, in weather and climate statistical analysis, larger data samples (and typically longer durations of time) are consequently necessary for reliable estimates of that statistic. Textbook statistical inference tests based upon the assumption of independence of the observations [217] may lead scientists to make inaccurate statistical estimates of a given underlying parameter. Generally, linear serial correlation is not detectable in observations of nonlinear systems with the auto-correlation function [58], in which case it may not even be possible to determine whether the assumption of independence is valid.

Wilks [216] found that accurate forecasts of serially dependent observed outcomes can consequently be themselves serially dependent, resulting in inflation of the sampling variance of the Brier score, and hence, inaccurate estimation of its true value (see Section 1.6.2). Statistical inferences of forecast skill made where sampling variances are inflated are *over*-confident, yielding overly precise confidence intervals and p-values for significance tests that are too small (i.e. an overly frequent occurrence of type I errors). Essentially, a larger sample size is required if data are serially correlated to obtain the same correct inferences of skill made with independent data. "Effective sample size" (ESS) (see Thiébaux and Zwiers [192]) corrections should be made, as in Wilks [216], to ensure that confidence intervals possess accurate probability coverage, otherwise estimates of forecast skill are likely to be misleading. An important result from Wilks [216] is the formulation of ESS corrections from the ratio of the analytical-to-empirical Brier score sampling variances. That particular analytical solution of the Brier score sampling variance is derived under the assumption that forecast-outcome pairs are i.i.d. so it can be used as an indicator of inflation of the empirical sampling variances. Figure 4.1 illustrates how the effect of serial dependence arises by showing a sample time series of serially correlated forecast skill scores and an i.i.d. bootstrap resample of serially independent forecast skill scores. The larger sampling variance of 8 subsample mean scores computed from the serially

correlated scores ($s^2_{IGN} = 0.15$) is larger than the serially independent scores ($s^2_{IGN_{boot}} = 0.06$)) demonstrating the inflationary effect of serial dependence on forecast evaluation statistics.

The investigation in chapter extends the study of Wilks [216] of the effects of serial dependence in sequential forecasts on estimates of the Brier score to a number of different forecast scenarios. In this case, the sampling properties of the ignorance score (see Section 1.6.2) under serial dependence is considered. It is shown for the first time that, while inflationary [2] effects can be exhibited in the sampling variances of the two scoring rules where there is linear serial correlation in sequential forecasts, serial dependence is neither a sufficient nor necessary condition for estimates of forecast skill to be inaccurate. Previously undiscussed cases have also been identified where serial dependence in evaluation data does not necessarily result in misleading estimates of forecast skill, and where nonlinear serial dependence in evaluation data, not detectable with an autocorrelation function, does result in misleading estimates of skill.

This chapter is structured as follows: Section 4.1 provides derivations of the analytical solutions of the sampling variances of the Brier score and ignorance estimates in a binary outcome scenario. The latter derivation is an original contribution in this thesis, and, in theory, can be used as a measure of the inflation of the empirical sampling variances under serial dependence, and to determine ESS corrections. The derivation of an analytical solution of the sampling variance is generally not straightforward for scoring rules, however, and the solutions need to be evaluated with sufficient sample sizes for them to be stable [17, 216].

Novel case studies illustrating the three possible scenarios described above where serial dependence either does or does not affect forecast skill estimates are presented in Sections 4.2, 4.3, and 4.4. A range of data-generating stochastic and dynamical systems, and forecast models are employed to demonstrate each scenario. The first scenario, in which linear serial correlation in evaluation data

---

[2]inflation of sampling variances relative to those under the assumption of independence

is transmitted to forecast evaluation statistics, is demonstrated in Section 4.2 by replicating Wilks's [216] result and evaluating forecasts of the trajectories of the Lorenz63 system [118]. Both examples illustrate that the effects of serial dependence on forecast skill estimation can be significant. An AR(1) process and stochastic testbed hurricane system are used in Section 4.3 to demonstrate where serial correlation in evaluation data is *not* transmitted to evaluation statistics, leaving statistical inference unaffected. This second case study shows Wilks's result does not always apply where forecasts and outcomes are serially correlated. The third scenario where nonlinear serial correlation in evaluation data results in sample variance deflation and hence, misleading estimates of forecast skill is illustrated in Section 4.4.

Estimation of the minimum sample sizes necessary for forecast skill estimates to converge onto their asymptotic "true" value under serial dependence and serial independence is examined in Section 4.5 in the context of a new concept called *time until convergence*. The relationship between the effect of serial dependence on forecast skill estimates and the predictability in state space of a dynamical system is also investigated for the first time.

Finally, an approximate method for ESS corrections by comparing estimates of the sampling variance under serial dependence and serial independence using a resampling method (Bradley et al. [17]) is proposed in Section 4.6. Computation of ESS corrections allow a forecast user to estimate the duration of time required to achieve statistically significant forecast skill estimates. For example, in "Weather Roulette" [67], possessing knowledge of how long it would take to prove the skill of a given forecast system may affect a punter's decision whether to immediately place bets using information from that forecast system, or wait until establishing statistical confidence in it (see also the "swindled statistician scam" in Chapter 5).

4.1 Sampling distributions of scoring rules

Forecast evaluation is routinely carried out to monitor and improve the quality of forecast systems, yet often the sampling uncertainty of a scoring rule is insufficiently accounted for [85]. Statistical inference of model output or forecast quality should only be made when sampling uncertainty is quantified, yet this is often omitted. The sampling variance of a scoring rule is dependent on both sample size and the statistical characteristics of the forecasts and outcomes [17]. In that sense, a scoring rule can be considered in the same way as standard statistical inference, where some underlying parameter or value, θ, is estimated, for example, by constructing a confidence interval for an empirical estimate $\hat{\theta}$ using a resampling method [85]. This is a simple and robust approach to determining sampling variance, but it can also be computationally inefficient. Illustration of the effects of serial dependence on forecast evaluation does not require inordinate sample sizes, however, so the empirical approach is opted for in this section.

An alternative approach which requires minimal computational effort is to derive the sampling variance of a particular scoring rule analytically using sampling theory [17]. Such a derivation is based on the assumption that the forecast-outcome pairs (p_i, Y_i) are independent random samples from their joint distribution. This assumption is commonly (and mistakenly) made in real world weather and climate forecasting [177], potentially resulting in misleading estimates of forecast skill. Derivations of the analytical sampling variance of the ignorance score is now presented, following Bradley et al. [17] and Wilks [216].

The sample estimator of the ignorance score (IGN) is expressed as

$$IGN = -\frac{1}{N} \sum_{i=1}^{N} log_2(p(Y_i)),\tag{4.1}$$

where $p(Y_i)$ is the probability assigned to outcome Y_i. The sampling variance

is then

$$Var[IGN] = Var\left[-\frac{1}{N}\sum_{i=1}^{N} log_2 p(Y_i)\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} Var[log_2 p(Y_i)]$$

$$= \frac{1}{N}Var[log_2 p(y)]. \tag{4.2}$$

The variance term on the RHS can be expanded as follows

$$Var[log_2 p(y)] = E[log_2^2 p(y)] - E[log_2 p(y)]^2$$

$$= E[log_2^2 p(y)] - IGN^2. \tag{4.3}$$

Therefore,

$$Var[IGN] = \frac{1}{N}[E[log_2^2 p(y)] - IGN^2], \tag{4.4}$$

where $E[log_2^2 p(y)]$ is numerically estimated from the outcome dataset of size N as

$$E[log_2^2 p(y)] = \frac{1}{N}\sum_{i=1}^{N} log_2^2 p(Y_i). \tag{4.5}$$

The derivation of Eqn. (4.4), based on the derivation of the Brier score sampling variance given by Bradley et al. [17], is an original contribution in this thesis. Wilks [216] utilises the fact that the sampling variance of the Brier score depends only on the moments of the joint distribution of the forecasts p and outcomes y to express it in terms of the parameters of a model. Expressions for ESS corrections can then be derived also in terms of the model parameters. Derivation of such ESS corrections is more difficult for IGN because Eqn. (4.4) depends on $E[log_2^2 p(y)]$ rather than the moments of the joint distribution, and is beyond the scope of this thesis. Instead, an alternative approximate method for ESS corrections is proposed here. This *approximate* method consists of finding the difference between sample sizes corresponding to a given empirical sampling variance computed respectively from serially dependent synthetic time series and serially independent bootstrap resamples. Section 4.6 provides a fuller explanation and demonstration of the approximate ESS correction method.

As previously stated, the analytical sampling variance solution in Eqn. (4.4) should be evaluated with sufficient sample sizes for them to yield stable results so that they are usefully accurate. Bradley et al. [17] found that, because of the inclusion of the higher moments of the joint distribution of the forecasts and outcomes in the Brier score, sample sizes are required to be fairly large. Wilks [216] determined that a sample size of $N = 3000$ yields stable enough results. The analytical sampling variance solution of IGN requires a larger sample size owing to its logarithmic function.

4.2 Case Study 1: Transmission of linear serial correlation to forecast evaluation statistics

Wilks's [216] key result is that positive serial correlation in evaluation data results in inflation of the variances of the sampling distributions of the Brier score where forecasts are sufficiently skilful. Inflation of the variance of the sampling distribution of the ignorance score is demonstrated here by both replicating Wilks's approach [216] using a probability model for forecast refinement distributions, and employing the Lorenz63 [118] system, which is used here for the first time in this context.

4.2.1 Linear-calibration/beta-refinement model

The stochastic "linear-calibration/beta-refinement" (LCBR) probability model [142, 213] used by Wilks [216] to study the effects of serial dependence on forecast evaluation statistics is also used here to frame the problem. The LCBR model provides a useful representation of the statistical properties of probability-of-precipitation binary forecasts in the USA over the period 1972-1987, and has been used to realistically simulate precipitation statistics. This simple stochastic model is not intended to accurately represent forecast statistics [216], but serves as an effective tool for examining the effects of serial

dependence on forecast evaluation sampling distributions.

Let p be a binary forecast produced from the probability model for beta-refinement distributions be defined by a probability density function given by

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \qquad (4.6)$$

where $\alpha, \beta < 0$ and $\Gamma(\cdot)$ denotes the gamma function. The parameters of the beta distribution control the sharpness of the forecasts (see Section 1.6), and the reliability of the forecasts is modelled using the linear function

$$\mu_{y|p} = a + bp, \qquad (4.7)$$

where $\mu_{y|p}$ denotes the conditional probability of outcome y given a particular forecast value p. A perfectly reliable, or calibrated, forecast is indicated where $a = 0$ and $b = 1$. To generate a synthetic time series of outcomes in this experiment, the occurrence of the binary event is determined as follows:

$$Y_i = \begin{cases} 1, & \text{if } u_i \leq a + bp(Y_i). \\ 0, & \text{if } u_i > a + bp(Y_i), \end{cases} \qquad (4.8)$$

where u_i is an independent uniform $[0, 1]$ random variable. Serial dependence is induced in the corresponding time series of forecasts by first transforming them to standard Gaussian variates (see Murphy [134]), that is

$$z_i = \Phi^{-1}[F(p_i)], \qquad (4.9)$$

where $\Phi^{-1}[\cdot]$ denotes the quantile function for the standard Normal distribution, and $F(p_i)$ denotes the CDF of the beta distribution. Next, a first-order autoregressive process is applied to the transformed forecasts to induce serial correlation, so that

$$z_{i+1} = \varphi z_i + \sqrt{1 - \varphi^2}\varepsilon_i, \qquad (4.10)$$

where φ is the lag-1 autocorrelation in the time series of the standard Gaussian variates, and ε_i denotes the Gaussian noise component of the autoregression.

After this step, the forecast values p_i can be derived by reversing the transformation in Eqn. (4.9). The parameters of the forecast-outcome time-series model described above determine the skill of the forecasts, the climatological probability of the event, and the degree of serial correlation in the sequential forecasts. Wilks [216] finds that, as the sharpness of the forecasts increases, the actual degree of induced lag-1 autocorrelation r_1 in the sequential forecasts diverges negatively from φ. The autocorrelation function (ACF) is used in all three case studies in this chapter to quantify serial correlation. Note that the ACF cannot be used to quantify nonlinear serial correlation which, if present in evaluation data, can result in inflated score sampling variances as illustrated for the first time in Section 4.4. The degree of inflation in the sampling distributions of the scoring rule induced by serial correlation in the forecast and outcome time series is assessed here by comparing the empirical statistical properties of the ignorance score computed from the correlated time series with bootstrap resamples of the time series. This is a slightly different approach to Wilks [216], who compares the empirical statistical properties of the Brier score with those of the analytical solutions. Estimates of scoring rule sampling variances made using the two approaches should be equal, however, since they are both made under serial independence.

ESS corrections are made by Wilks [216] using the equation derived by fitting the ratio of the analytical-to-empirical Brier score sampling variances with respect to the parameters of the LCBR model, given by

$$\frac{N'}{N} = \frac{1 - (1 - \mu_y)[b(1 - BS)r_1]^2}{1 + (1 - \mu_y)[b(1 - BS)r_1]^2}, \tag{4.11}$$

where BS is the Brier score and N' is the effective sample size. As previously discussed, Eqn. (4.11) is only applicable to the LCBR model and the Brier score. Sample size corrections for all other systems in this chapter using IGN are performed with the approximate method (described later in Section 4.5). In general, one does not expect analytical corrections to be at hand.

Figure 4.2 shows a clear inflation of the sampling variances of the IGN esti-
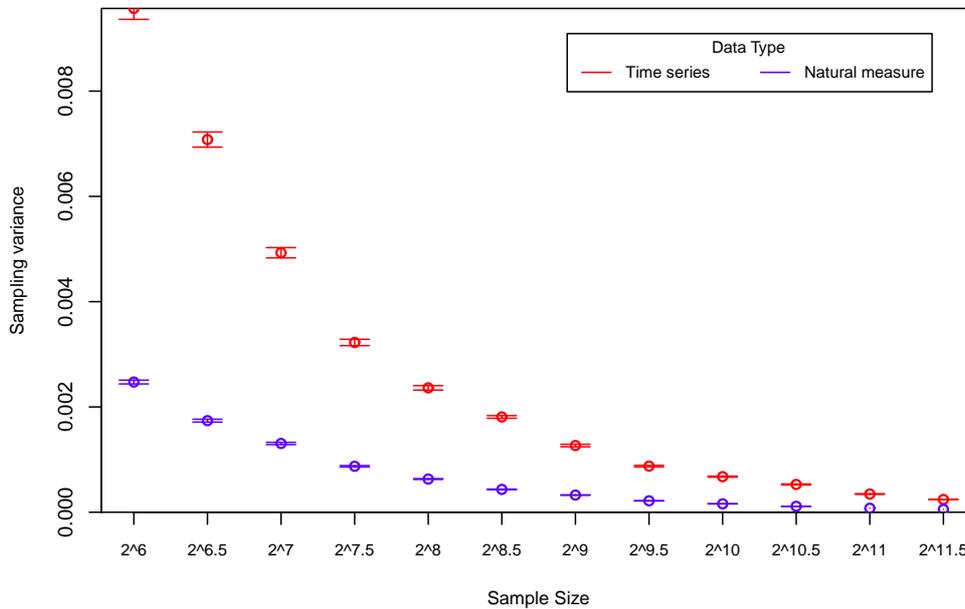
Figure 4.2: LCBR model forecast skill statistics under serial dependence: sampling variances of IGN estimates computed from $N = 2^{10}$ simulations correlated time series ($r_1(y) \approx$ 0.8; red circles) of reliable forecasts ($a = 0, b = 1$) of a low probability event ($\mu_y = 0.05$) and bootstrap resamples ($r_1(y) \approx 0$; blue circles), both with $5\% - 95\%$ uncertainty intervals. The sampling variances computed from the serially correlated IGN statistics exhibit inflation relative those computed from non-serial correlated IGN statistics. The forecasts are generated from a beta distribution with parameters $\alpha = 0.0333$, $\beta = 0.6333$.

mates of reliable forecasts ($a = 0, b = 1$) of a low probability event ($\mu_y = 0.05$) generated from the LCBR model. The difference between the sampling variance under serial dependence and serial independence decreases with increase in sample size as expected indicating the convergence of the score statistics onto the true score. To demonstrate the effect on statistical inference of the forecast skill estimates, Wilks's approach of computing the probability coverage of 95% confidence intervals is followed here. The probability coverage estimates are calculated as the relative frequency, out of $N = 2^{10}$ simulations, of the confidence interval including the true values of the BS and IGN, considered equal

Coverage of 95% Confidence Intervals for Ignorance Estimates
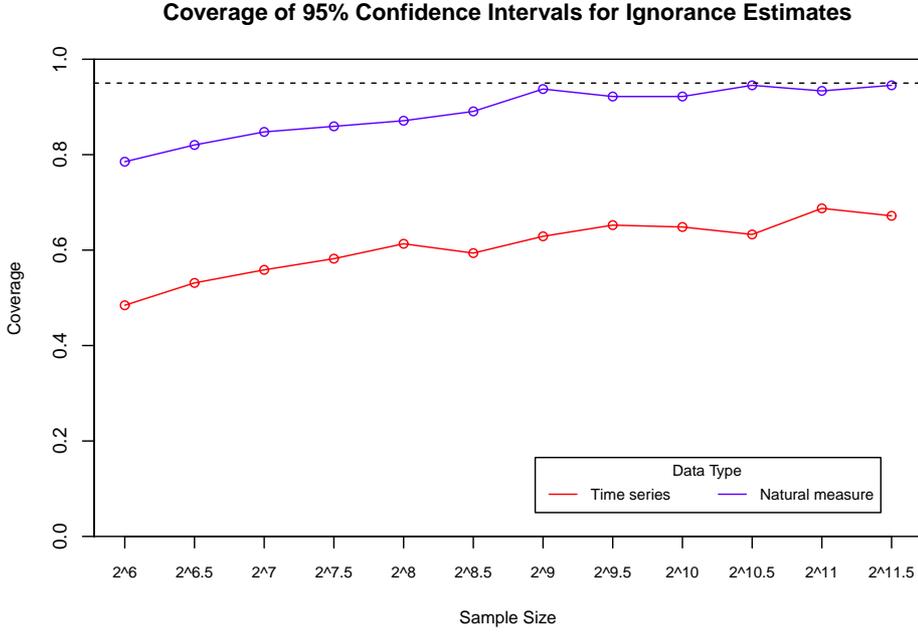
Figure 4.3: Statistical inference of LCBR model forecast skill under serial dependence: probability coverage of 95% confidence intervals for $N = 2^{10}$ IGN estimates computed from a serially correlated time series $(r_1(y) \approx 0.8)$ of reliable forecasts $(a = 0, b = 1)$ of a low probability event $(\mu_y = 0.05)$ and bootstrap resamples $(r_1(y) \approx 0$; blue circles), both shown with increasing sample size. The plot demonstrates that confidence intervals are too compact under serial dependence by showing that the probability coverage of the confidence intervals for the serially correlated IGN statistics is lower than those for the non-serially correlated IGN statistics. As N increases, the probability coverages of both converge onto the nominal 95% coverage (dashed line) but a larger sample size is required for the former to do so. The values of lag-1 autocorrelation, climatological probability, and model parameters are identical to Fig. 4.2.

to the expectations of those scores. The former is given with respect to the parameters of the LCBR model as

$$E[BS] = (\sigma_p^2 + \mu_p^2)(1 - 2b) - 2a\mu_p + \mu_y. \tag{4.12}$$

where μ_p and σ_p^2 are the first and second moment of the beta-refinement (forecast) distributions. $E[IGN]$ cannot be computed in the same manner so a sufficiently large sample size $(N = 2^{11.5})$ is used to determine the true IGN score.
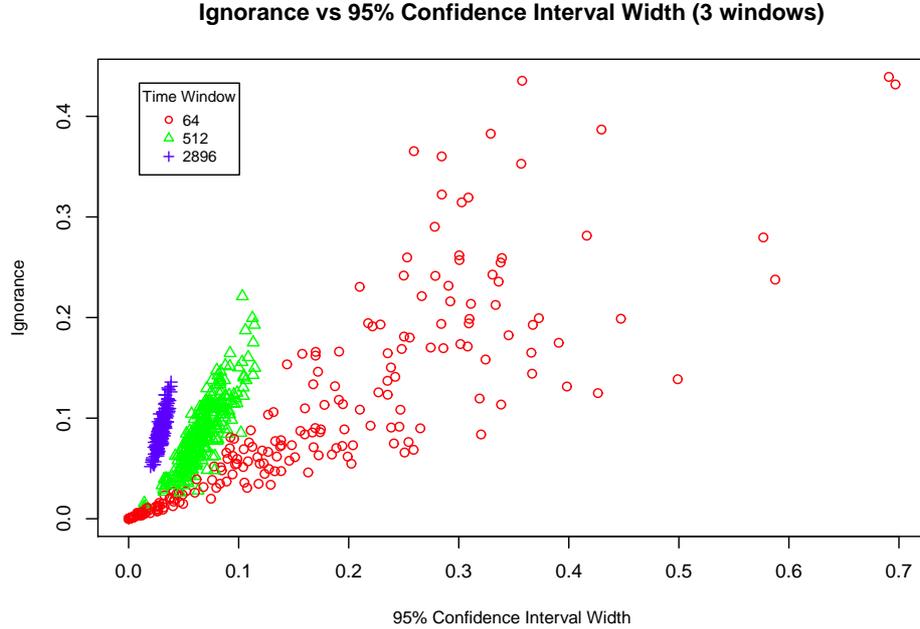
Figure 4.4: Statistical inference of LCBR model forecast skill under serial dependence: IGN estimates of correlated time series plotted against 95% confidence interval widths computed from the IGN statistics of a correlated time series $(r_1(y) \approx 0.8)$ of reliable forecasts $(a = 0, b = 1)$ of a low probability event $(\mu_y = 0.05)$. The plot shows how confidence intervals tend to be too narrow under serial dependence where forecasts are more skilful and where sample sizes are too small. The values of lag-1 autocorrelation, climatological probability, and model parameters are identical to Fig. 4.2.

Figure 4.3 compares the probability coverage of the 95% confidence intervals for the empirical IGN estimates for serially correlated $(r_1(y) \approx 0.8)$ and serially independent $(r_1(y) \approx 0)$ time series of evaluation data with increasing sample size. Both probability coverage curves converge onto the nominal 95% coverage (dashed line) with increase in sample size but the probability coverage of confidence intervals computed from the correlated time series is more insufficient, and even at a sample size of $N = 2^{11.5}$ lies below the nominal range. Figure 4.4 illustrates the relationship between 95% confidence interval width and both forecast skill and sample size. Note that the width of the confidence interval shrinks with increase in skill, and the upper bound on the widths also decreases

with increase in sample size. The excessive precision of confidence intervals under serial dependence is demonstrated by the exhibited relationship.

Repeating the experiment with calibration parameter $b = 0.8$ results in lower degrees of sampling variance inflation (plots not shown), reflecting the expected result from Wilks [216] that more skilful forecasts of serially correlated outcomes are also serially correlated. The experimental results from the LCBR model show agreement with Wilks's results demonstrating misleading statistical inference of forecast skill where serial correlation is transmitted from outcomes to forecast evaluation statistics, and exacerbation of the effect for more skilful forecasts.

4.2.2 Lorenz63
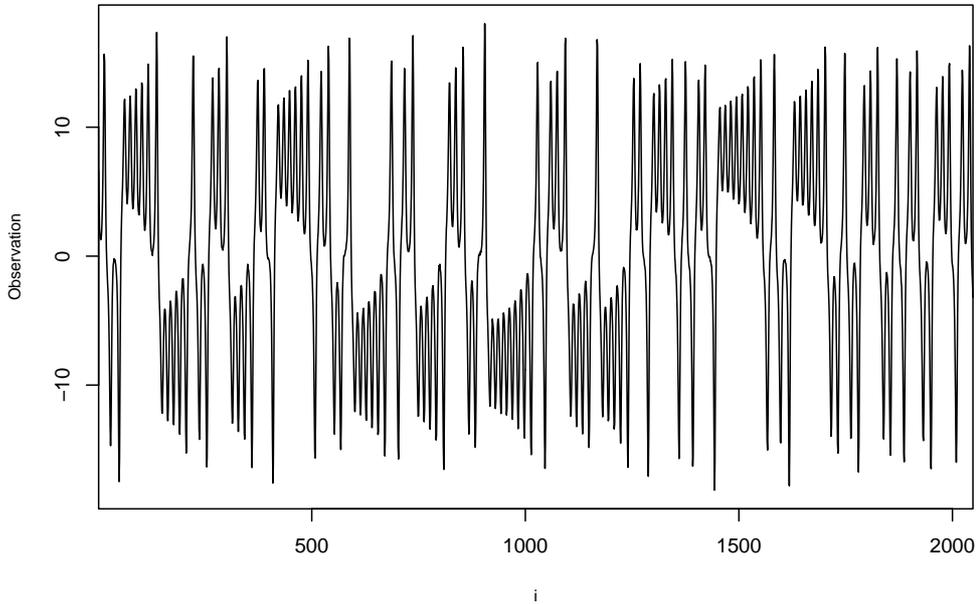
Figure 4.5: Lorenz63 observations: time series of x state variable observations illustrating the bimodal behaviour of the Lorenz63 attractor. The observations have a strong degree of linear serial correlation ($r_1(y) \approx 0.96$) measured over the whole sample size of $N = 2^{11}$ timesteps.

136

Induced inflation in scoring rule sampling variances and imprecision in forecast skill estimates under serial dependence, as demonstrated with stochastic simulations generated with the LCBR model, can also be demonstrated with a nonlinear dynamical system. The inflationary effect of serial dependence on ignorance score estimates is assessed here by evaluating sequential forecasts of the state trajectory of the Lorenz63 dynamical system [118]. The Lorenz63 system is recognisable by its double fixed point *attractor* (resembling butterfly wings) which occupy two distinct regions of state space. Consequently, the x-variable exhibits bimodal behaviour (see Fig. 4.5) which can result highly correlated sequential trajectory observations for sufficiently short time steps.

The forecasts in this experiment are produced using the KDB density construction method (see Section 1.8), while the initial conditions at each forecast initialisation and corresponding outcomes are sampled from the inverse of the stochastic observational noise model (see Section 2.1). Sequential forecast-outcome pairs are generated for a number of lead times and sample sizes, from which time series of forecast evaluation scores can be compiled. As with the LCBR model experiment above, reference sets of i.i.d. score estimates are created by bootstrap resampling from the score time series.

Representative examples illustrating the effect of serial dependence on forecast evaluation are shown in Figs. 4.6 and 4.7 for a forecast lead time of $\tau = 1.0$ Lorenz unit. Linear serial correlation in the outcome time series ($r_1(x) \approx 0.94$) is transmitted to the time series of score statistics ($r_1(IGN) \approx 0.5$), resulting in inflation of the variance of the score sampling distribution, lack of probability coverage in confidence intervals, and overconfidence in skill.

The transmission of serial correlation from data to forecast evaluations can be interpreted by considering that the underlying distribution of score statistics is time-dependent, implying that the autocovariance of the score is expected to be non-zero. In mathematical notation, the autocovariance $R(\tau)$ of a score S is
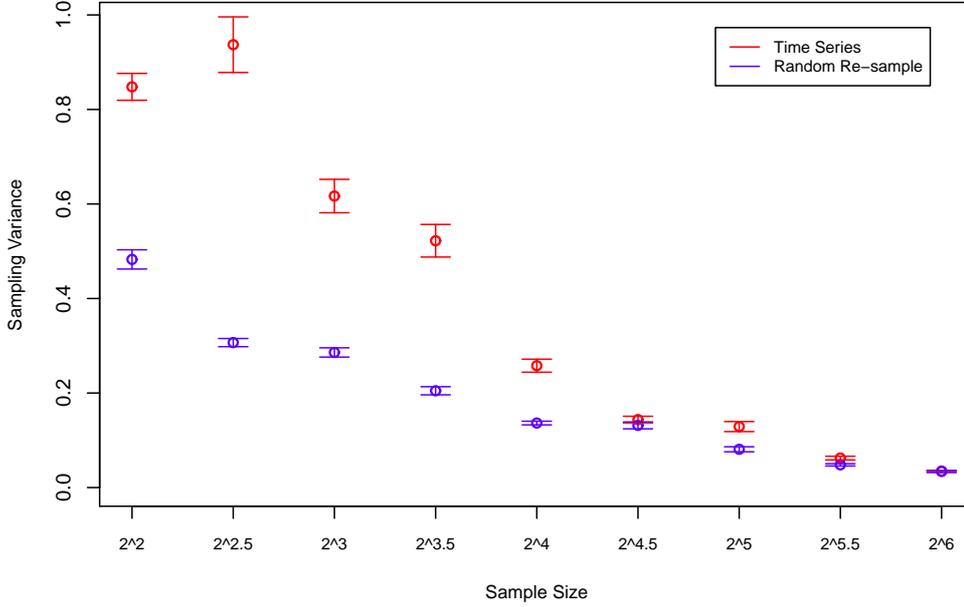
Figure 4.6: Lorenz63 forecast skill statistics under serial dependence: Sampling variances of a) ignorance estimates computed from forecasts of a correlated time series of Lorenz63 observations ($r_1(y) \approx 0.94$; red circles) and b) the natural measure of ignorance estimates ($r_1(y) \approx 0$; blue circles), both with $5\% - 95\%$ uncertainty intervals. There is a clear inflation of the sampling variances until at least a sample size of 2^5 showing that the serial correlation in the observations is transmitted to the score statistics.

given as

$$R(\tau) = E[(S_t - E[S_t])(S_{t+\tau} - E[S_{t+\tau}])] \tag{4.13}$$

$$= E[S_t S_{t+\tau}] - E[S_t]E[S_{t+\tau}] \tag{4.14}$$

$$\neq 0, \tag{4.15}$$

where $E[S_t] \neq E[S_{t+\tau}]$ are the means of the score distributions at time t and time $t+\tau$ (lag τ) respectively. The non-zero result arises under serial dependence since, only where S_t and $S_{t+\tau}$ are independent, is it true that

$$E[S_t S_{t+\tau}] = E[S_t]E[S_{t+\tau}]. \tag{4.16}$$

Although the inflation of the scoring rule sampling variance induced by serial
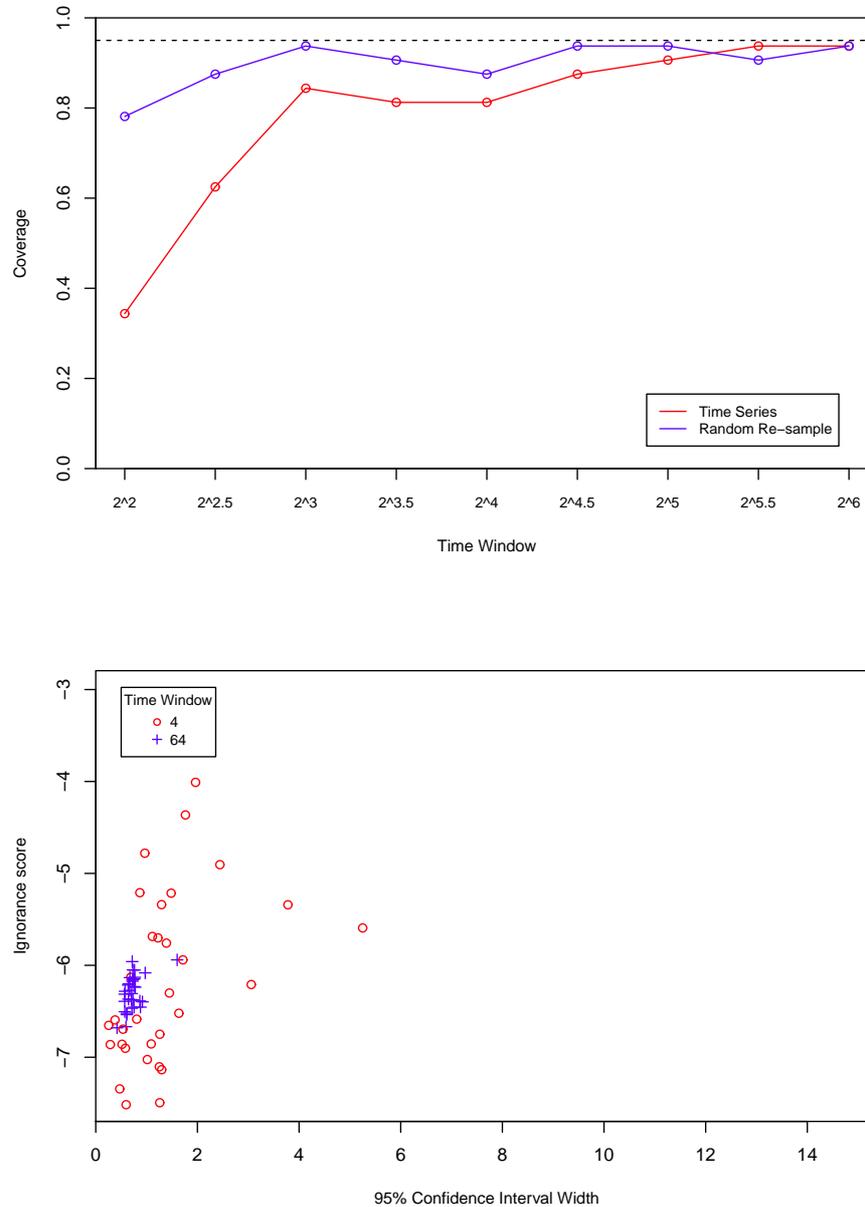
Figure 4.7: Statistical inference of Lorenz63 forecast skill under serial dependence: probability coverage of 95% confidence intervals for increasing sample size (top), and IGN estimates of correlated Lorenz63 forecast time series plotted against 95% confidence interval widths (bottom). The two plots show the tendency of confidence intervals to be too compact under serial dependence where forecasts are more skilful or sample sizes are too small.

dependence has only been demonstrated for probabilistic forecasts in this section, it can also easily be shown for point forecasts.

4.3 Case Study 2: Non-transmission of linear serial correlation to forecast evaluation statistics

While there are forecast evaluation scenarios where linear serial correlation in data can lead to misleading estimates of forecast skill (see Section 4.2 and Wilks [216]), the presence of linear serial correlation in an observational time series is not a sufficient condition for estimates of forecast skill to be misleading. For the first time, it is demonstrated that there are forecasting scenarios where the distribution of a forecast evaluation measure is not time-dependent so serial correlation in a time series of observations is not transmitted to the forecast evaluation statistics. Without the inflationary effect induced by serial dependence on the sampling variance of the scoring rule, ESS corrections are not required and statistical inference of skill can be made under the assumption of serial independence.

Two stochastic target systems are employed in this section to show how serial dependence can be transmitted from sequential evaluation outcomes (i.e. observational data) to the corresponding point forecasts if they are sufficiently skilful, but not to the forecast evaluation statistics. The first is a first-order autoregressive process, and the second is a testbed system designed to simulate Atlantic basin hurricane annual counts using a Poisson process.

4.3.1 AR(1) process

Consider a time series of observations s_t generated from a *first-order autoregressive* (AR(1)) process, first introduced by Yule [219] to model sunspots. An

observation s_t at time t is given by

$$s_t = \varphi s_{t-1} + \epsilon_t, \tag{4.17}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the normally distributed random noise component of the AR(1) process. Since the noise is a Gaussian process, the observations s_t are also Gaussian distributed. The model parameter φ controls the degree of autocorrelation in the time series, and the process is *weak-sense stationary* for values $|\varphi| < 1$, meaning that the mean $E[s_t]$ and covariance $Cov[s_t, s_{t+\tau}]$ are constants in time. In that case, as φ approaches a value of 1, the influence on s_t from the previous observation s_{t-1} increases.

Let X_t represent a 1 step ahead forecast of the observation s_t generated from an imperfect model using the observation s_{t-1} so that

$$X_t \sim \mathcal{N}(s_{t-1}, \sigma_\epsilon^2). \tag{4.18}$$

Hence, the forecasts are, like the observations, Gaussian distributed, and exhibit a similar degree of serial correlation determined by the parameter φ.

The effect of serial dependence is now assessed by examining the differences between forecast evaluation statistics from a number of numerical experiments. As in Section 4.2, the sampling variances of scoring rule estimates computed from a serially correlated time series of forecast-outcome pairs are compared with bootstrap resamples from the same time series for different sample sizes. Figure 4.8 shows the estimates of the sampling variances of the score statistics for the correlated forecast time series and bootstrapped forecasts over increasing time windows (i.e. sample sizes). In this case, however, the sampling variance estimates for both sets of forecast-outcome pairs at all time windows lie within 95% uncertainty intervals constructed for the sampling variance of score estimates computed from a time series of forecast-outcome pairs where each of the pair are both standard normal distributed and, hence, i.i.d.. The containment of the estimates within the uncertainty interval indicates that sampling variances of the time series and bootstrap resamples score estimates are statistically
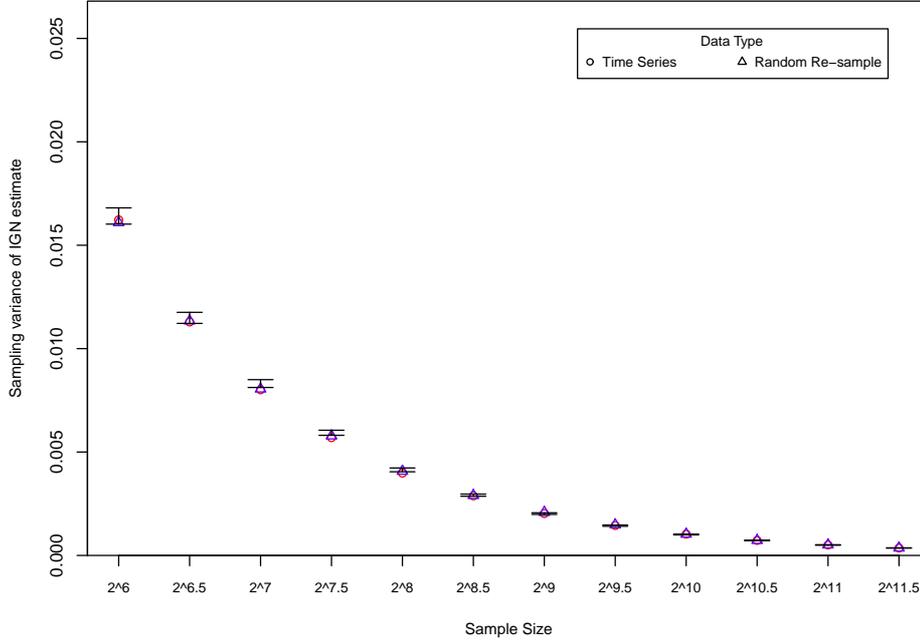
Figure 4.8: AR(1) forecast skill statistics under serial dependence: Estimates of sampling variances of IGN estimates for an AR(1) observation time series ($\varphi = 0.9$; red circles) and the bootstrapped observations (blue circles). Both sets of points all lie within 95% uncertainty intervals constructed from $N_{boot} = 2^7$ bootstrap resample estimates of the sampling variance of Gaussian distributed forecasts showing that there is no significant difference between either of the sampling variances and uncorrelated Gaussian forecasts. Each sampling variance estimate contains 2^8 IGN estimate samples.

indistinguishable both from the sampling variance of the standard normal forecast score estimates, and from each other. Hence, the distributions of the score statistics of the serially dependent sequential forecasts and serially independent bootstrapped forecasts can both be considered Gaussian and identical.

The indistinguishability of the sampling variances reflects the fact that the forecast errors of both datasets are normally distributed and independent (i.i.d.) (i.e. the score distribution is time independent) so that there is no serial correlation present in the skill score time series, and hence, no inflation of the sampling variances. The independence of the forecast errors satisfies Eqn. (4.16). The

lack of serial correlation in the score statistics is also evident in Fig. 4.9 where a 1-step time delay scatterplot reveals almost no linear relationship between ignorance at time t and time $t+1$. While an example of zero sampling variance

AR(1) IGN Lag−1 Delay Map (phi=0.9)
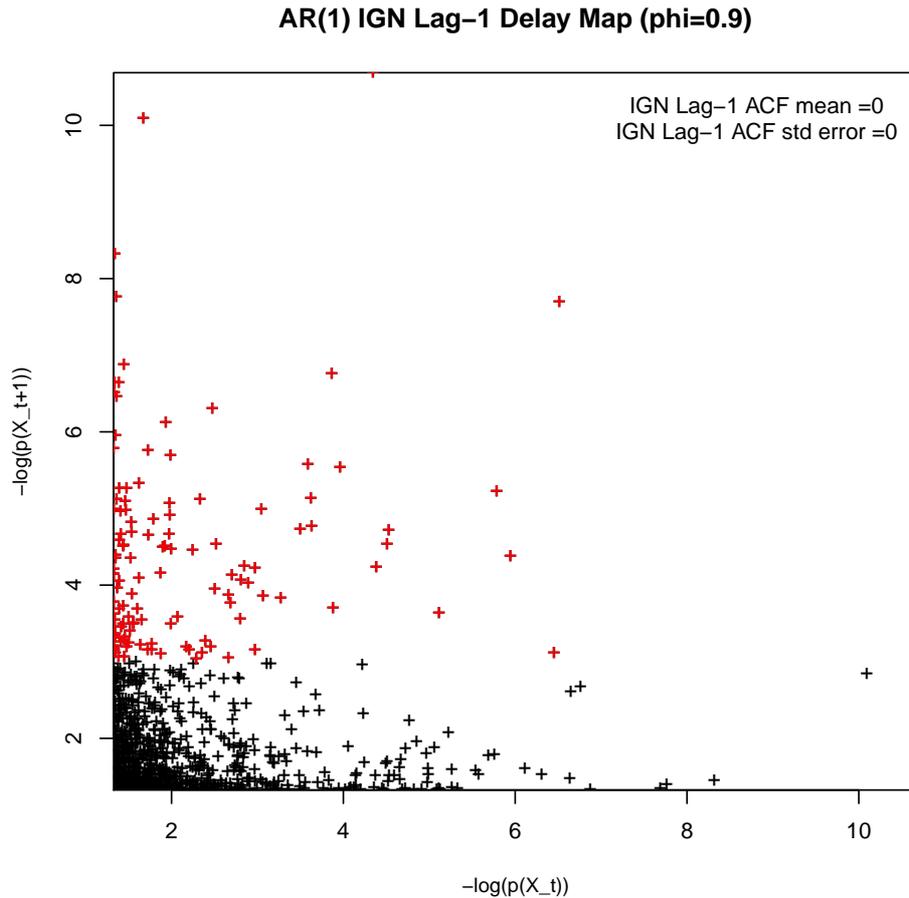
Figure 4.9: AR(1) forecast skill statistics under serial dependence: Example of a 1 step delay plot showing the lack of linear serial correlation in a single IGN time series of sample size $N = 2^{10}$ computed from serially correlated observations ($\varphi = 0.9$). The red coloured points, denoting ignorances scores $-log(p(s_{t+1})) > 3$ (signifying less skilful forecasts) at time $t+1$ (y-axis), also indicate that forecasts are more skilful at time t, highlighting the lack of serial correlation. The mean and standard error of the lag-1 autocorrelation values of the $N_{boot} = 2^8$ replications of time series are not significantly different from zero.

inflation has been demonstrated above, a counterexample is now also described where the linear serial correlation in the time series of observations and fore-

casts generated under the AR(1) process *is* transmitted into the corresponding scoring rule time series. If the forecast PDF is non-state dependent, so that a realisation of a score at time t is only dependent on the outcome s_t, then the sequential scores will also be serially dependent. Consider a "perfect" climato-
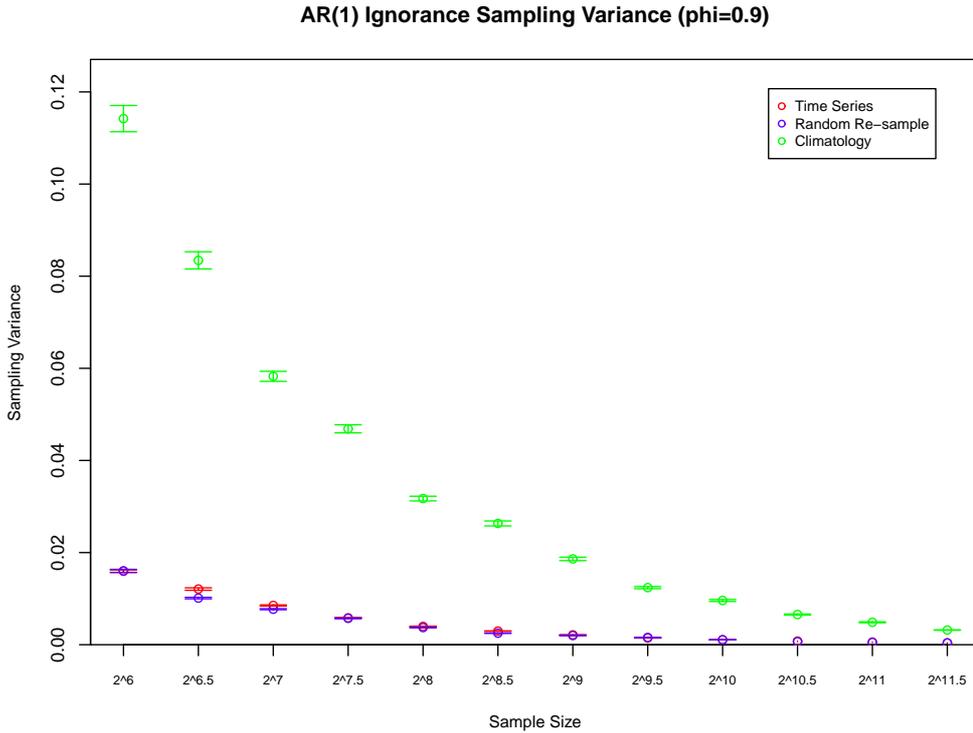
Figure 4.10: AR(1) forecast skill statistics under serial dependence: Mean sampling variance of the IGN for AR(1) time series of forecasts of serially correlated observations ($r_1(y) = 0.9$; red line), forecasts of bootstrapped observations ($r_1(y) \approx 0$; blue line), and climatological forecasts (green line), all with $5\% - 95\%$ uncertainty intervals computed from $N_{boot} = 2^7$ samples. There is a clear inflation of the climatological sampling variance

logical Gaussian forecast which is non-state dependent be defined as

$$X_{clim} \sim \mathcal{N}\left(0, \frac{\sigma_\epsilon^2}{1 - \varphi^2}\right), \tag{4.19}$$

since $E(s_t) = 0$. The variance of the climatological forecast distribution is

derived as follows:

$$
\begin{aligned}
Var(X_{clim}) &= Var(s_t) \\
&= Var(\varphi s_{t-1} + \epsilon_t) \\
&= \varphi^2 Var(s_{t-1}) + \sigma_\epsilon^2 \\
&= \frac{\sigma_\epsilon^2}{1 - \varphi^2},
\end{aligned}
\tag{4.20}
$$

since $Var(s_t) = Var(s_{t-1})$. Hence, the forecast distribution is time indepen-dent. The sampling variances of the ignorance statistics for the climatological forecasts, serially correlated forecasts, and bootstrapped forecasts over increas-ing sample sizes are shown in Fig. 4.10. The inflationary effect on the climato-logical forecast sampling variance is clearly visible from the fact that the green curve lies well above the other two curves. The demonstration of both accurate and inaccurate estimates of forecast skill with a single data-generating system (i.e. an AR(1) process) in this section highlights the importance of understand-ing how serial dependence is transmitted from a time series of observations to the forecast evaluation statistics. Both the data-generating system and the fore-cast model need to be considered when determining whether serial dependence will have an impact on the inference of forecast skill.

4.3.2 Testbed hurricane system

A stochastic testbed hurricane system is now introduced to examine the effect of serial dependence on forecast evaluation statistics in a scenario more analogous to real geophysical phenomena than the other systems employed so far in this chapter. Consider a hurricane system in which the mean number of storms follows a cycle of T_p years, while the number of storms in any given year is a random variable. The annual storm counts are generated according to a stochastic Poisson process given as

$$
Y_t \sim Pois(\lambda(t)),
\tag{4.21}
$$

where Y_t is the number of hurricanes in a given year t, and has a sinusoidal time-dependent mean parameter λ which is determined by the equation

$$\lambda(t) = A sin\left(\frac{2\pi t}{T_p}\right) + y_c. \tag{4.22}$$

A stochastic Poisson process has been used to simulate hurricane counts since it is a simple model for discrete response variables [2], and has be shown to be consistent with the behaviour of many underlying physical processes [46]. The parameters of Eqn. (4.22) have been set to realistically simulate real Atlantic basin CAT1-5 hurricane counts (i.e. $A = 2.5$, $y_c = 6$, and $T_p = 24$). With these parameter values, the lag-1 autocorrelation of a time series is measured to be $r_1(y) \approx 0.4$. Now let a forecast model of the annual simulated hurricane counts be defined by a "squared Gaussian" distribution so that it is structurally incorrect (i.e. imperfect), that is, for a given random variable

$$V_t \sim \mathcal{N}(\mu, \sigma^2), \tag{4.23}$$

the random variable

$$X_t = \lfloor V^2 + 0.5 \rfloor, \tag{4.24}$$

represents the number of annually forecast hurricanes where $\lfloor \cdot \rfloor$ is the floor function. In addition, the model parameters μ and σ have been fitted to each of the 24 phases of the hurricane system's cycle by minimising the relative entropy (see Section 2.5) of the forecast PDF p and the true PDF q. Hence, although the true PDF is unknown, it is assumed that the forecaster knows that there is a 24 year cycle and the fitting process can be regarded as the model training period. Expressed mathematically, the parameter estimates are given by

$$(\hat{\mu}, \hat{\sigma})_\phi := \arg \min_{\mu, \sigma} -q_\phi(Y_j) \sum_{j=0}^{M} log_2\left(\frac{p_\phi(Y_j)}{q_\phi(Y_j)}\right), \tag{4.25}$$

where $q_\phi(Y_j)$ and $p_\phi(Y_j)$ are the true and forecast probabilities respectively of the jth outcome occurring at phase ϕ.

A time series of synthetic annual forecast-outcome pairs $(p_\phi(Y_t), Y_t)$ is generated by sampling the outcome data from the Poisson distribution defined in

Eqn. (4.21) with the initial phase being selected at random from a discrete uniform distribution (i.e. $\phi_t \sim \mathcal{U}\{1, \ldots, 24\}$). The forecast probability $p_\phi(Y_t)$ is determined according to the fitted parameters $(\hat{\mu}, \hat{\sigma})_\phi$ of the forecast model.

Sampling variance of relative IGN estimate
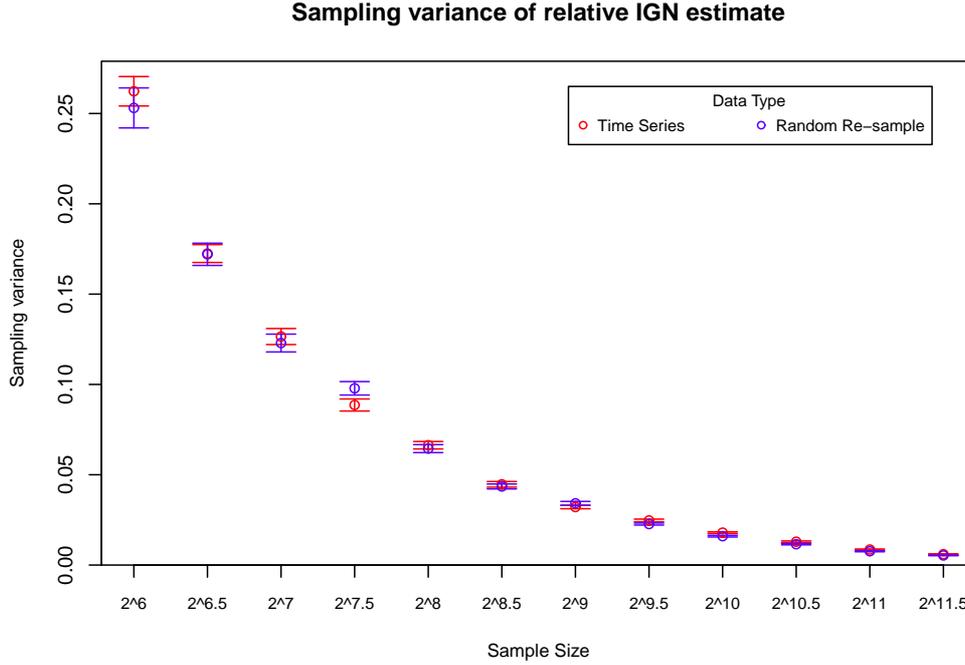
Figure 4.11: Hurricane forecast forecast skill statistics under serial dependence: Sampling variances of IGN estimates computed for 2^8 time series of serially correlated observations ($r_1(y) \approx 0.4$; red line) and bootstrapped observations ($r_1(y) \approx 0$; blue line).

The sampling variances of the forecast skill estimates computed from the generated correlated time series are compared with bootstrap resamples for increasing sample sizes to assess the effect of serial dependence on the forecast evaluation statistics. Figure 4.11 shows that, for time window lengths greater than ~ 64 years, there is no statistically significant evidence of score sampling variance inflation. Like the evaluation of the forecasts of the AR(1) time series in the previous section, the non-effects of serial dependence occur because the score statistics are serially independent.

The results presented in this section demonstrate that there are forecasting scenarios where serial correlation in data does not result in misleading estimates

of forecast skill. Even with high degrees of lag-1 autocorrelation in the time series generated from both the AR(1) ($r_1(y) \approx 0.9$) and the toy hurricane system ($r_1(y) \approx 0.35$), there is almost no lag-1 autocorrelation present in the scoring rule time series ($r_1(IGN) \approx 0$), and no induced inflation of the score sampling variance. Incidentally, the lack of serial correlation present in the forecast skill statistics can also be demonstrated for a nonlinear stochastic process. Consider again an autoregressive process, but now defined as

$$s_t = \varphi s_{t-1} + \epsilon_t^2. \tag{4.26}$$

so that the noise is no longer Gaussian, and the process is not linear. A similar degree of linear serial correlation is present in a time series generated from this process compared to the linear AR(1) process defined in Eqn. (4.17) at all values of $\varphi \in [0, 1]$. If forecasts are constructed in the same way as those in the AR(1) forecast scenario above (see Eqn. (4.18)), then the score statistics are i.i.d., and there is no induced inflation of the sampling variance of the score. Hence, linearity in the data generating process itself is not a sufficient condition for linear serial correlation to be present in the forecast evaluation statistics. A nonlinear process can also be employed to show that nonlinear serial correlation in an observational time series can result in misleading estimates of forecast skill. Case study 3 examines this scenario in Section 4.4.

4.4 Case Study 3: nonlinear serial correlation in data; linear serial correlation in skill score statistics

In the previous two case studies in Sections 4.2 and 4.3, examples have been given to demonstrate the how *linear* serial correlation in a time series of observations can have both an effect and a non-effect on estimates of forecast skill. It is shown in this section that there are systems that do not exhibit linear serial

correlation yet serial dependence still has an effect on estimates of forecast skill. As explained in Section 4.2, the skill score distribution may be time dependent given that forecast error is dependent on the state of the system. The dependence of the forecast skill statistics results in a degree of linear serial correlation in the skill score time series and, hence, to fallacious estimates of skill.

4.4.1 Logistic map

The logistic map is a well-known 1-dimensional nonlinear dynamical system which was popularised by May [126] as an ecological model of the dynamics of breeding populations in time. The mathematical form of the logistic map is expressed as

$$x_{i+1} = f(x_i) \tag{4.27}$$

$$= ax_i(1 - x_i), \tag{4.28}$$

where $x_n \in (0, 1)$ represents the state of the map. Figure 4.12 shows the logistic map with parameter value $a = 4.0$. Even though there is no measurable degree of linear serial correlation ($r_1(y) \approx 0$), a nonlinear relationship between sequential data is evident in the displayed curve. The effect of serial dependence in observations of a nonlinear dynamical system is demonstrated here by utilising a simple forecast model based on a Gaussian distribution to generate sequential forecasts of the state of the logistic map iterated forward in time. The initial state is uniformly sampled from the support of the logistic map $x \in [0, 1]$. Consider a 1 step-ahead Gaussian forecast PDF ρ_{fcst} at time $i + 1$ with a standard deviation σ_{fcst} which is dependent on the gradient of the logistic map (see Eqn. (4.28)) at time i. The expected ignorance, or entropy, of the forecast ρ_{fcst}, as defined by Roulston and Smith [172], is given as

$$E[IGN] = -\frac{1}{ln2} \int_{-\infty}^{\infty} \rho_{truth} ln\rho_{fcst}(x) \, \mathrm{d}x \tag{4.29}$$

$$= \frac{1}{2ln2} \left[ln2\pi + ln\sigma_{fcst}^2 + \frac{\sigma_{truth}^2 + (\bar{x}_{truth} - \bar{x}_{fcst})^2}{\sigma_{fcst}^2} \right]. \tag{4.30}$$
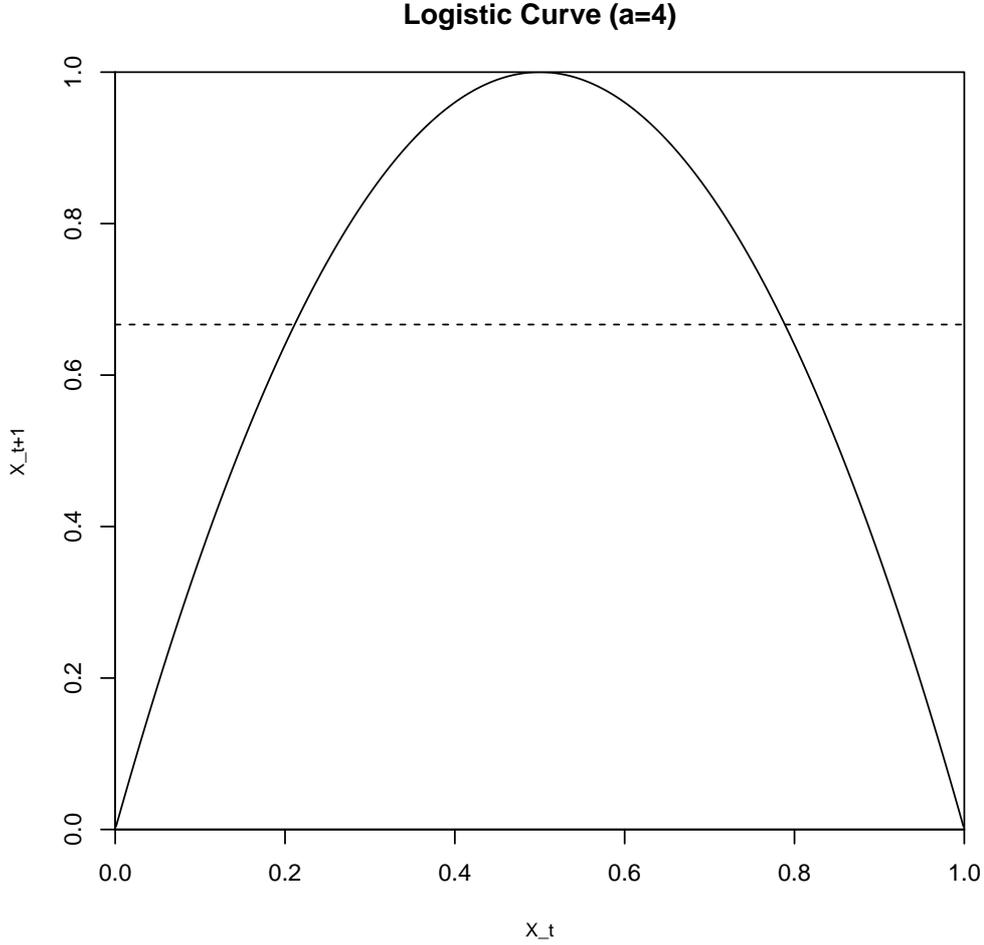
Figure 4.12: Logistic map: the logistic map given by $x_{i+1} = 4x_i(1 - x_i)$. The parabolic shape of the curve indicates the presence of serial dependence although there is zero lag-1 autocorrelation ($r_1(y) \sim 0$). The linear regression fit (dashed line) has a zero slope which also hints at the lack of a linear dependency between sequential observations.

Consider now that the forecast PDF is perfect so that there is no bias in the forecast (i.e. $\bar{x}_{fcst} = \bar{x}_{truth}$), and the variance of the forecast equals that of the truth (i.e. $\sigma_{fcst} = \sigma_{truth}$). The last term on the right hand side of Eqn. (4.30) then simplifies to unity. Now consider a variant of the $E[IGN]$ of a perfect Gaussian forecast, referred to as *Theoretical Ignorance Expected*, which evaluates the expected ignorance of a one step ahead forecast. The theoretical ignorance expected score (TIE) is so-called because it is based on the assumption

Ignorance (relative to I.C. uncertainty) Delay Map (r=4)
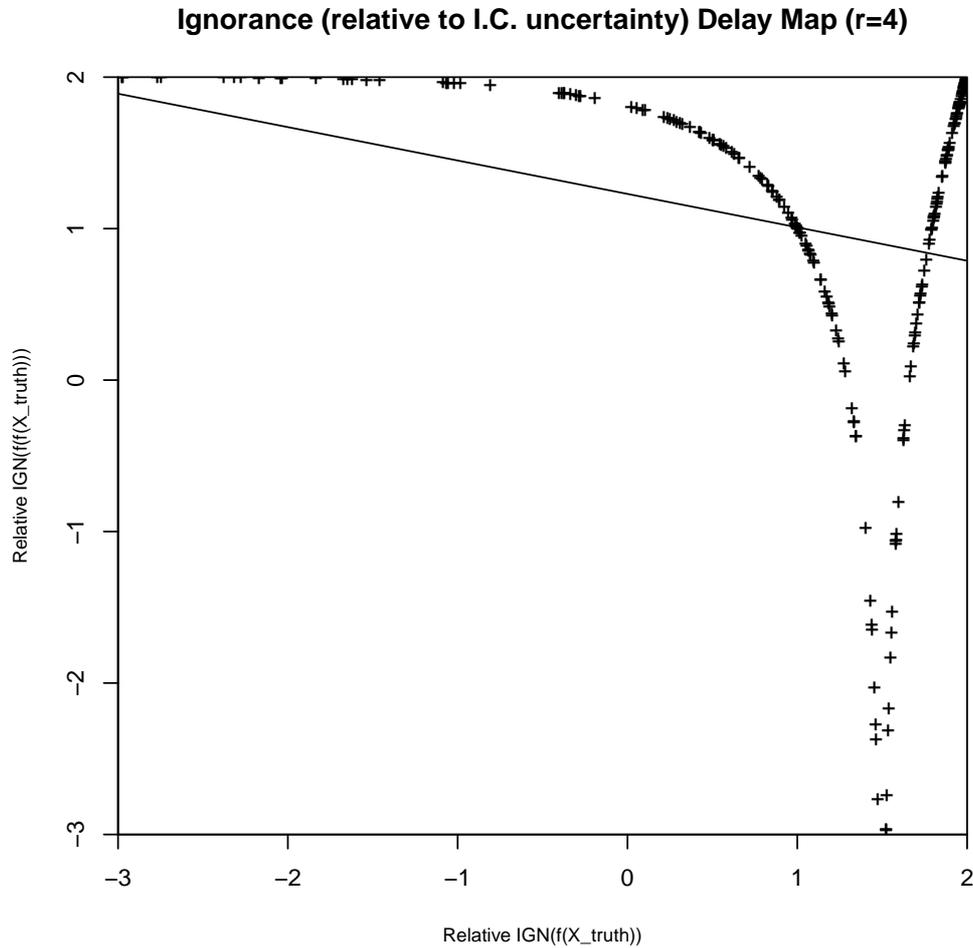
Figure 4.13: Logistic map forecast skill: theoretical ignorance expected (relative to TIE with $\sigma_{truth} = 1/128$) at $f(f(x))$ and $f(x)$ of a single logistic map time series ($\alpha = 4.0$) of sample size $N = 2^8$. A linear fit is shown as a dashed line and the value of the lag-1 ACF of the time series is $r_1(TIE) = -0.26$, both indicating a degree of negative linear serial correlation in the skill score time series.

that there is an arbitrary uncertainty in the underlying distribution of system states at time $i + 1$ which is dependent on the gradient of the logistic map at time i. The TIE score is expressed as

$$TIE = \frac{1}{2ln2}\left[ln(2\pi) + ln(\sigma_{truth}f'(x_i))^2 + 1\right],\qquad(4.31)$$

151

Theoretical Ignorance Expected Sampling Variance
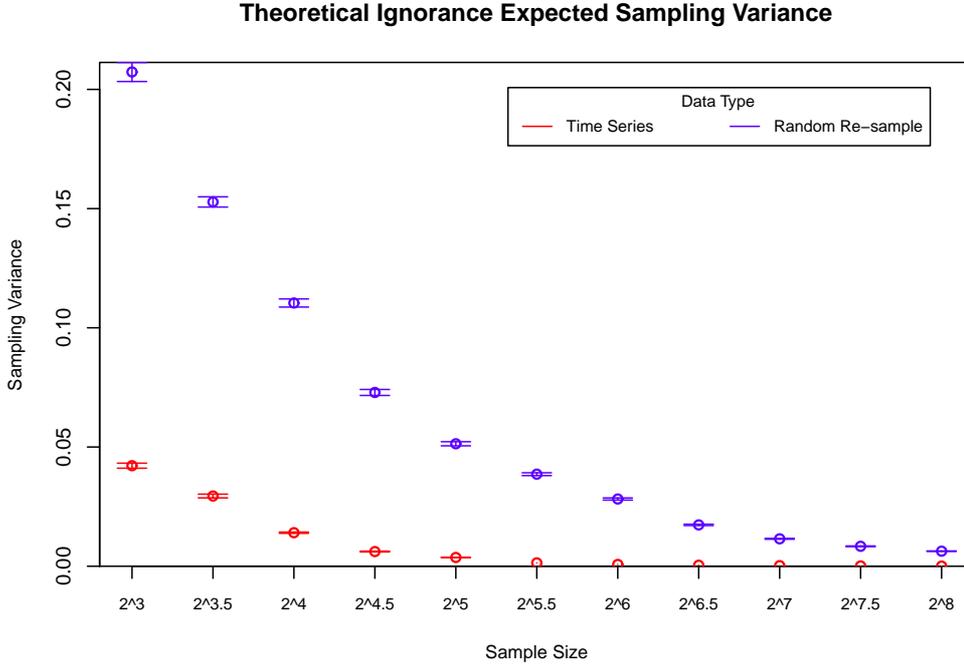
Figure 4.14: Logistic map forecast evaluation statistics under serial dependence: sampling variances of a) TIE estimates computed from forecasts of a correlated time series of Logistic map observations ($r_1(TIE) \approx -0.26$; red line) and b) TIE estimates ($r_1(TIE) \approx 0$; blue line) computed from forecasts of the natural measure of the Logistic map, both with $5\% - 95\%$ uncertainty intervals. There is a clear deflation of the sampling variances of a) until at least a time window length of 2^5 showing that the serial correlation in the observations is transmitted to the score statistics.

where $f'(x_i)$ is the first derivative of the logistic map at x_i, and σ_{truth} is the standard deviation of the underlying distribution of system states at time i. Example numerical results of the TIE of forecasts at time $i + 1$ relative to the TIE at time i, where there is zero uncertainty (i.e. $f'(x_i) = 1$), are shown in Fig. 4.13. The linear serial correlation in the skill score statistics is evident from the linear fit and the lag-1 ACF value $r_1(TIE) = -0.26$ computed from a time series of 2^8 iterations of the map. A comparison of Figs. 4.12 and 4.13 shows how serial dependence in an observation time series may be unmeasurable using the ACF, but can result in linear serial correlation in the forecast evaluation statistics.

Serial dependence in this case may prove to be particularly problematic for those wishing to demonstrate statistically significant forecast skill since it could easily be overlooked, resulting in misleading estimates of skill.

4.5 Convergence of information deficit under serial dependence

Consider two forecast models, one perfect and one imperfect, constructed to make predictions of the trajectory of a nonlinear dynamical system. The perfect model is only subject to initial condition (IC) and parameter uncertainty whereas the imperfect model is subject to structural imperfections, and IC and parameter uncertainty. If sequential probabilistic forecasts p and q are produced from the imperfect model and perfect model respectively, then there exists an expected *information deficit* in p relative to q which can be measured with relative entropy (see section 2.5), defined as

$$D(\mathbf{p}|\mathbf{q}) = \int_{-\infty}^{\infty} -(p(y) - q(y))log_2 p(y) \, \mathrm{d}y. \tag{4.32}$$

In real-world forecasting, q is not obtainable, hence the true value of the information deficit is also unknown. Indeed, precise numerical estimation of the information deficit of the forecast p is only possible for a sufficiently large sample of forecast-outcome pairs to sufficiently reduce sampling uncertainty. An alternative formulation is to estimate the information deficit [41] by contrasting the empirical ignorance of the forecast p with the ignorance expected of p if it were in fact perfect. The latter is referred to as *implied ignorance*. The information deficit can be interpreted as the difference in skill between the imperfect model and its internally perfect version, and is defined thus

$$ID = \left[\frac{1}{N} \sum_{i=1}^{N} -log_2 p(y) - S_{clim} \right] - \int_{-\infty}^{\infty} -p(y)log_2 p(y) \, \mathrm{d}y, \tag{4.33}$$

where

$$S_{clim} = \int_{-\infty}^{\infty} -p_{clim}(y)log_2 p_{clim}(y) \, \mathrm{d}y. \tag{4.34}$$

S_{clim} is the implied ignorance of the climatological forecast p_{clim} and the measure of zero skill of ignorance. The left hand term of Eqn. (4.33) in square brackets represents the empirical ignorance of the forecast while the the right hand term represents the implied ignorance of the forecast. The information deficit should converge onto its true value with increase in sample size in accordance with the law of large numbers (LLN) [189]. If serial dependence is present to some degree in the time series of observations, the sampling variance of the forecast evaluation statistic will be prone to inflation/deflation effects relative to a serially independent time series as shown in Section 4.2. In that case, the sample size required for a precise estimate of the true information deficit is further modified.

The degree of serial dependence in the information deficit statistics varies depending on the location on the Lorenz63 attractor since the degree of serial dependence also varies over state space. Moreover, the degree of serial dependence is also dependent on the level of forecast skill, as has been noted by Wilks [216] and in Section 4.2. So, the question is posed here for the first time: how long does it take to estimate the true information deficit where sequential observations are serially dependent and forecast skill varies for a given lead time? In the absence of a known analytical solution of the sampling variance of a log-based scoring rule (see Section 4.2), a *Monte Carlo* method has been employed here to sample sequential observations along the trajectory of the Lorenz63 attractor at different forecast lead times. The time durations (i.e. samples sizes) required for the estimates of the information deficit to converge to its true value determine the *time until convergence* (TUC). Comparing the time until convergence of the information deficit under serial dependence and serial independence also provides an indication of how much longer a statistician or forecaster must wait to demonstrate statistically significant forecast skill when forecasting red processes. The importance of this understanding is demonstrated in a fictitious betting scenario in Section 5.5.
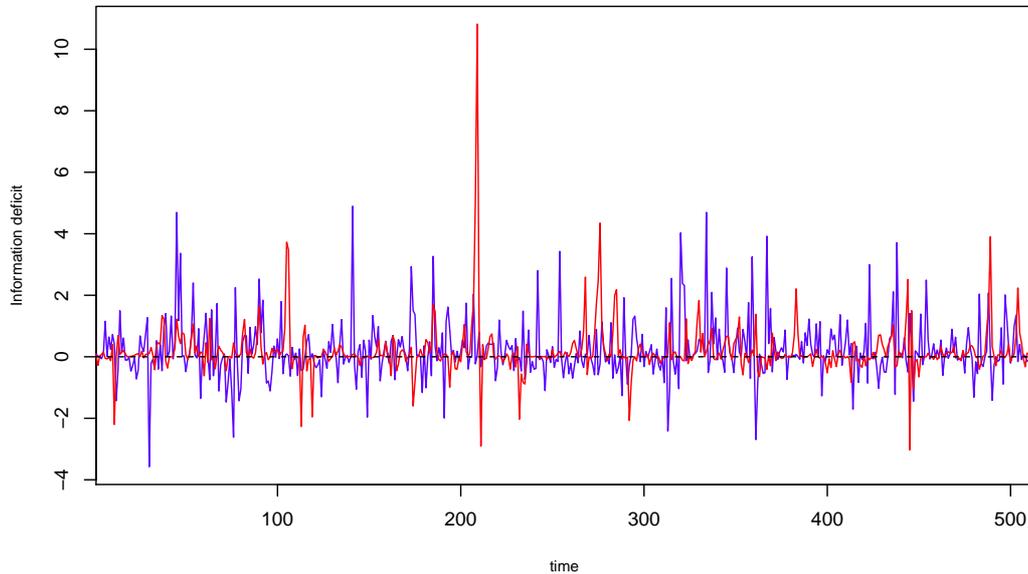
Figure 4.15: Information deficit time series: 2^9 step time series of information deficit statistics for Lorenz63 forecasts constructed using IN (blue line) and PDA (red line) data assimilation schemes ($\tau = 0.1$). The PDA forecasts are have a lower information deficit $ID_{PDA} = 0.18$ bits compared to the IN forecasts $ID_{IN} = 0.21$ bits over this time window. The differences between the information deficit values for the two forecast systems tend to be smaller than the corresponding differences in IGN, the values of which are $IGN_{PDA} = -5.30$ and $IGN_{IN} = -3.57$ for the same observation time series.

4.5.1 Experimental design

The Lorenz63 system is employed here to assess the TUC of the information deficit onto its true value as it exhibits serially correlated behaviour, and, being defined by ordinary differential equations (ODEs), has a *continuous-time* flow allowing for higher sampling rates (shorter forecast launch steps) than *discrete-time* systems such as the logistic map. Up until this point, investigation of the effect of serial dependence on forecast skill estimates in the context of the Lorenz63 system has been performed with analysis of a single state variable (x). The TUC of the information deficit is assessed here for several different forecast lead times to investigate the impact of sampling from various locations

on the attractor in *3-dimensional* space. The degree of serial dependence varies with location on the attractor. Both the perfect and imperfect Lorenz63 models are identical to those presented in Chapters 2 and 3 utilising the inverse noise data assimilation (DA) scheme[3]. A second DA scheme, called *pseudo-orbit data assimilation* (PDA) [89], is also utilised here. The PDA scheme consists of finding a sequence of system trajectories in the model state space which are consistent with the observational noise, and provides a better approximation of the initial conditions than the more simple and cost-effective inverse noise method [40]. This feature of PDA often produces more skilful forecasts than those produced from an inverse noise model. An example of a time series of ID statistics for forecasts produced using the PDA and IN schemes is shown in Fig. 4.15. Note that a robust evaluation of the relative performance of these two DA schemes is beyond the scope of this thesis. The aim is to examine the relationship between forecast skill and TUC by producing forecasts with different levels of skill. Table 4.1 lists the configurations of all of the TUC experiments using Lorenz63. The forecast launch step Δ determines the steps between forecast initialisation on the attractor, and α_{blend} denotes the blending parameter (see Section 1.5.3).

Table 4.1: Experimental Configurations for Lorenz63 Forecasts

Expt. No.	System	State variable	DA scheme	Noise model	Δ	α_{blend}	N_{ens}	τ^*
1	Lorenz63	x	Inverse Noise	$\mathcal{N}(0, 0.5^2)$	0.1	1.0	64	0.1
2	Lorenz63	x	PDA	$\mathcal{N}(0, 0.5^2)$	0.1	1.0	64	0.1

*integration step size $h = 10^{-2}$ [118].

---

[3]data assimilation is not the focus of this thesis. The numerical Lorenz63 forecast results have been kindly provided by Ed Wheatcroft and Hailiang Du at the Centre for the Analysis of Time Series, London School of Economics. The novel aspects contributed by this thesis research is the analysis of these results, not the construction of the forecasts and DA themselves.
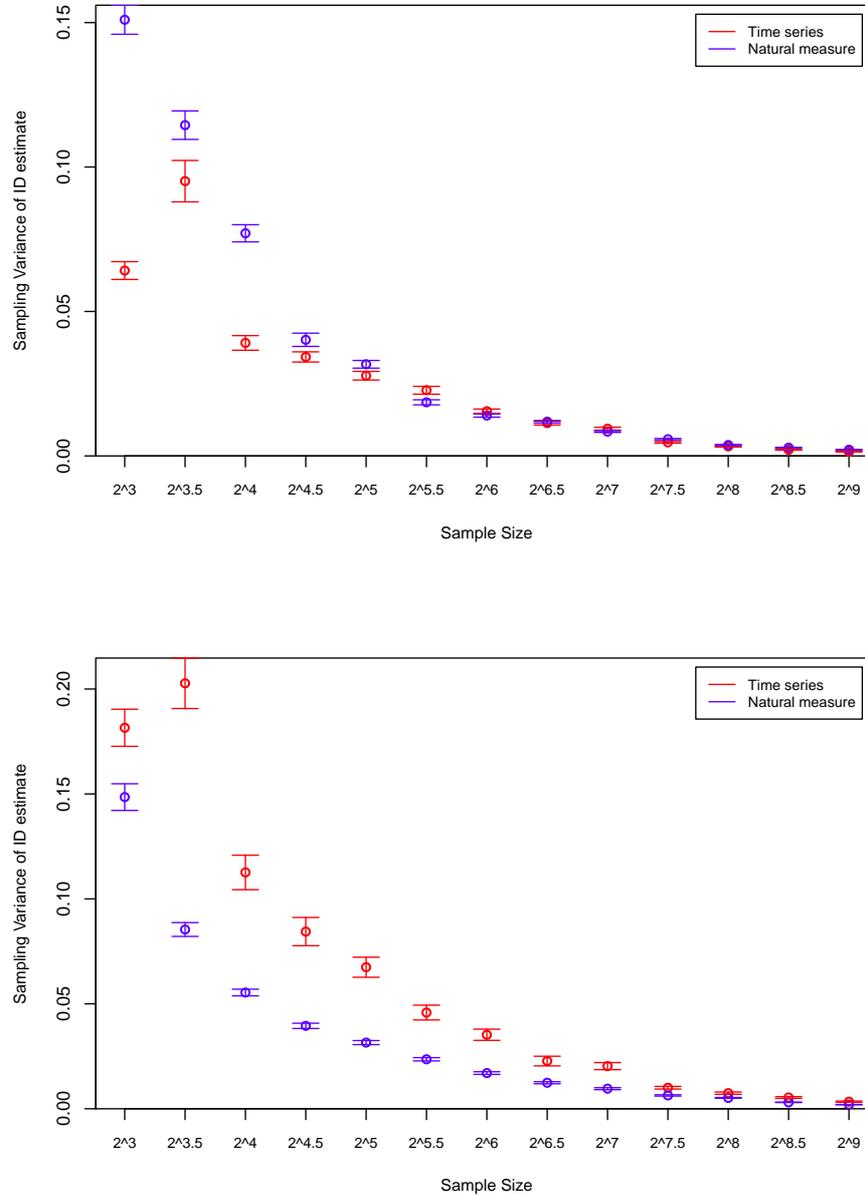
Figure 4.16: Sampling variance with IN and PDA: Sampling variances of ID estimates computed from IN (top) and PDA (bottom) forecasts ($\Delta = 0.1$, $\tau = 0.1$) of a) a correlated time series of Lorenz63 observations (red circles) and b) the natural measure of ID estimates (blue circles) with increasing sample size, both with $5\% - 95\%$ uncertainty intervals.

4.5.2 Numerical results

Figure 4.16 shows the evaluation statistics of the PDA (Expt. 2) forecasts sampled sequentially with those sampled from the natural measure using bootstrap resampling of the Lorenz63 attractor at lead time $\tau = 0.1$. While there is no effect of serial dependence in the observation time series ($r_1(y) \approx 0.86$) on the sampling variance of the information deficit (ID) for the IN forecasts ($r_1(ID) \approx 0.38$), it is clearly evident for the PDA forecasts. Lag-1 autocorrelation values computed from a sample size $N = 2^9$ are $r_1(ID_{IN}) \approx 0$ and $r_1(ID_{PDA}) \approx 0.38$, respectively. This difference in the sampling variance results is attributable to the differences in skill between the two sets of forecasts which is of the order of 0.03 bits for $N = 2^9$. Crucially, the time until convergence of the information deficit is longer for the PDA forecasts given the inflation of its sampling variance. This difference between the sampling variances of ID of the IN and PDA forecasts reflects an inversely proportional relationship between forecast skill and rate of convergence.

Given that predictability of the flow evolution of the Lorenz63 system varies depending on the location in state space, both the degree of serial dependence in observations, and forecast skill can vary over state space. Figure 4.17 shows the profile of the information deficit statistics of the PDA forecasts in 3-dimensional state space. All of the information deficit statistics are coloured coded depending on within which range of values they lie. For example, many of the forecasts constructed at locations with high unpredictability on the attractor (e.g. where the left and right lobes meet) tend to have larger information deficits, indicated by the red symbols. Many of the poorest forecasts lie around this location.

4.6 Approximate ESS corrections

As explained in section 4.1, expressions for effective sample size (ESS) corrections can be derived if the sampling variance of a scoring rule can expressed in
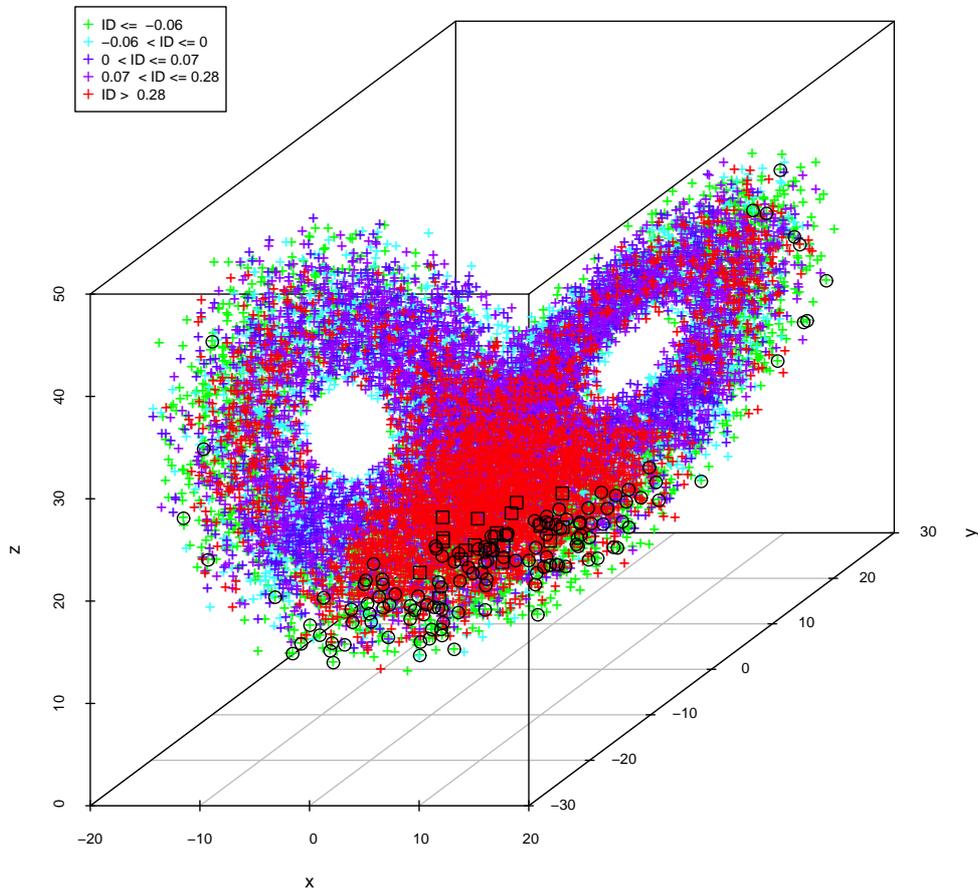
Figure 4.17: Forecast skill with PDA: forecast ID ($\Delta = 0.1$, $\tau = 0.1$) illustrating the degree of predictability of the Lorenz63 system in state space. The double fixed point attractors are clearly represented by the ID samples in state space. Black circles denote very skilful forecasts ($ID < -2$) while black squares denote very poor forecasts ($ID > 4$).

terms of a model's parameters, as in Eqn. (4.11). Wilks [216] used this ESS correction for the Brier score with a "linear-calibration/beta-refinement" (LCBR) probability model. Such expressions are more difficult to derive for the ignorance score (and information deficit) since $Var[IGN]$ does not depend on the moments of the joint distribution of forecasts p and outcomes y. Furthermore, derivations of ESS expressions for the other system-model configurations in

Sections 4.2, 4.3, and 4.4 is beyond the scope of this thesis.

Estimation of the time until convergence (TUC) can provide an approximate alternative method for making ESS corrections. By simply comparing empirical estimates of the sampling variance under serial dependence and serial independence using a resampling method [17], as has been done throughout this chapter, the required sample size correction can be estimated. The ratio of the sample sizes of skill scores computed under serial dependence and serial independence which correspond to the same skill score sampling variance indicates the ratio $\frac{N'}{N}$. The correction is given simply by the difference $N - N'$. For example, consider the sampling variances computed from serially dependent and serially independent (natural measure) information deficit estimates for the PDA forecasts shown in Fig. 4.16. A sampling variance of ≈ 0.04 corresponds to a sample size $N' \approx 2^6$ for the time series and $N \approx 2^{4.5}$ for the natural measure. This indicates a required extra increase in sample size of ≈ 41.

Of course, accurate estimates of the forecast skill are ultimately the aim so the TUC determines which minimum sample size is necessary. Referring again to Fig. 4.16, it is evident that the $5\% - 95\%$ uncertainty intervals for sampling variances of the PDA forecast information deficit estimates do not quite overlap by $N = 2^9$ so a larger sample size is required to be certain of obtaining correct estimates under serial dependence. On the other hand, convergence of the IN forecast information deficit estimates occurs at a sample size of $N = 2^5$ which reflects the lack of serial dependence transmitted to the ID statistics[4]. At the point at which they do, the estimates of ID under serial dependence and serial independence can be considered to converge ensuring that the estimates are accurate.

---

[4]The differences in the sampling variances between the sample size $N = 2^3$ and $N = 2^5$ reflect a small degree of negative lag-1 auto correlation in the ID time series at small sample sizes.

4.7   Forward View and Conclusions

Three case studies of the effects of serial dependence on estimates of forecast skill have been presented in this chapter using various system-model possibilities. In the first case study, the inflationary effect on the sampling variance of scoring rules resulting from the presence of serial correlation in the forecast evaluation statistics using Wilks's [216] probability model for refinement distributions (LCBR model) has been replicated, and also demonstrated using forecasts of the Lorenz63 nonlinear dynamical system. The second case study shows for the first time that linear correlation is not necessarily transmitted to the forecast evaluation statistics where evaluation data are serially correlated. Two stochastic target systems (AR(1), testbed hurricane system) have been employed to show that inflation of the score sampling variance does not occur if the distribution of score statistics is not time-dependent. The third case study has described, for the first time, a forecasting scenario where a deflationary effect on a scoring rule's sampling variance can occur in the presence of serial correlation which is not linear in data generated using the logistic map. Together, the results of these three case studies reveal a previously unreported complexity of forecast evaluation under serial dependence, and highlights how forecasters should exercise caution when making statistical inferences of forecast skill.

To address the limitations imposed by serial dependence on evaluating forecasts of geophysical phenomena, forecasters are advised to make sample size (ESS) corrections dependent on the degree of inflation of the sampling variance of the scoring rule. In addition to serial correlation in observational data, the degree of score sampling variance inflation can be dependent on other forecast properties such as forecast skill, forecast calibration and climatological frequency (of a binary event) as explained in Wilks [216], and also on the given scoring rule. Wilks fits an empirical relationship between the ESS correction and analytical-to-empirical sampling variance ratio with respect to all of the above properties. Since the properties are only defined for the LCBR model, however,

derivation of an empirical fit for ESS corrections for all other system-model configurations in this chapter has not been possible. Furthermore, analytical solutions for the sampling variances of some scoring rules are likely to be difficult to obtain for multi-categorical or continuous outcome scenarios.

The empirical fit approach for ESS corrections used by Wilks [216] is also not tractable where serial correlation in the observational data is not linear but results in linear serial correlation in the score statistics, as shown in Section 4.4. Serial correlation which is not linear cannot be factored into the ESS correction equation which is dependent on lag-1 autocorrelation. Moreover, given that serial correlation which is not linear is undetectable using the autocorrelation function (ACF), a forecaster may not even be aware that their estimates of forecast skill are inaccurate.

An alternative approximate method for determining ESS corrections has been proposed and demonstrated in Section 4.6. This method consists of determining for which sample sizes the sampling variances are equal for the dependent and independent datasets, and take the difference as the correct sample size.

Determination of the *time until convergence* (TUC) of a scoring rule towards its asymptotic "true" value is another approach to assessing the effect of serial dependence on statistical confidence in forecast skill. An examination of the TUC of the information deficit has illustrated how the effect of serial dependence can vary depending on which region of state space is being observed and predicted when evaluating forecasts of a dynamical system. This insight highlights how the relationship between serial dependence in observations and the state of the target system being observed can be of importance when aiming to obtain accurate estimates of forecast skill.

The results presented in this chapter illustrate the effect of serial dependence on forecast evaluation, and provide useful guidelines on how to compensate for the effect and arrive at accurate estimates of forecast skill.

The following are novel contributions in this chapter:

1. derivation of the analytical sampling variance of ignorance score estimates for binary forecasts

2. demonstration of how misleading estimates of forecast skill can result from the presence of serial correlation in evaluation data (with both a stochastic and nonlinear dynamical system)

3. explanation of how the presence of serial correlation in evaluation data is neither a necessary nor sufficient condition for misleading estimates of forecast skill (with stochastic systems)

4. illustration of how misleading estimates of forecast skill can occur where serial correlation is not present in evaluation data but is present in forecast evaluation statistics (with a nonlinear dynamical system)

5. examination of the *time until convergence* of score estimates to their asymptotic "true" value

6. proposal and demonstration of an empirical method for effective sample size (ESS) corrections where serial correlation is present in evaluation statistics

Chapter 5

Techniques and Unresolved Challenges for Hurricane Forecasting

This chapter brings into focus a number of statistical aspects of hurricane forecasting which warrant consideration by forecasters to ensure best practice when constructing and evaluating forecasts.

Section 5.1 gives an overview of the challenges of hurricane forecasting, and briefly outlines the role that forecasting plays in mitigating hurricane impacts. The various hurricane forecast predictands (i.e. forecast variables) relevant to the (re)insurance industry and policy-makers are then introduced in section 5.1.3. Among the categories of predictands considered most important in operational hurricane forecasting is annual hurricane counts [82, 158], which are the subject of the modelling/forecasting studies presented in sections 5.1 to 5.5.

Sections 5.2 and 5.4 introduce two novel yet simple statistical forecast systems designed for constructing year-ahead predictions of annual hurricane counts. The first forecast system is based on a method of constructing probability density functions (PDFs) conditioned on the state of a key environmental index relating to hydro-meteorological conditions which modulate hurricane activity.

This modelling technique is herein referred to as *synoptic conditioning* [32, 46]. The other forecast system is based on an empirical forecasting approach which uses temporal analogues from the historical hurricane time series to construct forecast PDFs, and is referred to as *conditional analogue* forecasting [183, 200]. An innovative *"top-hat" kernel estimation* method is introduced to smooth the forecast PDFs which would otherwise be coarse having been constructed from limited samples of small-count data. This method is a generalisation of Bröcker and Smith [26]. The synoptic conditioning and conditional analogue forecast systems are both trained and evaluated with synthetic data generated from the hurricane model of section 5.2 using conventional scoring rules and via a betting scenario called "Hurricane Roulette", which has been creatively adapted from "Weather Roulette" from Hagedorn and Smith [67]. In both evaluation exercises, the forecast systems demonstrate relative skill over climatological forecasts, indicating their potential usefulness as easily constructible statistical forecast models. Both of the forecast systems are put to the test in chapter 8 where they are used to produce forecasts of the 2013 hurricane season.

An investigation into the limitations of statistical inference with small-count data, relevant for hurricane prediction, is presented in section 5.3. A forecast predictand of particular interest to the (re)insurance and civil planning industries is annual U.S. hurricane landfall counts owing to their direct relation to impacts on livelihoods and property [82, 117]. Much research in those industries has been focussed on statistical modelling of U.S. landfalls because of the inability of dynamical models to simulate hurricane tracks [28, 46, 64, 175]. It is demonstrated in this chapter that robust statistical inferences of annual landfall counts are not realistic, providing a cautionary guide for users of such statistical predictions.

Finally, a unique perspective of forecast quality is presented in section 5.5 by examining the relationship between forecast skill and forecast value, and highlighting important distinctions between the two concepts. Forecast value, or utility, is a complex quantity because it is dependent on the forecast user, and

consequently is not often distinguished from forecast skill. This can lead both to the avoidance of forecast systems which have not demonstrated statistically significant skill and to pressure to use systems with "skill" but no value relative to the task at hand. A novel approach to distinguishing between forecast skill and forecast value is described in the context of Hurricane Roulette, and used to show that a decision-maker is not obligated to wait until proving forecast skill before utilising a forecast system to realise its value. All of the concepts discussed in this chapter form part of the statistical framework of forecast construction and evaluation proposed in this thesis as best-practice guidance for forecasters and decision-makers alike.

The key new contributions included in this chapter are: the proposal of two statistical models based on conditional probability forecasting for producing one year-ahead predictions of hurricane counts, discussion of the limitations of conventional statistical inference methods for small-count variables, and a unique analysis of the statistical relationship between forecast skill and forecast value which highlights that the two concepts need not be considered identical.

5.1 Hurricane forecasting: its limits and the role in reinsurance

Forecast systems have been developed using dynamical or statistical techniques, or a combination of both, to make predictions of North Atlantic basin tropical cyclones on a range of timescales. Depending on lead time, the forecast predictand can range from a single weather system forming over an hourly timescale (see Chapter 6) to annual tropical cyclone counts to the power dissipation index (PDI), an aggregate measure of tropical cyclone activity, on decadal timescales [204]. Tropical cyclone forecasts are typically probabilistic, issued in the form of a PDF, or are point forecasts accompanied by some estimate of uncertainty.

5.1.1 Limitation of models

The skill of tropical cyclone forecasts is currently curbed by various limitations on forecast models. These limitations vary depending on the forecast lead time. On daily timescales, dynamical models are employed to model tropical cyclone formation and tracks, but they have a tendency to overforecast tropical cyclone formation [11], and tracking their movement across the North Atlantic basin is difficult [175], despite improvements in skill in recent years [144]. The NHC uses dynamical models only as objective guidance in their formation predictions (see chapter 6) while statistical-dynamical models have been used to forecast tropical cyclone intensity (i.e. windspeed) with less success [39]. In addition to limitation of model biases, an insufficient amount of time has passed to demonstrate forecast skill since the operational inception of these models [175].

Operational seasonal tropical cyclone forecasts are abundant [57, 65, 143, 147, 198], and although they are mostly produced using statistical methods, dynamical models are able to simulate cyclonic-like disturbances, and have achieved some degree of skill [28]. Many predictors, such as Atlantic and global tropical sea surface temperatures (SSTs), have been identified as important for statistical predictions of seasonal TC activity [208], including the El Niño-Southern Oscillation (ENSO), a mode of climate variability in tropical Pacific SSTs and sea level air pressure, which has emerged as a key predictor in seasonal statistical tropical cyclone models [28, 64]. Unfortunately, accurate predictions of the phase and strength of ENSO are not considered possible before the boreal spring preceding the hurricane season. This limitation is referred to as the "spring predictability barrier" [211] . Aside from physical constraints on TC predictability, accurate seasonal forecasting is further complicated by the lack of reliability of the historical hurricane data archive.

At the longer end of the forecast lead timescale, dynamical models are currently unable to resolve mesoscale weather systems such as tropical cyclones so skilful predictions of hurricane activity out to multi-decadal timescales have not

so far been demonstrated using these models alone [53, 99, 162, 204]. One alternative is to adopt a downscaling approach to simulate TCs with high-resolution regional models by forcing them with boundary conditions taken from global coupled models [204, 99]. Another method is to exploit the statistical relationships between local and remote large-scale climate processes that influence tropical cyclone development with the tropical cyclone activity itself [205, 208]. Otherwise, univariate modelling using a timeseries of hurricane activity is a parsimonious approach although it is limited, like seasonal statistical predictions, by the size of the reliable historical data archive.

There is disagreement in the literature [49, 100, 149, 181, 204] which of the dynamical, statistical, or statistical-dynamical methods has provided the most accurate approach to modelling TC activity on seasonal to multi-decadal timescales due to the lack of out-of-sample evaluations. Camargo et al [28], however, state that the seasonal forecast skill of the best performing operational dynamical models is comparable to that of their statistical model counterparts. On the other hand, Vecchi et al [204] point out a limitation in statistical models by highlighting the very different projections of the Atlantic cyclone power dissipation index (PDI) when regressed on absolute SSTs and relative SSTs separately. They conclude that additional empirical research will unlikely yield a unique, statistically significant hypothesis of the SST-Tropical cyclone relationship, and that that non-statistical theories and models should be used in conjunction with statistical techniques to ensure that there is a physical basis for the modelling of tropical cyclones using environmental covariates. Of course, statistical association does not imply physical causality, which should be taken into account in any purely statistical analysis of hurricane activity.

5.1.2 Role of forecasting in (re)insurance

Although North Atlantic basin hurricanes are typically not the largest or most intense storms globally, they sometimes make landfall in heavily populated re-
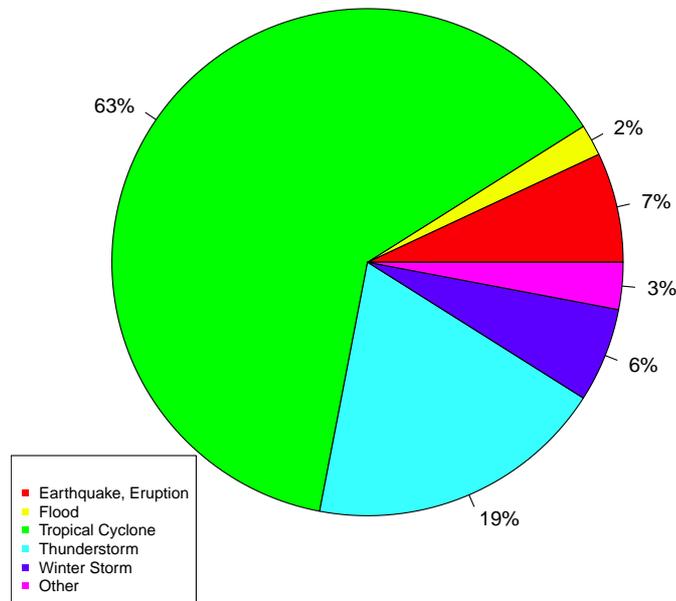
Figure 5.1: Distribution of insured losses caused by U.S. natural catastrophes 1950-2011: the distribution of insured losses (normalised to 2011 dollars for inflation) incurred by the insurance industry due to U.S. natural catastrophes during 1950-2011. Tropical cyclones have caused 63% of the total insured losses. Data source: TOPICS GEO 2011, Munich Re

gions of the Caribbean Sea, Mexico and the eastern seaboard of North America where they can inflict a huge amount of devastation. The impact on the insurance industry has been significant; over half of the insured U.S. natural hazard losses paid out for by the industry between 1950 and 2011 have been due to tropical cyclone damage ($\sim US\$230$billion[1] - see Fig. 5.1) [43, 167]. Not only do hurricanes bring powerful winds to coastal and inland areas (reaching speeds of up to 200mph), but also flooding rain and storm surges which are huge waves created by the offshore winds. There has been an increasing trend in hurricane

---

[1]normalised to 2011 dollars for inflation

losses since the middle of the 20th century, and in recent years the insurance industry has been heavily impacted by hurricanes such as Sandy, Katrina, and Andrew (nearly US\$400 million in insured losses went unpaid after the occurrence of the latter [43, 104]). Clearly, skilful forecasts of hurricane activity on all timescales would hold potential value for hurricane risk management (e.g. pricing annual insurance premiums) [116, 117, 156].

Arguably, the insurance sector does not currently hold the view that longer term (i.e. seasonal to multi-decadal timescales) hurricane predictions are more reliable than its existing predictive methods which consist of climatological baseline forecasts [116]. There is a perceived lack of practised forecast evaluation, and a single poor performance by a forecasting system is enough to negatively affect the overall perspective of those in the insurance sector. Hence, long-term predictions are not widely believed to be skilful and are generally not utilised. Such scepticism is, to some extent, justified since there is a limited historical data archive archive available for operational forecast evaluation, and proving statistically significant skill (before proving the value) of long-term hurricane forecasting is not currently possible. Communication of forecast skill and its translation into forecast value also poses a challenge when the relationship is nonlinear. For example, if there is an exponentially increasing relationship between historical hurricane losses and hurricane intensity then a forecasting system which predicts both CAT1 and CAT5 storms equally well may be of significantly higher value in the latter case owing to a greater potential for prevention of loss, at least to property [112, 157]. Nevertheless, the potential value to the various sectors affected by storms may be substantial if sufficient forecast skill is demonstrated on seasonal to decadal timescales.

5.1.3 Hurricane forecasting scenarios

There are a number of operational forecast products available for a range of hurricane-related predictands on different spatial and temporal scales. Several

of these predictands are of particular interest to the (re)insurance sector, policy-makers, and those of the public who are vulnerable to hurricane risk [34], and are the subject of the statistical analyses and forecast construction, evaluation, and recalibration in this thesis. These specific forecast predictands are summarised below.

Hurricane formation, track and intensity

Operational forecasts of the activity of individual hurricanes on daily timescales[2] and local spatial scales are regularly issued by several forecasting centres around the globe such as the National Hurricane Center and Joint Typhoon Warning Center. Forecast products for several predictands including tropical cyclone formation, best-track, and wind speeds are available from these centres, and have shown improvements in skill out to longer lead times over the past 20 years [144, 166]. These forecasts are vitally important for the planning and decision-making of the (re)insurance industry and government because they provide predictive information on the location and severity of hurricanes that make landfall. The information is usually fed into catastrophe (CAT) models to assess short-term risk and make loss projections (Trevor Maynard, pers. comm., January 2010).

Annual Atlantic basin hurricanes

North Atlantic basin hurricanes are represented by one of the longest available historical data archives of any extreme geophysical phenomenon with earliest windspeed records dating back to 1851. As a result, basin hurricanes are commonly analysed, modelled and forecast to make predictions of hurricane activity on anywhere from seasonal to multi-decadal timescales. Seasonal forecasts are potentially important for users in (re)insurance because the timescale corresponds to the cycle of annual insurance premium renewals. Forecasts of annual

---

[2]out to 120 hours forecast lead time

counts with longer lead times (i.e. years to decades ahead) would also be of benefit to those in risk management and (re)insurance for long-term planning applications such as capital and asset management [116]. Seasonal forecast information has only been minimally utilised in (re)insurance as guidance to date, however, because of lack of proven skill (see section 5.1.2). Annual Atlantic basin hurricane counts are predominantly studied in this thesis because of their relevance for (re)insurance, and the relative robustness and availability of annual count data at least since the 1960's.

Annual U.S. landfalls

North Atlantic basin hurricanes that strike the coast of the United States, Central America, and Atlantic basin island nations during their lifetime have the potential to wreak some of the worst devastation to life and property [13, 117, 157]. A hurricane is classified as a *landfall* when all or part of the hurricane eye wall crosses the coastline [43]. U.S. landfalls are of concern to the insurance community because these inflict the worst damages where populations are dense and economic value is relatively high. The most intense of these (CAT3-5) inflict 85% of the total damage in the continental U.S. ($\sim US$\$150billion[3]) yet comprise only 24% of the total U.S. landfall counts [157]. Predictions of the number of U.S. landfalls and their intensities on a range of timescales strongly influence (re)insurance property premiums [82, 117] so clearly skilful forecasts have potential value to that industry [34, 45].

Environmental indices

There are a number of environmental atmospheric and oceanic factors that have been used to as predictors of hurricane activity owing to their role in modulating hurricane variability. Those factors relating to SSTs in the tropical Atlantic are of fundamental importance since they are a measure of the avail-

---

[3]normalised to 2005 dollars for inflation

able heat energy required for a tropical cyclone to form and develop [51, 178]. PDI and ACE are modelled on Atlantic and global tropical mean SSTs [206] because they are significantly correlated with inter-annual SST variability. The El Niño-Southern Oscillation (ENSO), a dominant mode of atmospheric-oceanic oscillation of Pacific Ocean sea level pressure and SSTs, is considered important because of its relation for global tropical SST anomaly patterns and vertical windshear over the Atlantic which both influence TC activity [15, 64, 28]. El Niño (warm) episodes are associated with passive hurricane seasons and La Niña (cold) episodes are associated with active hurricane seasons. Accurately predicting the phase of ENSO before the beginning of the hurricane season is thought to be important for achieving skilful seasonal forecasts of the Atlantic basin TCs, but is difficult before the boreal spring (see Webster and Hoyos [211]). Other modes such as North Atlantic Oscillation, Atlantic Multidecadal Oscillation (AMO), and Quasi-biennial Oscillation are also used in hurricane modelling but since they are indices often of very similar geophysical variables it is difficult to deconvolve their relative importance [207].

5.1.4 Challenges to hurricane forecasting

While there have been improvements in the techniques and skill of statistical hurricane forecasting [28, 65, 149], there remain a number of unresolved challenges for achieving skilful forecasts of annual hurricane counts and for robust evaluation of these forecasts. Each of these challenges is illustrated and guidelines are provided throughout the remainder of this chapter so that forecasters and forecast users alike can capitalise or, at least, avoid any adverse consequences. The challenges are summarised as follows:

- **small-count data:** when modelling a variable which is represented by small-count data, standard Gaussian-based statistical techniques need to be modified to reliably fit model parameters and perform robust statistical inference [2]

- **forecast uncertainty:** forecasts of variables on specific temporal and spatial scales of material value for users may not be skilful; those which are skilful may not have direct relevance

- **evaluation:** there exist several limitations with seasonal hurricane data (i.e. short historical archive, serial dependence in data) which limit the significance of any measurement of forecast skill

- **profit vs proof:** making a decision to utilise a forecast can be difficult if a decision-maker believes that it has value but its skill is unproven

5.2 Synoptic conditioning hurricane forecast system

A novel hurricane forecasting methodology, designed to exploit the dependency of TC activity on the ENSO phase, is now presented and evaluated. The methodology, henceforth referred to as *synoptic conditioning*, consists of a bivariate modelling of historical annual Atlantic basin hurricane counts conditioned on the phase of ENSO (see section 5.1.3) during the peak period of the hurricane season (August-October). Conditional probability distributions are constructed from the historical modelling so that probabilistic forecasts of seasonal hurricane counts can be issued depending on the expected August-October (ASO) ENSO phase.

Real observational data have been used to calibrate the forecast system so as to demonstrate its plausibility in real-world forecasting. The hurricane data are sourced from the National Hurricane Center's (NHC) HURDAT database[4] while the ENSO data are sourced from the Climate Prediction Center's (CPC) website[5], both hosted at the National Oceanic and Atmospheric Administration

---

[4]http://www.aoml.noaa.gov/hrd/data_sub/re_anal.html

[5]http://www.cpc.ncep.noaa.gov/products/precip/CWlink/MJO/enso.shtml#current

(NOAA). ENSO is classified here using the Oceanic Niño Index (ONI), a measure of 3 month running mean SST anomalies in the critical Niño-3.4 region[6] based on 30-year climatological records updated every five years to remove a long-term warming trend in SSTs. El Niño (warm) and La Niña (cold) episodes occur when the index rises above or below a threshold of ±0.5, respectively. The performance of the synoptic conditioning (SC) forecast system is assessed in a betting scenario in section 5.2.2.

5.2.1 Testbed ENSO-hurricane system

Models based on Poisson processes are considered the canonical method for annual hurricane counts [46]. Let the hurricane system be defined by a stochastic Poisson process with a mean parameter dependent on the ENSO phase ϕ_t so that the seasonal hurricane count Y_t is a Poisson-distributed random variable expressed as either

$$Y_t \sim \begin{cases} Pois(\lambda_A), & \text{if } \phi_t = A. \\ Pois(\lambda_B), & \text{if } \phi_t = B. \end{cases} \tag{5.1}$$

where the parameters λ_A and λ_B are the 1966-2012 means of hurricane counts during events A (El Niño) and B (non-El Niño) respectively. Hence, the seasonal CAT1-5 hurricane counts are distributed according to one of two probability distributions P_A or P_B dependent on the ENSO phase.

5.2.2 Hurricane roulette

In this section, a conceptual framework for communicating forecast skill called Hurricane Roulette [67] is defined. Hurricane Roulette is a betting scenario where a betting client (the "punter") the is offered odds by a cooperative insurer (the "house") defined by the unconditional climatological PDF p_{clim} of annual CAT1-5 hurricane counts at the start of each hurricane season. The client places her bet by distributing all of her wealth (Kelly betting strategy [95]) on

---

[6]Latitude: 5°N-5°S, Longitude: 120°W-170°W [71]

K possible annual hurricane count outcomes y_k, $k = 1, \ldots, K$ over a number of games (seasons) $t = 1, \ldots, N$. The stake on each outcome is determined by the client's forecast PDF P which is conditioned on the ENSO phase ϕ_t. The odds $o_t = \frac{1}{p_{clim}(Y_t)}$ against the actual outcome Y_t set by the cooperative insurer determines the client's return, or pay-off, on each annual bet. Four important assumptions are made in this betting scenario:

1. the client is motivated to maximise her wealth over N games of Hurricane Roulette;

2. both the decision-maker and cooperative insurer possess perfect forecast models which are calibrated with the same historical annual hurricane count records, and these models are unmodified over N games;

3. the client only has access to *a priori* information about the forthcoming peak season (ASO) ENSO phase at the time of each annual bet;

4. for simplicity, ENSO episodes are classified into two phases: El Niño (less-active hurricane season) and non-El Niño (active hurricane season) so that there are two conditional probability distributions.

Co-operative insurer

Consider two events A and B (representing El Niño and non-El Niño phases, respectively) which are mutually exclusive (i.e. $A \cap B = \emptyset$) and complete (i.e. $A \cup B = S$ where S is the entire sample space). The unconditional climatological probability distribution of seasonal hurricane counts p_{clim} is a convex linear combination of individual probability distributions relating to two mutually exclusive and complete events so that

$$P_{clim} = \alpha P_A + \beta P_B, \qquad (5.2)$$

where α and β are the probabilities that the cooperative insurer assigns to each event A and B occurring, respectively. Considering that only the client has *a*

priori information about the ASO phase, the best information available to the cooperative insurer is the historical climatological frequencies of the El Niño or non-El Niño episodes. The values computed from the Oceanic Niño Index (ONI) data during the period 1966-2012 are $\alpha = 0.34$ and $\beta = 0.66$.

Client

In the games of Hurricane Roulette, the client places her stake according to one of the conditional distributions $p(Y|A)$ and $p(Y|B)$, which represent her beliefs of the probabilities of hurricane number outcome Y occurring given either the occurrence of event A (El Niño) or event B (non-El Niño), respectively. Following Bröcker and Smith [26], conditional distributions $p(Y|A)$ and $p(Y|B)$ can be expressed as

$$p(Y = y|A) = \delta P_A(y) + (1 - \delta)P_{clim}(y), \tag{5.3}$$

and

$$p(Y = y|B) = \epsilon P_B(y) + (1 - \epsilon)P_{clim}(y), \tag{5.4}$$

where δ and ϵ reflect the confidence that the client has in her *a priori* information (i.e. how she weights the probabilities of events A and B occurring). These linear combinations are akin to "linear pooling" as described in Section 2.4.1. Hence, the client's forecast system has been perfectly calibrated so, if she has knowledge of the ENSO phase ϕ_t at the start of a hurricane season (i.e. $\delta = 1$ and $\epsilon = 1$), her probability forecast is perfect, that is $p_t(y) = P(y)$.

5.2.3 Forecast skill

Given that both the forecast models employed by the client and cooperative insurer in this betting scenario are structurally correct, the only limitation on issuing a perfect forecast $p_t(Y = y) = P_A(y)$ at round, or "season", t is incomplete knowledge of the ENSO phase ϕ_t. The client relies on her *a priori* information to determine her forecast PDF each season while the cooperative insurer issues a time independent PDF.

To assess the skill of the client's forecast system, the expected ignorance of her forecasts relative to the cooperative insurer's climatological forecast model after each season over the period 1966-2012 (i.e. $N = 47$) is evaluated. The phase ϕ_t at each season is specified according to the historical ONI record while the hurricane count is drawn at random from a Poisson distribution (see Section 5.2.1). The relative expected ignorance at each round of Hurricane Roulette is given as

$$
E[IGN]_t =
\begin{cases}
-\sum_{j=0}^{M} P_A(y_j) log_2 \left(\dfrac{p(y_j|A)}{p_{clim}(y_j)} \right), & \text{if } \phi_t = A. \\
-\sum_{j=0}^{M} P_B(y_j) log_2 \left(\dfrac{p(y_j|B)}{p_{clim}(y_j)} \right), & \text{if } \phi_t = B.
\end{cases}
\tag{5.5}
$$

where M is the maximum hurricane count, and p_{clim} is randomly selected according to α and β.

There are two possible values of $E[IGN]_t$ for each $\{\delta, \epsilon\}$ pair depending on the phase ϕ_t. Under the assumption that the client has perfect knowledge of the occurrence of the ENSO events A and B, she is able to select precisely from which distribution she should draw her forecast $p_t(Y = y)$. Suppose that the rounds of Hurricane Roulette are played numerous times by independent clients (i.e. in different simulations) with the same forecast system. Figure 5.2 shows the quantiles of expected ignorance of the 2048 clients' probabilistic forecasts over $N = 47$ rounds from 1966-2012 to show the distribution of possible forecast skill. All quantiles lie on either $E[IGN] = -0.23$ where $\phi_t = A$ or $E[IGN] = -0.05$ where $\phi_t = B$ since the client uses two forecasts PDFs and the cooperative insurer uses a single forecast PDF. Hence, the client's forecasts have expected skill in both El Niño and non-El Niño years. In fact, even with much lower confidence in her forecasts (e.g. $\delta = 0.4$, $\epsilon = 0.7$), the client still has expected forecast skill. Clearly, more skill is gained in El Niño years because the cooperative insurer puts less weight on P_A than it does on P_B.

Now consider the empirical skill of the client's forecast system where seasonal hurricane count outcomes y_t are drawn at random. The empirical relative

178
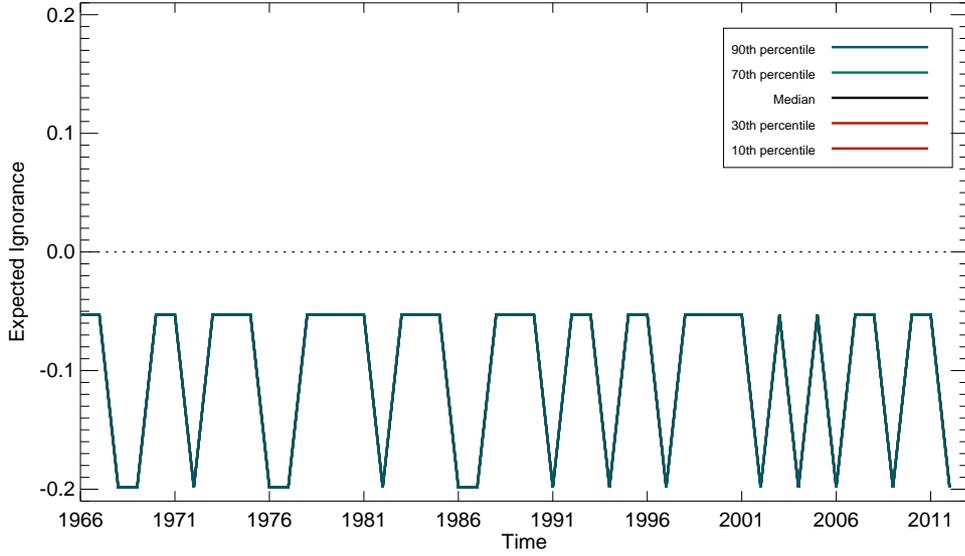
Figure 5.2: Expected skill of SC forecasts: the distribution of the expected relative ignorance of 2048 clients' forecasts when betting against the cooperative insurer's climatological forecasts from 1966-2012 with parameters $\delta = \epsilon = 1$. The client's forecasts consistently have expected skill over the cooperative insurer's forecasts. Note the collapse of all the isopleths onto two different values, indicating either an El Niño or non-El Niño phase, contrasting with Fig. 5.3.

ignorance at each betting round t is given as

$$IGN_t = \begin{cases} -log_2\left(\frac{p_t(y_t|A)}{p_{clim}(y_t)}\right), & \text{if } \phi_t = A. \\ -log_2\left(\frac{p_t(y_t|B)}{p_{clim}(y_t)}\right), & \text{if } \phi_t = B, \end{cases} \qquad (5.6)$$

where y_t is the hurricane count outcome at time t. The relative ignorance of the client's forecasts with parameter values $\delta = \epsilon = 1$ is illustrated in Fig. 5.3. Clearly, there is more uncertainty in the skill of the 2048 clients' forecasts than the expected forecast skill shown in Fig. 5.2, due to sampling uncertainty arising from the stochastic hurricane process. The median of IGN is consistently negative over the whole time series. The client's forecast system exhibits superior forecast skill to the cooperative insurer's climatological forecast system as shown in Figs. 5.2 and 5.3, but the question is: whose forecast system will emerge victorious in the game of Hurricane Roulette? The following section

provides a novel demonstration of how the client's profit and loss exhibits a particular symmetry with information theoretic skill measures such as ignorance, and hence, how forecast skill can be equivalent to forecast value in this case.
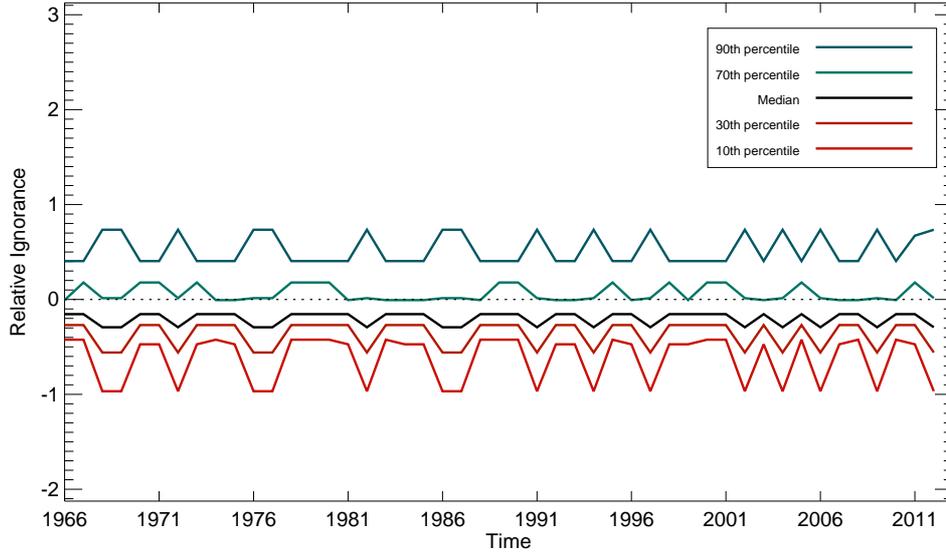
Figure 5.3: Empirical skill of SC forecasts: the distribution of the empirical relative ignorance of 2048 clients' forecasts when betting against cooperative insurer's climatological forecasts from 1966-2012 with parameters $\delta = 1$ and $\epsilon = 1$. The median is constantly negative indicating that the skill of the majority of clients' forecasts is greater than the cooperative insurer.

5.2.4 Results of Hurricane Roulette

In the standard version of Hurricane Roulette, the client places Kelly bets [95] (distributing her wealth proportionate to the probability she assigns to each outcome); this will to maximise the expected growth rate of wealth. There are other possible versions of Hurricane Roulette including different betting scenarios, reflecting various players' attitudes towards risk or profit targets [67].

Given an arbitrary initial investment c_0, the cooperative insurer offers odds $o_t = \frac{1}{p_{clim}(y_t)}$ at the start of each hurricane season so that the capital retained by the client c_t is the product of the odds and the client's stake s_t on the seasonal

hurricane count outcome y_t, that is

$$c_t = o_t(y_t) \times s_t(y_t) \qquad (5.7)$$

$$= \frac{p_t(y_t)}{p_{clim}(y_t)} \times c_{t-1}. \qquad (5.8)$$

Therefore, the client's capital after each round t is simply her initial capital
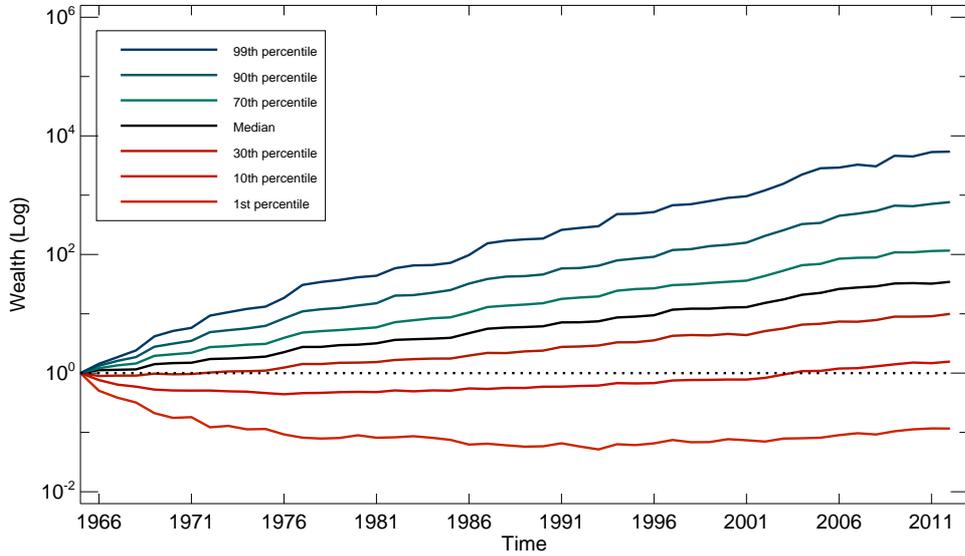
Figure 5.4: Client's accumulation of wealth: the distribution of clients' profit $c_t - c_0$ in rounds of Hurricane Roulette over the period 1966-2012 computed from 2^{11} simulations. 90% of clients have profited within 50 years by betting on the synoptic conditioning forecast system against the cooperative insurer's climatological forecast.

multiplied by the ratio of her and the cooperative insurer's forecast probabilities at round t. Depending on the ENSO phase ϕ_t, this return ratio is defined by

$$u_t = \begin{cases} \frac{p(y_t|A)}{p_{clim}(y_t)}, & \text{if } \phi_t = A. \\ \frac{p(y_t|B)}{p_{clim}(y_t)}, & \text{if } \phi_t = B. \end{cases} \qquad (5.9)$$

The client's capital c_t is thus governed by the return ratio u_t and the capital
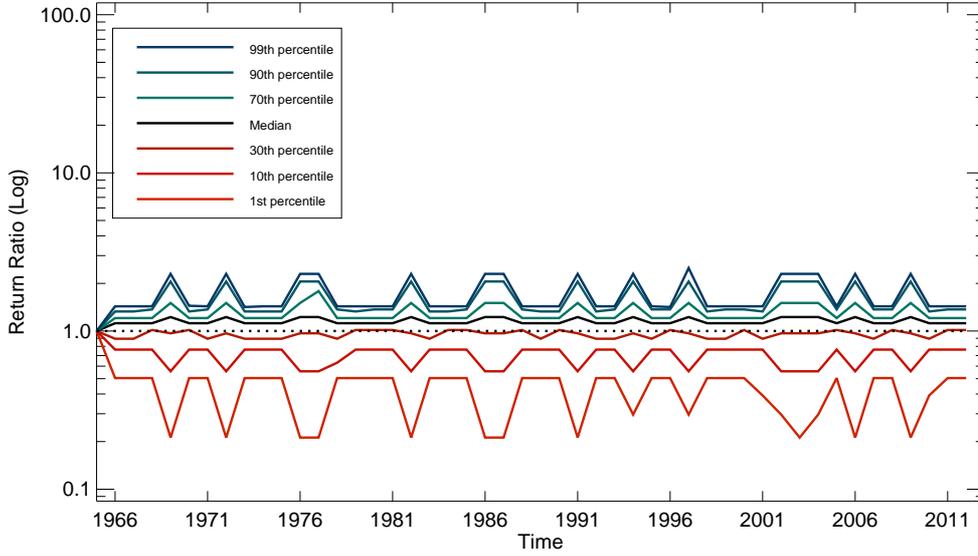
Figure 5.5: Client's wealth: The distribution of clients' return ratios u in rounds of Hurricane Roulette over the period 1966-2012 computed from 2^{11} simulations. The median lies above the $u = 1$ line indicating that most clients have profited by betting on the synoptic conditioning forecast system against the cooperative insurer's climatological forecast. Also given the log scale, the average (arithmetic mean) wealth of a punter is well above zero (i.e. the house has also lost). The bumps reflect where forecast PDFs are sharper (i.e. El Niño phases where the Poisson mean parameter λ_A is smaller) resulting in more extreme incidences of forecast skill.

acquired in the previous round, that is

$$c_1 = u_1 \times c_0 \tag{5.10}$$

$$\vdots$$

$$c_t = u_t \times c_{t-1} \tag{5.11}$$

$$= u_t \times u_{t-1} \times c_{t-2} \tag{5.12}$$

$$= u_t \times u_{t-1} \times \ldots \times u_1 \times c_0 \tag{5.13}$$

$$\vdots$$

$$c_N = u_N \times u_{N-1} \times \ldots \times u_1 \times c_0 \tag{5.14}$$

$$= U^N \times c_0, \tag{5.15}$$

where U is the geometric average of the client's return ratio u after N rounds.

The standard version of Hurricane Roulette described above can be defined in terms of the logarithmic ignorance score [67]. The growth rate $G(N)$ of the client's capital over N rounds is the logarithm of the geometric return ratio average U [209], that is

$$G(N) = \frac{1}{N} log U \tag{5.16}$$
$$= \frac{1}{N} log \frac{c_N}{c_0}. \tag{5.17}$$

The correspondence with relative ignorance (e.g. [172]) is evident in Eqn. (5.17). Hence, the growth rate reflects a proper score (inasmuch as ignorance is a proper score, see section 1.6). $G(N)$ can be used to compare the relative performance of two forecast systems. The numerical results of 2048 games (independent clients) of Hurricane Roulette from 1966-2012 are shown in Figs. 5.4 and 5.5.

5.3 Statistical inference with small-count data

Two properties of annual hurricane count data pose a challenge when making statistical inferences of hurricane activity; one is that they are discrete, and the other is that the counts tend to be low (there are approximately 6 Atlantic basin CAT1-5 hurricanes each year on average). Standard Gaussian-based data-analytic methods are not appropriate for small-count data analysis because they are based on asymptotic theory which is only valid for large and evenly distributed samples [2, 3]. There are number of specialised methods, however, which can be employed; for example, "exact" inference is often considered for estimating p-values and confidence intervals that does not require large samples or values of the count variable [180]. In the context of hurricane count data analysis, obtaining robust statistical inference is particularly challenging with sub-categories of Atlantic basin hurricanes such as U.S. landfalls which have even lower annual counts.

Risk Management Solutions (RMS), a risk management consultancy, has taken the approach of constructing a wide variety of statistical models to make medium-term predictions of U.S. landfalls including varying baseline periods and change-point analysis to account for changes in the underlying distribution of U.S. landfall activity [83, 34]. One such method is to infer U.S. landfall activity from predictions of Atlantic basin hurricanes [97], of which there is naturally more available data, providing a stronger signal-to-noise ratio. The signal-to-noise ratio effectively measures the relative strength of the signal and corresponding noise for some quantity, and is given by

$$\frac{A_s^2}{A_n^2}, \tag{5.18}$$

where A_s is the amplitude of the signal (e.g. the rate (mean) of annual U.S. landfall counts) and A_n is the amplitude of the noise (e.g. the variance of annual U.S. landfall counts).

RMS models U.S. landfall rates based on the assumption that they have been constant since 1948 [34, 202]. Inference of U.S. landfall counts using a constant landfall fraction model is potentially limited, however, by the relatively small historical counts of basin-wide hurricanes (the 1966-2012 average is 6.2 CAT1-5 hurricanes).

A simple experiment is now presented to illustrate the challenge for robust statistical inference of U.S. landfall fractions. The standard statistical model for fraction statistics is the binomial distribution [2]. Villarini et al. [208] employ a binomial, or *logistic*, regression to model fractions of basin tropical cyclones making landfall over the U.S. on predictors such as Atlantic SSTs. A binomial model is used here to highlight the limitations imposed by small-count data on statistical inference of a variable such as the U.S. landfall fraction. Let Y denote a random variable representing the annual number of CAT1-5 Atlantic basin hurricanes making landfall over the U.S., while π is the U.S. landfall fraction rate computed for some historical period. Hence, under the binomial assumption, $Y \sim \mathcal{B}(n, \pi)$. Typically, approximate *Wald* confidence intervals are

constructed for Y/m (i.e. the fraction of Y landfalls out of n annual Atlantic basin hurricanes), is defined as

$$\hat{\pi}_k \pm z_{\alpha/2}\sqrt{(\hat{\pi}_k(1 - \hat{\pi}_k))}. \tag{5.19}$$

$z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. This confidence interval suffers from poor probability coverage when sample sizes are small and for parameter values π close to 0 or 1. A rule of thumb for the Wald interval to perform well is that $n\pi \geq 5$ and $n(1 - \pi) \geq 5$.

The exact *Clopper-Pearson* confidence interval, constructed by inverting a two-tailed binomial test, is considered a better alternative [3]. The upper and lower limits of the Clopper-Pearson confidence interval, expressed as a function of the parameters n and π, are given by

$$\sum_{Y=x}^{n} \binom{n}{Y} \pi_0^Y (1 - \pi_0^{n-Y}) = \alpha/2, \tag{5.20}$$

and

$$\sum_{Y=0}^{x} \binom{n}{Y} \pi_0^Y (1 - \pi_0^{n-Y}) = \alpha/2. \tag{5.21}$$

The Clopper-Pearson confidence interval defined by Eqns. (5.20) and (5.21) can be used for statistical inference of U.S. landfall fractions by providing a range of expected landfall counts for each Atlantic basin hurricane count category. Figure 5.6 shows 95% Clopper-Pearson confidence intervals ($\alpha = 0.05$) computed for a range of Atlantic basin hurricane count categories for a given parameter value $\pi = 0.22$ which represents the landfall fraction rate from 1966-2012. The limitation of employing likelihood intervals to infer U.S. landfalls from Atlantic basin hurricanes is evident in Fig. 5.6, particularly for smaller basin count categories. The discreteness of the binomial distribution leads to conservative confidence intervals (and higher probability of type II errors in hypothesis tests) [3, 145]. The computed coverage probability can be much larger than the nominal confidence interval unless n is very large.

The discretisation of U.S. landfall counts leads to conservative exact intervals [3], thereby making precise estimates of landfall activity difficult to achieve
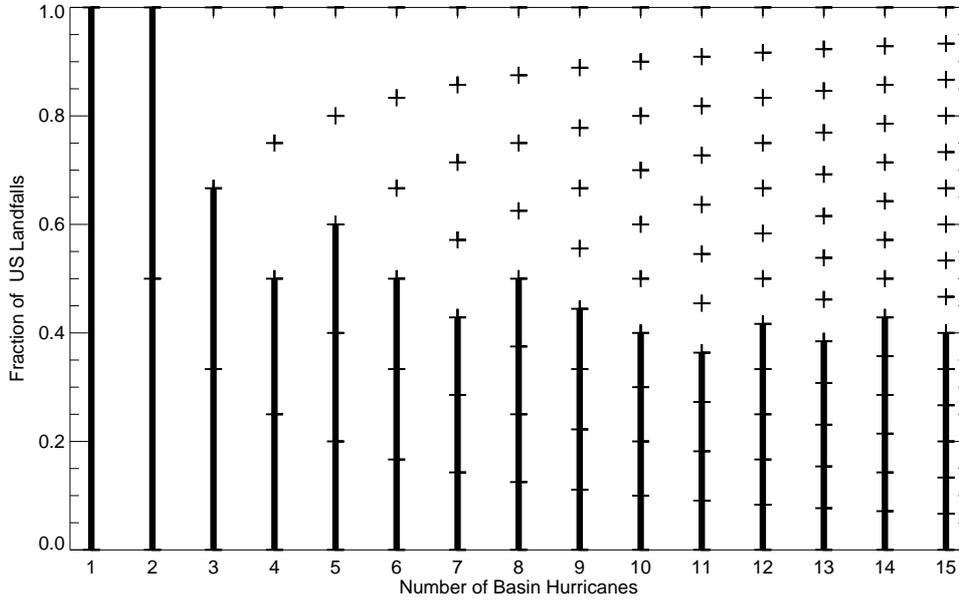
Figure 5.6: Statistical inference of U.S. landfall fractions: 95% Clopper-Pearson confidence intervals with parameter $\pi = 0.22$ estimated from the U.S. landfall fraction rate over the period 1966-2012. '+' symbols denote set of possible fractions for each landfall count category. The lack of precision in the likelihood intervals demonstrates the limitation of statistical inference with small count data.

with exact confidence intervals. In addition, the Clopper-Pearson confidence intervals above are only applicable to a landfall fraction rate assumed constant for all basin count categories. So, what of approximate confidence intervals for U.S. landfall fraction estimates computed individually for each category from the real hurricane data? The score confidence interval, provides probability coverage close to nominal levels, even for smaller sample sizes [3]. The score-test confidence interval is given by

$$\left(\hat{\pi} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{([\hat{\pi}(1 - \hat{\pi}) + z_{\alpha/2}^2/4n]/n)} \right) / (1 + z_{\alpha/2}^2/n). \qquad (5.22)$$

Figure 5.7 contains score confidence intervals constructed from the real hurricane data for each basin count category. Evidently, the score confidence intervals are more precise than the Clopper-Pearson confidence intervals on account

186

of their design to provide probability coverage closer to the nominal level.
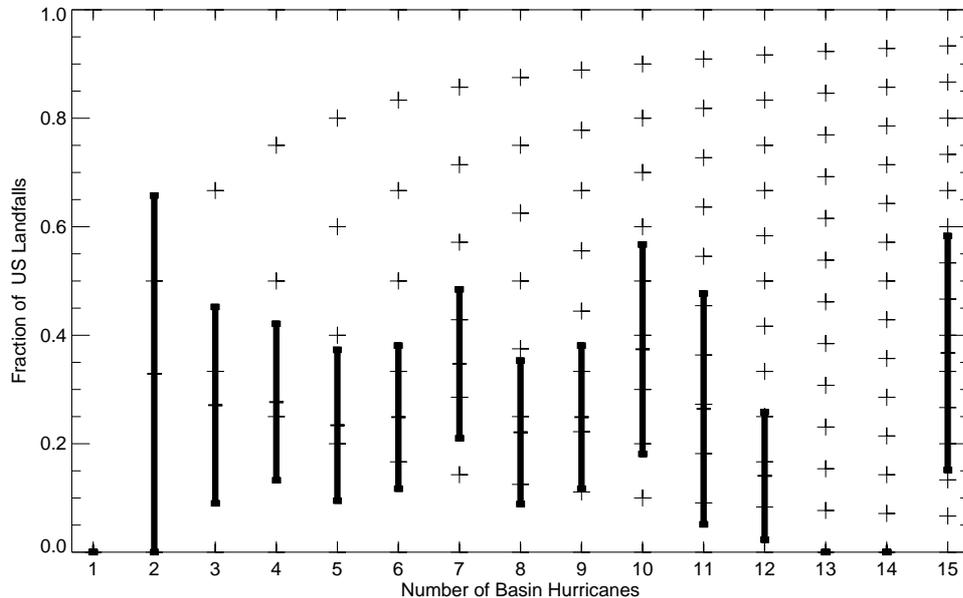
Figure 5.7: Statistical inference of U.S. landfall fractions: 95% score confidence intervals computed from 1966-2012 U.S. landfall fraction rate. '+' symbols correspond to the fractions for each landfall count category.

Several studies [208, 37, 101, 186] have taken the statistical modelling of landfalling hurricanes to smaller spatial scales, and hence smaller U.S. landfall fractions, by modelling the dependency of U.S. landfalls on sub-regions of the U.S. coastline. The limitations on statistical inference imposed by small-count data, however, are even more problematic for smaller subsets of Atlantic basin hurricanes. One storm risk research institute, Risk Prediction Initiative (RPI), hosts a seasonal hurricane forecast competition where entrants are invited to submit probabilistic forecasts of seasonal U.S. landfalls (of various intensity categories) at 6 coastal regions of the U.S. The competition was opened as an initiative to standardise and compare the skill of forecasts issued by commercial, academic, and amateur forecasts alike. Forecasters are asked to assign predictive probabilities to each U.S. landfall outcome between 0 and 5, occurring in each sub-region. Such predictions are affected by the limitations on statistical

inference with small-count data discussed above, and, as a result, comparisons of forecast skill are neither realistic nor informative.

5.4 Empirical conditional analogue hurricane forecast system

A novel yet simple and computationally inexpensive approach to producing seasonal hurricane forecasts based on an analogue forecasting technique [195, 200, 183] is presented in this section. Analogue forecasting (AF) has emerged from the belief that weather patterns are self-repeating, and so, present initial conditions, if observed to be similar to those in the past, are likely to evolve in the same way. Studies of interannual and interdecadal variability of Atlantic basin hurricane behaviour suggest that there may be periodic and self-repeating patterns of hurricane activity [61, 31]. Analogue forecasting has a long history in weather prediction owing to its directly empirical nature and straightforward application [195]. The continued improvement of dynamical models, and requirement of sufficiently large datasets to achieve skilful analogue forecasts, however, have confined its use to longer time scales [195, 200]. Nevertheless, the analogue method has been shown to demonstrate some degree of skill by limiting the geographical region for which forecasts are produced so that good previous matches are more likely [200]. The same reasoning is followed in the development of the Atlantic basin hurricane forecast methodology described here given the relatively local scale of the main development region (MDR) [62] where most hurricanes form.

In forecasting scenarios where the predictand is a continuous variable, the selection of candidate analogues is made by, for example, measuring the correlation [10, 12] between the current state s_n and a previous state s_m, or minimising their Euclidean distance [195, 200], that is

$$d(s_n, s_m) = \arg\min_{s \in \mathbb{R}} ||s_n - s_m||. \tag{5.23}$$

In the case of forecasting discrete predictands such as annual hurricane counts, Euclidean distance metrics may not be necessary since it is entirely reasonable to make $s_m = s_n$ the selection criterion to find the best candidate analogues.

Consider an historical time series of Atlantic basin hurricane outcomes y_t, $t = 1, \ldots, N$ which is to be utilised to produce a forecast of the outcome y_{N+1}. In a simple version of 1-year lead time AF, analogues y^A of the outcome y_N are located in the remaining subset so that the collection of analogue indices (years) $t = i$ is given by

$$I^A := \{i; y^A = y_N\}. \tag{5.24}$$

Given a 1-year lead time forecast, hurricane outcomes occurring in the subsequent years to those of the selected analogue outcomes, referred to as *images* [12] of the analogues, are then collected from the remaining subset. So, if the images are denoted y^{AI}_{i+1} then I^{AI} is the collection of years $t = i + 1$ in the time series for which all images of the selected analogues belong so that

$$I^{AI} := \{i + 1; y^A = y_N\}. \tag{5.25}$$

Finally, histograms of the set of images y^{AI}_i with indices I^{AI} might then be used to produce a point forecast by, for example, computing the mean or mode from the histogram, or to produce a probabilistic forecast by translating the histogram into a forecast PDF with, for example, a kernel dressing method. The latter probabilistic method is referred to as *Conditional Analogue* (CA) *forecasting* in this thesis, although it is based on the "Random Analogue Prediction" method of Smith [183]. CA forecasting can be deployed for a single observation analogue, but also for an ordered sequence of consecutive observations, called a *series analogue*. In addition, the analogue selection criteria time window can be extended to beyond just one year preceding the year $t = N + 1$, as described in the simple example above, so as to sample more information from the dataset[7]. Let the analogue selection criteria for a forecast for the year $t = N + 1$ be a d element *base* vector of observations preceding the year t,

---

[7]it is perhaps intuitive that sampling more information from the historical dataset by

defined by $\mathbf{y}_{t-1} = \{y_{t-1}, \ldots, y_{t-d}\}$ where d is the analogue window length. The two analogue methods are now formally defined.

Single analogue method

To sample forecast information conditioned on the d hurricane outcomes preceding the year $t = N + 1$ using the single analogue method, the following steps are taken:

1. the base vector for the year $t = N + 1$ according to window length d is defined as

$$\mathbf{y}_{t-1} = \{y_{t-1}, \ldots, y_{t-d}\} \tag{5.26}$$

2. analogues y_i^A are located in the dataset according to the base vector $\mathbf{y}_t$ conditioned on each window length $1, \ldots, d$ to obtain the sets of indices

$$I_1^A := \{i; y_i^A = \mathbf{y}_{t-1,1}\} \tag{5.27}$$

$$\vdots$$

$$I_d^A := \{i; y_i^A = \mathbf{y}_{t-1,d}\} \tag{5.28}$$

3. images $\mathbf{y}_i^{AI}$ of the analogues y^A are collected to build the sets of indices

$$I_1^{AI} := \{i + 1; y_i^A = \mathbf{y}_{t-1,1}\} \tag{5.29}$$

$$\vdots$$

$$I_d^{AI} := \{i + d; y_i^A = \mathbf{y}_{t-1,d}\} \tag{5.30}$$

4. the sets of indices $I_1^{AI}, \ldots, I_d^{AI}$ are combined into one climatological distribution of hurricane outcome images for all selected analogues which represents the raw conditional forecast information.

---

extending the analogue time window would lead to better calibrated forecasts but, of course, this may be dependent on the memory (i.e. serial dependence of observations) of the hurricane system.

Series analogue method

The procedure for locating series analogues is similar to that of the single analogue but requires that the analogue consists of the entire ordered series corresponding to the base vector (i.e. d hurricane outcomes); for example, if $d = 3$ then the series $\mathbf{y}_{t-1} = \{y_{t-1}, y_{t-2}, y_{t-3}\}$ must occur in the same order elsewhere in the dataset to be considered an analogue. This method is based on the *delay reconstruction* of chaotic dynamical systems [92]. The following steps are taken:

1. the base vector for the year $t = N + 1$ according to window length d is defined as

$$\mathbf{y}_{t-1,d} = \{y_{t-1}, \ldots, y_{t-d}\} \tag{5.31}$$

2. series analogues $\mathbf{y}_i^A$ are located in the dataset according to the base vector $\mathbf{y}_t$ conditioned on each window length $1, \ldots, d$ to obtain the sets of indices

$$I_1^A := \{i; \mathbf{y}_i^A = \mathbf{y}_{t-1,1}\} \tag{5.32}$$

$$\vdots$$

$$I_d^A := \{i; \mathbf{y}_i^A = \mathbf{y}_{t-1,\{1,\ldots,d\}}\} \tag{5.33}$$

3. images $\mathbf{y}_i^{AI}$ of the series analogues $\mathbf{y}_i^A$ are collected to build the sets of indices

$$I_1^{AI} := \{i + 1; \mathbf{y}_i^A = \mathbf{y}_{t-1,1}\} \tag{5.34}$$

$$\vdots$$

$$I_d^{AI} := \{i + 1; \mathbf{y}_i^A = \mathbf{y}_{t-1,\{1,\ldots,d\}}\} \tag{5.35}$$

4. the sets of indices $I_1^{AI}, \ldots, I_d^{AI}$ are combined into one climatological distribution of hurricane outcome images for all selected series analogues which represents the raw conditional forecast information.

To clearly demonstrate and compare the single analogue and series analogue methods, consider the following simple example: to construct a year ahead

prediction of the 1990 season hurricane count from the 1950-1989 historical dataset using both analogue methods and choosing a window length $d = 3$. The base vector is $\mathbf{y}_{1989,3} = \{y_{1989}, y_{1988}, y_{1987}\} = \{7, 5, 3\}$. For the first year (i.e. 1989) of the base vector, the single analogues (and series analogues, since $y_i^A = \mathbf{y}_i^A$) occur in previous years 1958, 1959, 1963, 1966, 1981, and 1985. The images of the analogues in 1959, 1960, 1964, 1967, 1982 and 1986 are then identified and collected, yielding $\mathbf{y}_i^{AI} = \{7, 4, 6, 6, 2, 4\}$. The procedure is repeated for the years 1988 and 1987 according to steps 2 and 3 in the analogue procedures defined above. All of the analogues and their images are listed in table 5.1 for window length $d = 3$.

Table 5.1: Single and series analogue methods for forecast of the 1990 hurricane season

	Analogue Method			
	Single		Series	
Year	Analogues y_i^A	Analogue images $\mathbf{y}_i^{AI}$	Analogues $\mathbf{y}_i^A$	Analogue images $\mathbf{y}_i^{AI}$
1989	7	7, 4, 6, 6, 2, 4	$\{7\}$	7, 4, 6, 6, 2, 4
1988	5	3, 5, 9, 4	$\{7, 5\}$	4
1987	3	4, 4, 6, 4	$\{7, 5, 3\}$	4

The sets of accumulated analogue images for window length $d = 3$ using the single analogue and series analogue methods are

$$\mathbf{y}_i^{AI} = \{7, 4, 6, 6, 2, 4, 3, 5, 9, 4, 4, 4, 6, 4\}, \tag{5.36}$$

and

$$\{\mathbf{y}_i^{AI}\} = \{7, 4, 6, 6, 2, 4, 4, 4\}, \tag{5.37}$$

respectively. Hence, the single analogue method samples more information from the dataset than the series analogue method. In fact, as a rule,

$$\#I^{AI}(single) \geq \#I^{AI}(series), \tag{5.38}$$

since the likelihood of locating an ordered sequence is lower than locating a single value. The series analogue, however, is designed to exploit serial dependence in observations more effectively. At this point, conditional probability forecasts $p(y_{1990}|y_{1989}, y_{1988}, y_{1987})$ could be constructed from histograms of the collected analogue images listed in Eqns. (5.36) and (5.37). Given the small sample of analogues, however, there is little information contained in the histograms, a problematic issue when the available dataset is as small as the reliable HURDAT hurricane archive. To address this limitation, several post-processing techniques can be implemented to optimise the forecast including finding the optimal window length d_{opt}. These techniques are discussed in section 5.4.1.

Note that, for both the single and series analogue methods, duplicate years may be accumulated up to d times if the base vector $\mathbf{y}_t$ is coincidentally repeated elsewhere in the historical dataset. Duplicated hurricane outcomes result in more forecast probability mass being placed on self-repeating observed states of a system. The strength of CA forecasting, like all analogue methods, lies in its utilisation of self-repeating patterns in observations. If a target system exhibits such patterns, sequences of observations may be expected to contain information. This is the case with testbed hurricane system (see section 4.2) which is used to evaluate the skill of the CA forecasting method later in section 5.4.2.

5.4.1 Probabilistic forecast construction with discrete data

Kernel density estimation is a nonparametric method of translating forecast ensembles (see section 1.8) into continuous PDFs [179, 26]. A different approach is required when estimating probability mass on discrete variables. A histogram is a straightforward and commonly used probability mass estimator [179] but lacks smoothness, particularly if the number of data is limited. Aitchison and Aitken [4] adapt the nonparametric approach to discrete multivariate binary data using a "cubical binomial" kernel function. A more simple *top-hat* kernel

density estimation approach has been developed and applied in this thesis. A weight function $w.$, similar to a kernel function, is used to determine the probability mass assigned to a discrete outcome y_k, and its two adjacent outcomes y_{k-1} and y_{k+1}, respectively. The weight function defined on y_k is given by

$$w_0(y_k) = \frac{2\kappa + 1}{3}, \kappa \in [0, 1), \tag{5.39}$$

while, defined on y_{k-1} and y_{k+1}, it is given by

$$w_{-1}(y_k) = \frac{1 - w_0}{2}, \tag{5.40}$$

and

$$w_{+1}(y_k) = \frac{1 - w_0}{2}. \tag{5.41}$$

Like other kernel estimators, the *bandwidth*, or smoothing parameter, κ controls the "spread" of the symmetric weights; 0 indicates a uniform distribution on $\{y_{k-1}, y_k, y_{k+1}\}$, and 1 indicates 100% weight on the central outcome y_k. In addition, the weight function satisfies the condition

$$\sum_{j=-1}^{1} w_j(y_k) = 1. \tag{5.42}$$

The main difference between the top-hat weight (THW) function and standard kernel functions is that the THW function only determines the weights placed on existing histogram probability mass or relative frequencies rather than acting purely as probability mass estimator. Therefore, the probability mass assigned to outcome, y_k, is

$$\hat{p}_\kappa(y_k) = w_{-1}p(y_{k-1}) + w_0p(y_k) + w_{+1}p(y_{k+1}). \tag{5.43}$$

As with other kernel density estimation methods, the choice of the bandwidth κ is crucial to obtain accurate estimates of the true p, it such a thing exists, and hence, to construct skilful forecasts. A sensible approach to optimising κ is to minimise some cost function such as a scoring rule

$$(\hat{\kappa}) := \arg\min_{\kappa} -\frac{1}{N}\sum_{t=1}^{N} S(p(y_i; \kappa)). \tag{5.44}$$

194

The optimisation step has been executed out-of-sample [4] with either a training set of forecast-outcome pairs or K-fold cross-validation (CV) [69] (see 1.8). Two additional steps have been performed when deploying the THW method to address the finite range of the kernel weights and to account for counting statistics:

1. a finite probability of $\frac{1}{N+1}$ is placed on outcomes lying outside the range of the raw histogram i.e. $p(y_k < y_{min} = \frac{1}{N+1})$ and $p(y_k > y_{max} = \frac{1}{N+1})$. $\frac{1}{N+1}$ represents the probability of a missing outcome occurring once less than an outcome occurring just once, having probability $p(y_k) = \frac{1}{N}$;

2. if there are hurricane number bins where $p(y_k) = 0$ after kernel dressing then a probability mass of $\frac{1}{N+1}$ is distributed across all kernel dressed bins and then all bin probabilities are normalised to form the pdf

After kernel dressing with the THW method, all probabilities are normalised to retain a probabilistic forecast PDF [67], that is

$$\sum_{k=1}^{K} \hat{p}_\kappa(y_k) = 1. \tag{5.45}$$

Normalisation consists of re-scaling probabilities $p(y_k)$ with the sum of the pre-normalised total such that

$$p(y_k) = \frac{p(y_k)}{\sum\limits_{k=1}^{K} p(y_k)}. \tag{5.46}$$

The final step in the histogram post-processing procedure is to blend (see section 1.8) the normalised forecast PDF with the unconditional climatological hurricane outcome distribution to ensure that it performs at least as well as a climatological forecast. The blending parameter α controls the weighting between the forecast p and climatology p_{clim} to produce a "final" forecasting probability given by

$$p(y) = \alpha \times p(y) + (1 - \alpha) \times p_{clim}(y), \tag{5.47}$$

where $\alpha \in [0, 1]$.

All of the parameters of the CA forecast system κ, α, and d can be optimised by minimising a scoring rule over a training set of forecast-outcome pairs for the given range of parameter values (i.e. $d = 1, \ldots, D$, $\kappa \in [0, 1)$, $\alpha \in [0, 1]$). An example of a simple iterative algorithm to find precise optimal values of a single parameter (i.e. κ or α) is given in Algorithm 1 where the parameter is denoted by θ.

Algorithm 1 can be modified to optimise both κ and α concurrently by inserting a second loop, and is executed for a range of values of d to find the optimal window length d_{opt}. The three optimised parameters can then be plugged into

$$\hat{p}_{t+1}(y|\kappa_{opt}, \alpha_{opt}, d_{opt}) = \alpha_{opt} \times p_{\kappa_{opt}}(y|d_{opt}) + (1 - \alpha_{opt}) \times p_{clim}(y), \quad (5.48)$$

to produce a year ahead hurricane forecast. The quality of the forecast is dependent on the size of the training set. The CA forecast system is calibrated and evaluated in section 5.4.2 with large datasets of synthetic hurricane data, ensuring a well-calibrated forecast model. In real world hurricane forecasting, the reliable historical record is limited in size, but still may show skill relative to climatology if hurricane activity exhibits repeating patterns. A testbed hurricane system with periodic behaviour is defined in the next section to evaluate the CA forecast system. In addition, a Bayesian forecast model which is designed to exploit such periodic behaviour is also introduced, and used as a benchmark model to evaluate the relative skill of the CA forecast system (see section 5.4.2).

5.4.2 Assessing the skill of the conditional analogue forecast system

Consider the testbed hurricane system which was introduced in section 4.3 to simulate annual CAT1-5 hurricane counts. The number of storms in a given

---

Algorithm 1: Forecast parameter θ optimisation

1: $\epsilon = c$ // set step value ϵ

Ensure: $\theta_{min} \leq \theta_{opt} \leq \theta_{max}$ // set bounds on θ

2: $\theta_L = \theta_{min}$

3: $\theta_U = \theta_{max}$

4: **for** $k = 1$ to L **do**

5:     $\epsilon_i = \frac{\epsilon}{10^{i-1}}$

6:     $\theta_{test} = \{\theta_L, \theta_L + \epsilon_i, \theta_L + 2\epsilon_i, \ldots, \theta_U\}$ // create array of test values

7:     $M = length(\theta_{test})$

8:

9:     **for** $j = 1$ to M **do**

10:       **for** $i = 1$ to N **do**

11:         $IGN_{i,j} = -log_2(p_{\theta_j}(Y_i))$

12:       **end for**

13:       $\widehat{IGN}_j = \frac{1}{N} \sum_{i=1}^{N} IGN_{i,j}$

14:     **end for**

15:

16:     $I = which(\widehat{IGN}_j = min(\widehat{IGN}_j))$

17:     $\theta_{opt} = \theta_I$

18:

19:     **if** $\theta_{opt} - \epsilon_i > \theta_{min}$ **then**

20:       $\theta_L = \theta_{opt} - \epsilon_i$

21:     **else**

22:       $\theta_L = \theta_{min}$

23:     **end if**

24:     **if** $\theta_{opt} + \epsilon_i < \theta_{max}$ **then**

25:       $\theta_U = \theta_{opt} + \epsilon_i$

26:     **else**

27:       $\theta_U = \theta_{max}$

28:     **end if**

29: **end for**

30: **return** θ_{opt}

---

year is drawn randomly from a Poisson distribution with a time-dependent mean parameter, $\lambda(t)$, which follows a sinusoidal cycle over time to simulate observed patterns of hurricane behaviour [31, 61, 98] in the Atlantic basin. Hence, if Y_t is a hurricane number at year t then

$$Y_t \sim Pois(\lambda(t)), \tag{5.49}$$

where

$$\lambda(t) = A sin\left(\frac{2\pi(t + \phi_0)}{T_p}\right) + y_c, \tag{5.50}$$

where A is the amplitude, y_c is the count offset, T_p is the period, and ϕ_0 is the initial phase of the oscillatory system. The unconditional climatological forecast used to measure zero skill is constructed by taking the sum of the system PDFs over all phases ϕ of the sinusoidal oscillation i.e. if $p_\phi(y)$ is the probability according to the system Poisson distribution, then

$$p_{clim}(y_k) = \frac{1}{T_p}\sum_{\phi=1}^{T_p} p_\phi(y_k). \tag{5.51}$$

The climatological forecast is used in the blending stage of producing a forecast PDF with the CA forecast method (see 5.4.1). Figure 5.8 displays an example of a time series of synthetic storm counts generated from the testbed hurricane system.

A Bayesian Hurricane Forecast Model

Given prior knowledge of the underlying periodic behaviour of the testbed hurricane system it is possible to construct a Bayesian model to produce benchmark forecasts. The standard Bayesian approach is to condition a probabilities of future unknown events (*posterior*) on information that is known so that

$$posterior \propto likelihood \times prior. \tag{5.52}$$

Consider that a Bayesian forecaster employs an imperfect hurricane forecast model to make predictions of hurricane counts at year t but knows that the
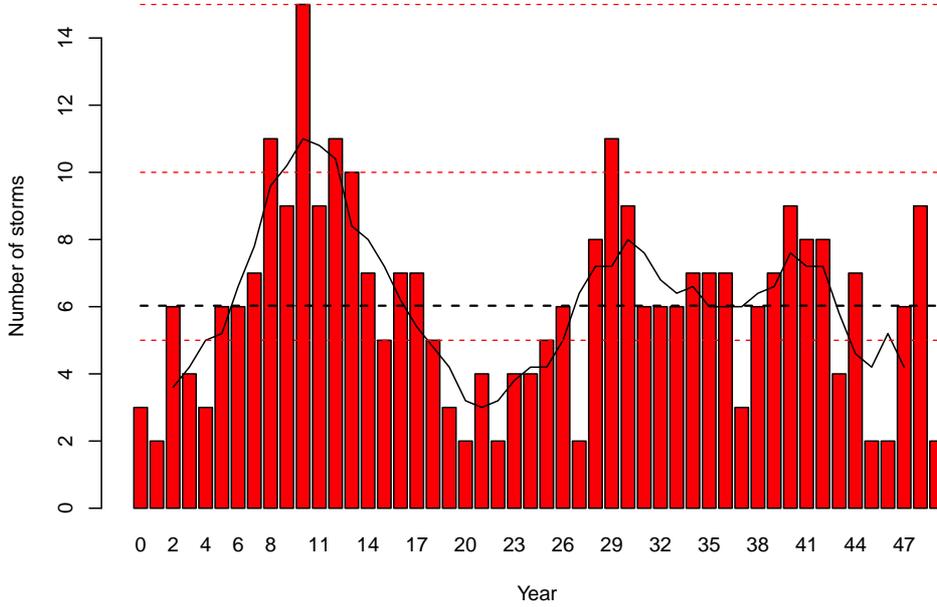
Figure 5.8: Atlantic basin hurricane counts: Example 50 year time series of synthetic CAT1-5 Atlantic basin hurricane counts. The mean (dashed line) corresponds to the real-world dataset average, and the solids line represents the 5-year running mean.

hurricane system has observes a periodic cycle with phases $\phi \in \{1, 2, ..., T_p\}$. Let the imperfect forecast model be defined by a *squared* Gaussian distribution. That is, if V represents a random variable drawn from this distribution, then

$$V \sim \mathcal{N}(\mu, \sigma^2), \tag{5.53}$$

then

$$X = \lfloor V^2 + 0.5 \rfloor, \tag{5.54}$$

where X represents the number of annually forecast hurricanes and $\lfloor \cdot \rfloor$ is the floor function. The model parameters μ and σ have been fitted to each of the T_p phases of the hurricane system's cycle by minimising the relative entropy (see section 2.5) of the forecast PDF p and the true PDF q, that is

$$(\hat{\mu}, \hat{\sigma})_\phi := \arg \min_{\mu, \sigma} -q_\phi(y_k) \sum_{k=0}^{K} log_2 \left(\frac{p_\phi(y_k)}{q_\phi(y_k)} \right), \tag{5.55}$$

199

where $q_\phi(y_k)$ and $p_\phi(y_k)$ are the true and forecast probabilities respectively of the kth outcome occurring at phase ϕ. The Bayesian forecaster selects her prior belief of a hurricane outcome y_k occurring as the unconditional forecast mass $p(y_k)$, and the likelihood function as $p(\phi|y_k)$ so that the posterior probability on hurricane outcomes y_k given phase ϕ is expressed as

$$p(y_k|\phi) = \frac{p(\phi|y_k) \times p(y_k)}{1/T_p}, \tag{5.56}$$

where $1/T_p$ represents the unconditional probability of the hurricane system having phase ϕ at any given time t.
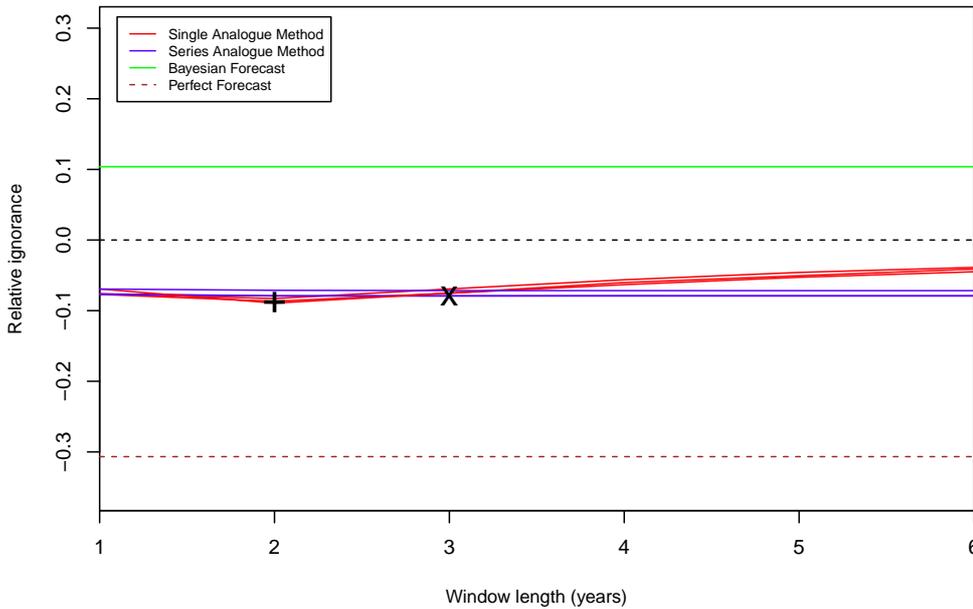
Figure 5.9: Forecast skill of CA forecasts: Ignorance scores computed for three training sets of single (red lines line) and series (blue lines) analogue forecasts at increasing window lengths. The score minima are shown for both the single (plus) and series analogue methods (cross). The single and series analogue methods both demonstrate skill relative to climatology, and better than the Bayesian forecast (green line), but are outperformed by the perfect model forecast (brown line).

CA forecast training and evaluation

To perform a robust calibration of the CA forecasts, three training sets of $N = 2^{12}$ synthetic annual hurricane counts are drawn with the initial phase ϕ_0 selected at random. The other parameter values are set to $A = 2.5$, $T_p = 24$, and $y_c = 6$. The performance of the calibrated CA forecasts is then evaluated with ignorance relative to the climatological forecast p_{clim} with a further evaluation set of size $N = 2^{11}$, and compared with the performance of the Bayesian model and a perfect model. Figure 5.9 shows the results of the evaluation stage. The single and series analogue forecast systems exhibit comparable skill which is superior to both the climatological forecast and the Bayesian forecast model, although they are, not surprisingly, less skilful than the perfect forecast model. The single analogue forecast system has maximum skill at window length of $d = 2$ years, but its skill evidently deteriorates with increase in window length, and is outperformed by the series analogue forecast system at longer window lengths. The series analogue forecast system performs consistently well across window lengths and has maximum skill at $d = 3$ years.

5.5 Forecast skill and forecast value

Establishing statistical confidence in results from data analysis is important for climate scientists aiming to detect temporal climate trends, and equally for forecasters who wish to prove that their forecast system has reliable skill. Proving out-of-sample skill of an annual hurricane forecast system is unrealistic on less than decadal timescales, however, because of the slow rate at which new evaluation information is collected. These timescales are too lengthy for those decision-makers who operate on the same timescales as the forecast lead times [116]. The temporal limitation on proof of skill has led to the belief that hurricane forecast information is of little economic value for decision-makers, and that baseline climatological expectations are a better predictive tool until the

skill of forecast systems can be established [156]. Such a belief stems from a confusion of skill with value, however, often coupled with the use of naïve statistical tests. While, of course, the degree of statistical uncertainty increases with decrease in sample size (time duration) [85, 18], it is shown in this thesis that there is a fundamental difference between the skill of a forecast and its value. Hurricane counts appear to reflect slowly changing hydro-meteorological conditions (e.g. the AMO), and the evaluation of both skill and value is complicated by long timescales. It is shown in this section, however, that these factors do not compel a risk tolerant decision maker to wait decades until skill is "proven". Forecasts may well have statistical skill without adding any value for decision-makers. At the same time, imperfect forecast systems can possess non-trivial value long before one might establish that their skill was statistically significant. The concept of profiting before proving forecast skill is explained and demonstrated in this section.

Relatively little consideration has been given to measuring the economic value of forecasts in particular, however, because of the complex, multi-disciplinary nature of the task [54, 94, 112, 193]. In addition to meteorology, the fields of economics, psychology, statistics, management science, and operations research are all relevant when evaluating forecast value. A detailed investigation of forecast value is beyond the scope of this thesis although the relationship between forecast skill and forecast value is discussed in section 5.5.3.

The evaluation and comparison of forecast skill and value in this section is framed in a betting scenario, referred to as the "Swindled Statistician Scam" [79], which unfolds as follows: a wily underwriter approaches a statistician with a business deal. The statistician will produce a probability forecast of the number of CAT1-5 Atlantic basin hurricanes in the coming year, and the underwriter will use her market contacts to bet on the forecast. As soon as the statistician can prove the forecast really does have skill, the underwriter will pay royalties. Will this leave the statistician swindled out of a small fortune?

Recall the testbed hurricane system based on a stochastic Poisson process

defined in Eqn. (5.50), and the imperfect forecast model of that system defined in Eqns. (5.53) and (5.54) in section 5.4.2 to generate synthetic datasets of storm counts. The system parameter values are set to the same values used in section 5.4.2 (i.e. $A = 2.5$, $T_p = 24$, and $y_c = 6$) along with a random selection of the initial phase ϕ_0.
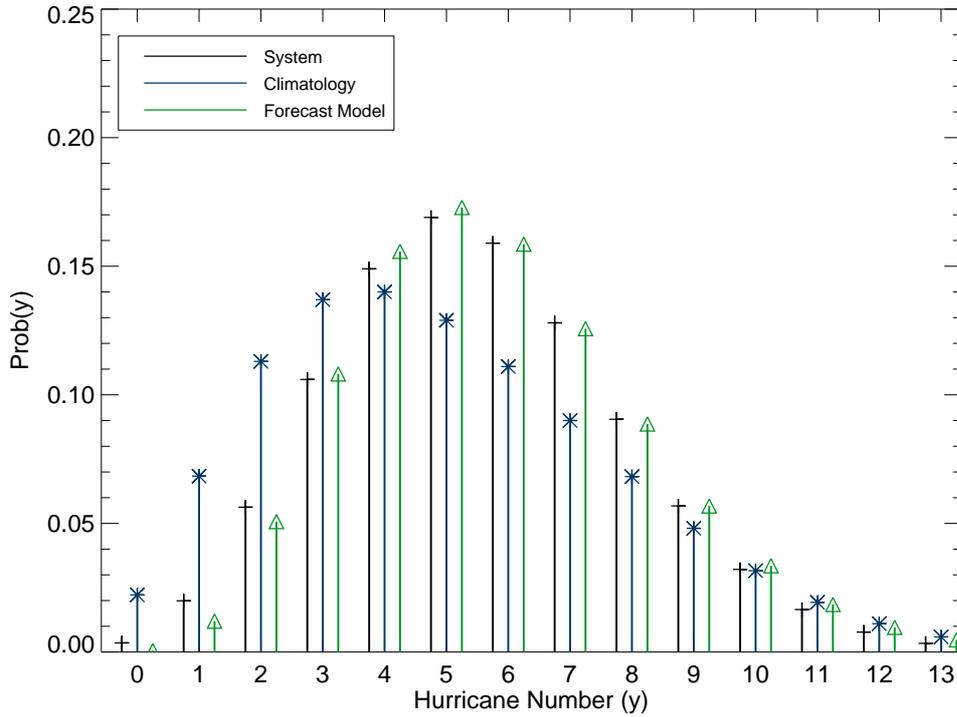
Figure 5.10: System, forecast, and climatology: probability distributions for the system (black), and an imperfect model (green) for phase year $\phi = 12$ of the 24 year cycle. The climatological PDF (computed over all values of ϕ) is also shown in blue. The imperfect model PDF appears is a better fit than the climatological PDF with respect to the difference between the expected ignorance of the two (i.e. $E[IGN_{fcst}] - E[IGN_{clim}] = -0.11$).

The model parameters μ and σ are, as before, fitted to each of the $T_p = 24$ phases of the hurricane system's cycle[8] by minimising relative entropy of the forecast PDF p and the true PDF q. An example, showing q, p, and the

---

[8]There is a variety of proposed values for hurricane cycles in reality [61, 31]. The demonstration here holds for any value of $T_p \gtrsim 8$.

climatological PDF (equally-weighted sum of the 24 system phase PDFs) at phase $\phi = 12$ is illustrated in Figure 5.10. A Monte Carlo approach is adopted to compare the outcomes of "waiting" or "betting". Firstly, the performance of the imperfect hurricane forecast model is assessed, and the duration of time required to attain statistically significant forecast skill from 2^{11} simulations utilising standard hypothesis testing (p-values) is measured. Following that, the value of the forecast model is assessed in N games of Hurricane Roulette, where the imperfect (but time dependent) model probabilities are used to place bets against odds set by the cooperative insurer using the correct (but not time dependent) climatological probability distribution. The results can be reported in either bits of information or as an expected annual return (see Hagedorn and Smith [67]).

5.5.1   Time to forecast skill

Jolliffe ([85]) discusses the importance of including statistical uncertainty in forecast verification through the use of confidence intervals and hypothesis testing. Attempts to quantify uncertainty in forecast skill statistics are hindered, however, when there is only a small amount of available evaluation data. Wilks [215], Jolliffe and Stephenson [86], Bradley et al [18] and Seaman et al [177] all discuss the limitations imposed by small data samples on forecast evaluation which cause large sampling variability, and hence statistical uncertainty in empirical measures of forecast skill. Recall from section 1.6 that the forecast evaluation problem is a distributions-oriented approach [142] where the correspondence between forecasts and outcomes is modelled explicitly by their joint distribution. Higher dimensionality in the joint distribution of forecasts and outcomes (i.e. the range of possible forecast values is large) (see Murphy [134]) results in further increased sampling variability, and hence, increased duration of time to prove forecast skill. Confidence intervals and null hypothesis significance tests (NHST) are commonly used to detect statistical significance

of forecast skill, but are based on the assumption of independence in forecast-outcome pairs (see chapter 4), asymptotic normality of the sample score, and do not perform well for small sample sizes (Agresti, [3]; Jolliffe, [85]). Although other options for small samples exist (e.g. nonparametric bootstrap intervals, Bayesian intervals), the standard hypothesis test is employed here because the likelihood of rejection is (erroneously) higher than other methods (see Nicholls [146] for a critique of hypothesis tests), and can therefore be used to demonstrate the *minimum* duration of time required to prove forecast skill.

The duration of time required for the statistician to skill of his hurricane forecast system is now assessed using relative ignorance (IGN) and the Pearson linear correlation coefficient, denoted r, which is defined as

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}, \tag{5.57}$$

where X denotes either the mean or median probability forecast, and Y is the annual hurricane outcome. A comparison is also made between statistical inference with these two different scoring rules.

Figure 5.11 illustrates the distributions of p-values resulting from hypothesis tests $H_1 : IGN < 0.0$ with increasing sample size (numbers of years) after $N = 2^{11}$ forecast evaluation simulations (or independent statisticians). It is evident that to establish statistically significant forecast skill (at the 95% level; p-value=0.05) would take 64 years for $\sim 91\%$ of independent statisticians in the case of the correlated time-series. The effects of serial dependence on skill score sample statistics (see chapter 4) imply that an even longer duration of time would be required to prove the statistical significance of the forecast skill than indicated by the results of the former case. Bootstraps of forecast-outcome pairs can easily be used to demonstrate longer time durations under serial independence (not shown). The time durations required to establish forecast skill using linear correlations are even longer. Statistical significance is attained by 96%
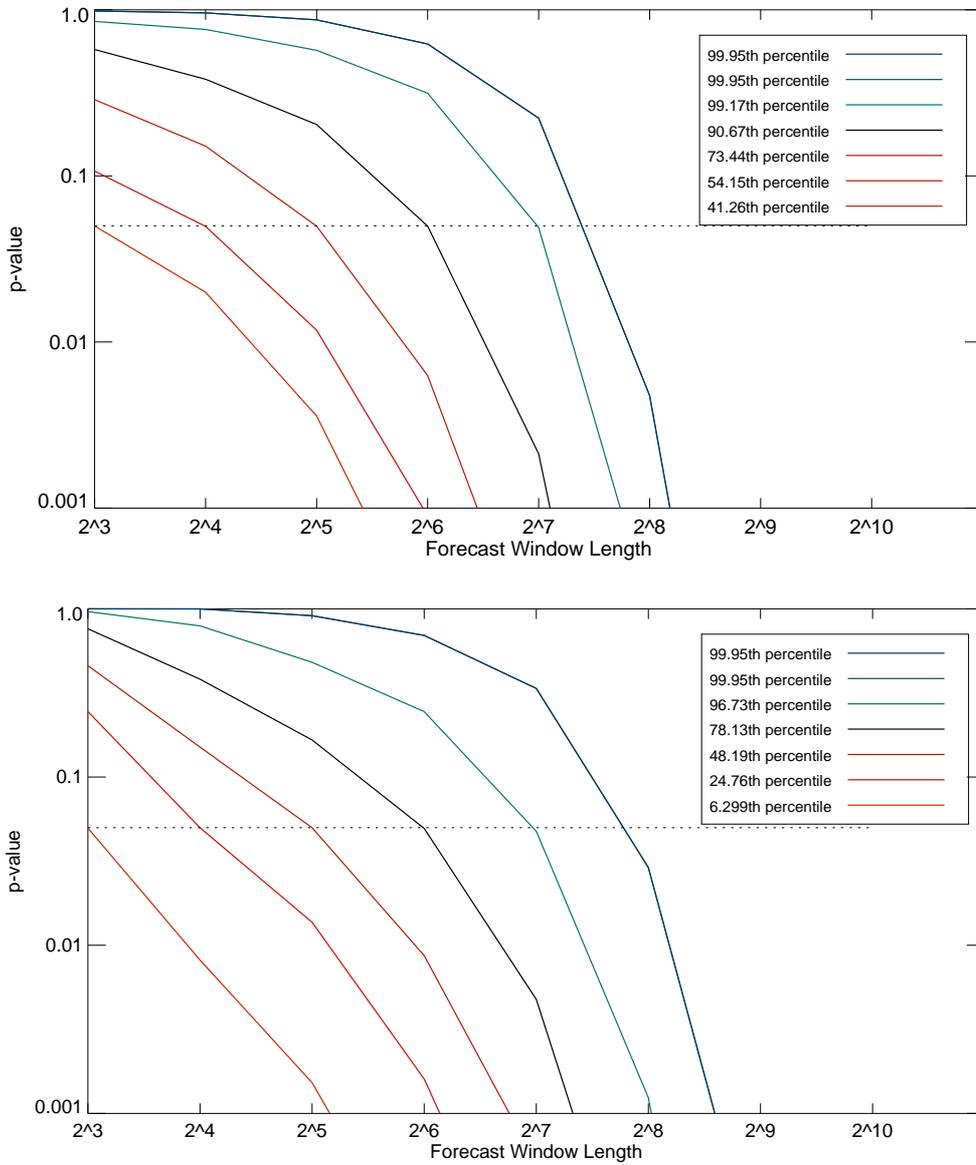
Figure 5.11: Time to forecast skill: Distribution of forecast skill p-values ($H1 : IGN < 0$) of 2^{11} independent statisticians (simulations) as evaluated with IGN (top) and r (bottom). 91% of the statisticians have established statistically significant skill (p-value ≤ 0.05) by 64 years with IGN while 78% have established statistically significant skill using r_{mean}.

of independent statisticians using mean forecasts after 64 years with hypothesis $H_1 : r > 0.3$, and 78% using mean forecasts after 64 years with hypothesis $H_1 : r > 0.4$ ($r = 0.4$ is considered to be a minimum value of skill for Atlantic basin hurricanes, and by Owens and Landsea [149]). Table 5.2 summarises the

206

results are almost identical for the use of median forecasts. It is also noteworthy that the p-values are mostly larger and have a wider spread at the smallest sample sizes (time duration) ($< 2^6$ years) - especially with $H_1 : r > 0.4$, compared to the relative ignorance results. The larger sampling uncertainty highlights the limitation of using linear correlations as a skill measure on these shorter timescales (this is discussed in section 5.5.2).

Table 5.2: Hypothesis tests of forecast skill

Score	$H : 1$	p-value	%age after	
			64 years	128 years
Relative ignorance	$IGN < 0$	0.05	$\sim 91\%$	$\sim 99\%$
Linear corr (forecast mean)	$r_{mean} > 0.3$	0.05	$\sim 96\%$	$\sim 99\%$
Linear corr (forecast mean)	$r_{mean} > 0.4$	0.05	$\sim 78\%$	$\sim 97\%$

5.5.2 Time to forecast value

The time duration required to demonstrate the value of the hurricane forecast system is now investigated in the Hurricane Roulette scenario, and then compared with the time to establish its skill as estimated in the previous section. The concept of "time to value" has been conceived in this thesis. Recall the scenario where the underwriter has agreed to pay royalties to the statistician once he has demonstrated statistical significance of the skill of his hurricane forecasts while she uses them to place bets on the outcomes of each hurricane season in a game of Hurricane Roulette (see section 5.2). Hurricane Roulette is recapitulated here as follows: at the start of each annual hurricane season the underwriter is offered odds defined by the climatology PDF. She then places her bet by distributing all of her current wealth (see Kelly betting strategy [95]) according to the forecast probabilities assigned to K possible annual hurricane

count outcomes y_k, $k = 1, \ldots, K$. The actual hurricane outcome determines the pay-off on each annual bet.
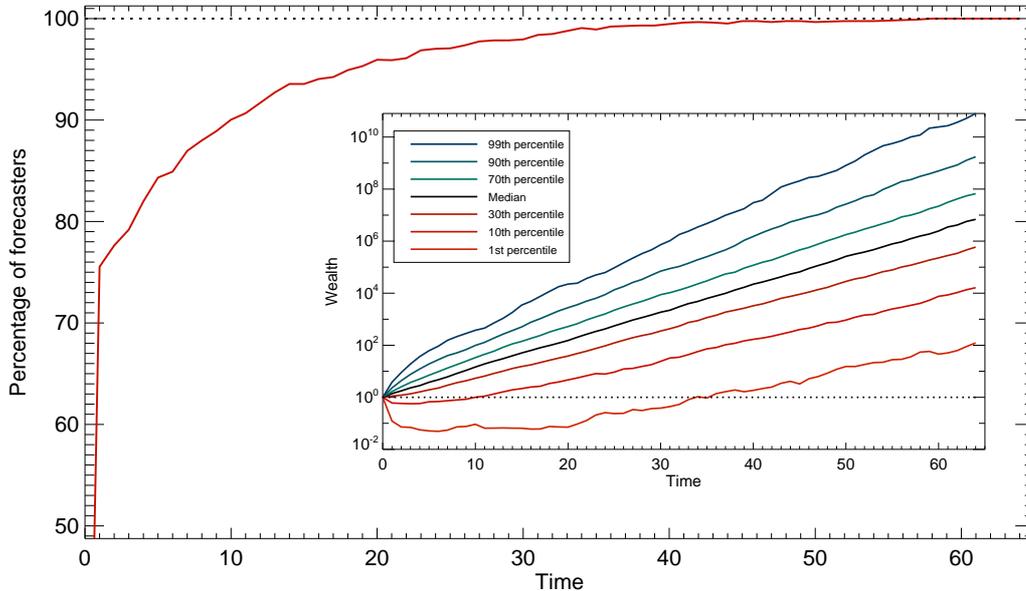
Figure 5.12: Time to forecast value: Percentage of 2^{11} independent underwriters expected to make a profit with time when betting against climatology using the imperfect model in a game of hurricane roulette (main plot), and frequency distribution of underwriters' wealth with time (inset plot). 99% of underwriters make a profit by 35 years, much earlier than the time for 99% of the statisticians to prove the skill of the forecast system (> 100 years for *IGN*).

The results of the Monte Carlo simulations of Hurricane Roulette are illustrated in Fig. 5.12. The percentage of 2^{11} independent underwriters who are likely to profit, and the frequency distribution of their wealth over time indicate that the underwriter is very likely to have made a non-trivial profit before two system cycles (i.e. 48 years) have even completed (NB: the initial phase, ϕ_0, is selected at random for each simulation to avoid bias). A comparison of the distributions of p-values in Fig. 5.11 to Fig. 5.12 reveals that the underwriter is highly likely to profit by betting on the statistician's hurricane forecast system before he is capable of proving the statistical significance of its skill using NHSTs.

5.5.3 Relationship between forecast skill and forecast value

A clear distinction between forecast skill and forecast value has been made in the previous sections, but is it actually possible to precisely quantify the relationship between the two metrics? Studies invariably conclude that the relationship is often complex [84, 171, 170]. For example, Richardson [170] found that an ensemble prediction system (EPS) which demonstrated little skill could still be of value to some users, and that there is more sensitivity in value (to increase in forecast ensemble size) than skill. Still, by comparing the results of the previous two sections, the intention is to provide guidelines here for using forecast scoring rules, and the relevance of their properties for forecast utility or value.

Figure 5.13 shows scatterplots of both estimates of wealth and relative ignorance (u and IGN) outcomes from 2^{11} Monte Carlo simulations of Hurricane Roulette. The strong degree of correspondence between ignorance and betting returns is evident, and is reflective of the fact that the Kelly betting variant of Hurricane Roulette is, like ignorance, proper [25]. Conversely, the relationship between forecast mean-outcome linear correlations r_{mean} and wealth exhibits a positive relationship, but there is more uncertainty than in the IGN plot. In fact, the relationship is not monotonically positive indicating that linear correlation coefficient cannot be considered a proper measure of forecast quality. There is no general answer to the question of whether a skilful forecast can be expected to be of value in application unless the quantities being forecast are based on the particular actions being taken.

5.6 Forward view and conclusions

A number of challenges posed by limited historical datasets and small-count data for forecasters aiming to make accurate predictions of hurricanes (or indeed other small-count predictands) have been discussed in detail in this chapter. These include the limitations on statistical inference with small-count data, the
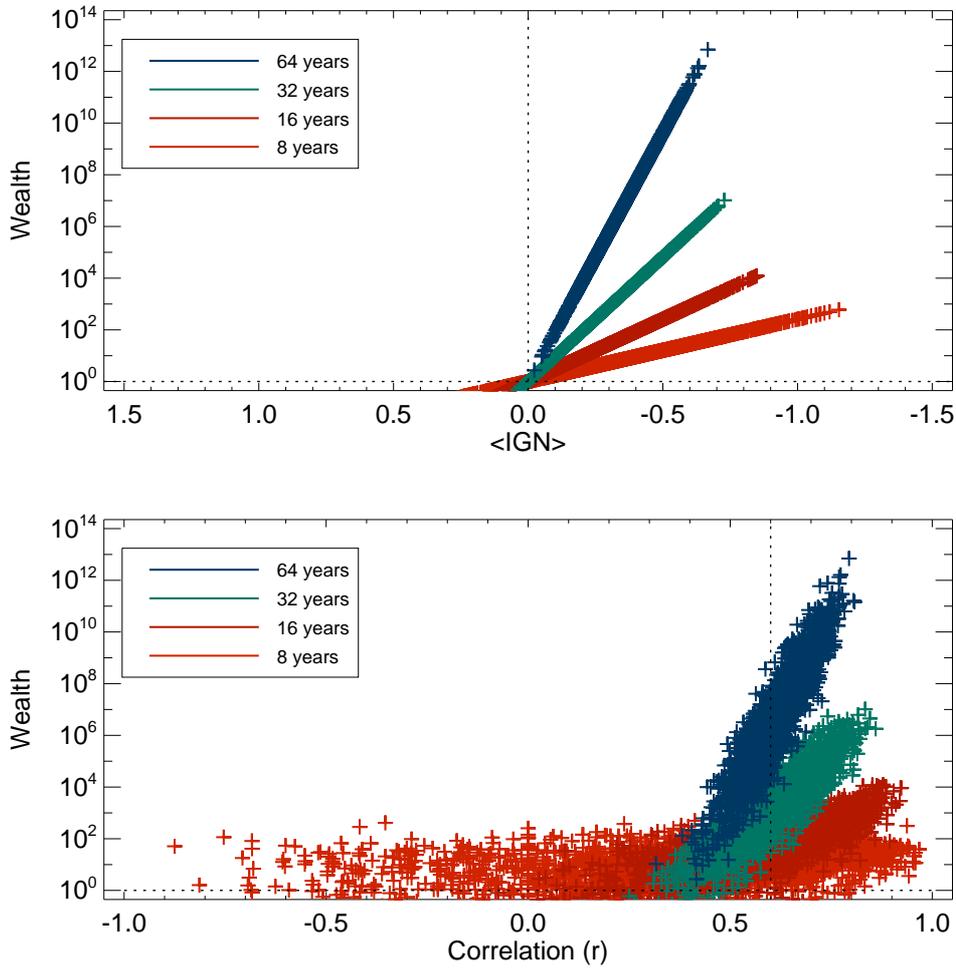
Figure 5.13: Bettor's wealth: Scatter plot of wealth vs ignorance (top) and wealth vs forecast mean-verification correlations (bottom) for 2^{11} underwriters who bet using the imperfect hurricane forecast model over different time windows. The vertical dotted line shows the threshold of relative skill (better than climatology) while the horizontal dotted line indicates the profit line. The relationship between IGN and wealth is strictly monotonic while the relationship between linear correlation r and wealth is not, highlighting the importance of employing proper scoring rules. NB: the x-axis in the top plot is negatively orientated.

difficulty in proving forecast skill with limited data that are collected at a slow rate, and deciding whether to utilise forecast information which lacks statistical confidence, but is potentially valuable.

Two novel univariate and bivariate statistical predictive techniques to exploit

the reliable data record have been proposed while best practices for robust forecast construction and evaluation have also been examined in depth. This research forms a part of the statistical framework for best-practice forecast construction, recalibration and evaluation proposed in this thesis. Firstly, a synoptic conditioning (SC) hurricane forecast system is described for the first time where forecast probabilities of hurricane count outcomes are conditioned on the El Niño-Southern Oscillation (ENSO), an index of periodic tropical Pacific SSTs and sea level air pressure. The influence of the El Niño phenomenon on tropical cyclone variability is well documented, and is considered an important predictor for seasonal hurricane predictions. The potential skill and value of the SC forecast system have been demonstrated in a Hurricane Roulette [67] betting scenario.

A second new forecast system for annual hurricane count predictions based on a univariate analogue forecasting has also been described. This technique is predicated on the basis that interannual and interdecadal hurricane activity exhibits periodic cycles. Both a single and series analogue method have been formulated where single or series of occurrences are found in the historical hurricane record to construct histograms and construct conditional forecast probability distributions. A novel top-hat kernel density estimation method has been introduced to smooth the constructed forecast PDFs, which are also blended with the climatological distribution to optimise the skill of the forecast PDF. Evaluation of forecasts produced of synthetic annual hurricane counts has shown that the forecast has higher skill than both a Bayesian forecast and the climatological forecast.

Insights into the limitations on making accurate predictions of hurricane counts with small-count data using conventional statistical inference have also been discussed. Skilful statistical forecasts of counts of annual hurricanes which make landfall over the North American coastline would be of tremendous value to the (re)insurance industry and government agencies. Given the small counts of this category of hurricanes, however, arriving at robust predictions through

statistical analysis is very difficult.

The relationship between forecast skill and forecast value has been examined with key distinctions illustrated in a "profit before proof" betting scenario. The "swindled statistician scam" demonstration is based on the assumptions that the statistician has constructed a skilful forecast system, and is not incentivised to do otherwise. The key purpose of the demonstration is to show that forecast skill and forecast value need not be confused. The results of the wealth and skill scatterplots in Section 5.5.3 indicate that there is an increasing monotonic relationship between the ignorance score and the profit made in a game of roulette (see Hagedorn and Smith [67]); which is not so evident in the wealth-linear correlation results. The predictive intervals in the wealth-correlation plots at all time windows are significantly larger than the wealth-ignorance scatterplots. This indicates that linear correlations are a less precise measure of the value of a forecast in the hurricane roulette/ Kelly betting scenario. The relationship between the two forecast evaluation measures is evident which shows that there exists a weaker trend at shorter time windows, and further reflects the unreliability of linear correlations as a corresponding evaluation of capital gain in the case of hurricane roulette.

The following are novel contributions or innovations in this chapter:

- formulation and evaluation of a new statistical conditioning hurricane forecast system utilising information about environmental conditions

- investigation of implications for statistical inference of U.S. landfall predictions where storm counts are small, and data are sparse

- formulation and evaluation of a new statistical empirical conditional analogue hurricane forecast system using temporal single and series analogues

- development and implementation of a novel top-hat kernel dressing method designed for forecast PDF smoothing with count data

- examination of the relationship between forecast skill and forecast value in an evaluation/betting scenario

Chapter 6

Evaluation and Reinterpretation of Atlantic Basin Tropical Cyclone Forecasts: 2012 Season

Chapters 2 and 3 have explored the evaluation and recalibration of binary forecasts in the context of a low-dimensional nonlinear dynamical system where it was concluded that forecast recalibration provides a straightforward and computationally cheap option for improving forecast quality. In this chapter, that forecast evaluation and recalibration framework is deployed for the first time in a novel real-world hurricane forecasting case study.

The subject of the evaluation/recalibration case-study in this chapter is the National Hurricane Center's (NHC) 48-hour tropical cyclone genesis binary forecasts from the 2012 hurricane season. The reliability of the NHC tropical cyclone genesis forecasts is assessed in Section 6.2 using reliability diagrams with consistency bars. Although reliability diagrams are published annually by the NHC to monitor the performance of its tropical cyclone genesis forecast system, it is argued that they are not in format which clearly quantifies forecast reliability. The performance of the 2012 NHC tropical cyclone (TC) genesis is reinterpreted here using reliability diagrams with consistency bars and on

probability paper to account for sampling uncertainty. The forecast system is shown to be mostly reliable over the 2012 season with some margin for improvement at several forecast probability categories. In Section 6.3, forecast recalibration is applied to the 2012 NHC TC forecasts (for the first time to the author's knowledge) to investigate whether or not it improves their reliability. Forecast recalibration evaluated in-sample is effectively meaningless. Hence, recalibration of the NHC TC forecasts is considered truly out of sample using the equivalent forecasts from the 2011 hurricane season as the training set, and via leave-one-out cross-validation [155]. It is shown that the reliability of the 2012 forecasts is decreased when calibrated with the 2011 forecasts, while it arguably increases when calibrated with the leave-one-out scheme. The relative improvement using the latter approach is attributable to the fact that the training and evaluation data are sourced from the same dataset (i.e. year), avoiding year-to-year variability in the joint forecast-outcome distribution.

In Section 6.4, an important characteristic of the NHC TC genesis forecasts which complicates the interpretation of forecast reliability is identified and investigated. The actual time duration between forecast and tropical cyclone formation, or "Time Until Event", varies between forecasts issued during the 2012 hurricane season. Given that forecasts are issued sequentially while a specific weather disturbance is tracked, the Time Until Event naturally decreases as the time of forecast issuance approaches the time at which that disturbance develops into a tropical cyclone. Consequently, there is a bias of reliability towards forecasts with shorter times until event. This concept has not been previously reported in the literature. The relationship between the Time Until Event and forecast probability category of the 2012 NHC TC forecasts is investigated, leading to an innovative proposal of diagrams to be included as supplementary to reliability diagrams in varying Time Until Event scenarios.

After the provision of technical background on NHC tropical cyclone genesis forecasts in Section 6.1, the remaining chapters contain the following new contributions: Section 6.2 presents an evaluation of the NHC TC genesis forecasts

from the 2012 hurricane season using reliability diagrams with consistency bars and on probability paper. These forecasts are then recalibrated for the first time in Section 6.3 by deploying the simple translation algorithm described in Chapter 2 using TC genesis forecast-outcome data from 2011 and leave-one-out cross-validation. Section 6.4 presents an investigation of variation in the Time Until Event of the NHC TC forecasts which has not been previously accounted for when assessing their reliability. Useful supplementary information for reliability diagrams is proposed for forecasting scenarios where Time Until Event is applicable.

6.1 NHC tropical cyclone genesis forecast overview

Throughout each hurricane season, the U.S. National Hurricane Center (NHC) publishes regular "Tropical Weather Outlooks" which report on significant regions of disturbed weather in the Atlantic basin, and their likelihood of development within two days. The outlooks include probabilistic forecasts of the development of these regions into tropical cyclones out to 48 hours as part of the NHC's operational remit [175]. The forecasts are subjective insofar as a duty forecaster assigns a probability of tropical cyclone genesis using observational data, objective reanalysis and global dynamical model output as guidance; the forecaster signs off on each forecast.

Every 6 hours during the Atlantic basin hurricane season, the NHC issues probability forecasts of the transition of a region of disturbed weather into a tropical cyclone (TC) up to 48 hours ahead. "Tropical Weather Outlooks" (TWO)[1], consist of a text forecast and a web display of satellite images,

---

[1]TWOs are issued at 2:00 AM EDT, 8:00 AM EDT, 2:00 PM EDT, 8:00 PM EDT and 1:00 AM EST, 7:00 AM EST, 1:00 PM EST, and 7:00 PM EST during the Atlantic hurricane season from June 1st until November 30th. "Special TWOs" are occasionally issued at asynoptic times if important changes to weather disturbances occur since the previous issuance (these are not included in this forecast evaluation exercise)

and are published daily on the NOAA website[2]. See Fig. 6.1 for an example of a
"Graphical Tropical Weather Outlook" image. Developing regions of disturbed
weather are circled and given a colour-coded probability of development into
a tropical cyclone (cyclogenesis) within 48 hours. Probabilities below 30% are
coloured yellow, $30-50\%$ are coloured orange, and above 50% are coloured red.
Each region is tracked for as long as there is a considered likelihood of cyclo-
genesis, and a probability forecast is issued up to every 6 hours. If a specific
region develops into a tropical cyclone then it is assigned a pictorial symbol
denoting its current classification of intensity. A tropical depression (a trop-
ical cyclone with maximum sustained winds of 38 mph or less) is assigned a
red-coloured symbol resembling a "$L\times$". If the cyclone develops further into a
tropical storm (a tropical cyclone with maximum sustained winds of 39 to 73
mph) then a red-coloured vortex-shaped symbol with an unfilled centre is as-
signed. Finally, to denote the cyclone's development into a hurricane (a tropical
cyclone with maximum sustained winds of 74 mph or higher), a red-coloured,
filled vortex-shaped symbol is used. The sequence of symbol assignment for
each forecast does not necessarily occur in this strict progressive order since
a tropical cyclone may develop rapidly within the 6-hour period in between
forecasts, and one or more stages may be skipped in the process. Probability
forecasts are also issued during the process of dissipation of the cyclone (cyclol-
ysis) if it is judged that a secondary cyclone may subsequently develop within
48 hours. In this case, the secondary cyclone is considered to be a separate
event occurrence to the first one. Tropical cyclone genesis forecasts are issued
"ad hoc" by a human forecaster (using model output and observational data
as guidance http://www.nhc.noaa.gov/about gtwo5.shtml?) generally in 10%
probability increments (i.e. 10%, 20%, 30%,...). Each probability increment can
be considered an individual forecast probability category. In the cases where 0%
probabilities are denoted on a graphical TWO, however, a "near-zero" probabil-

---

[2]details are provided at http://www.nhc.noaa.gov/aboutgtwo.shtml? and
http://www.nhc.noaa.gov/archive/gtwo/atl/latest
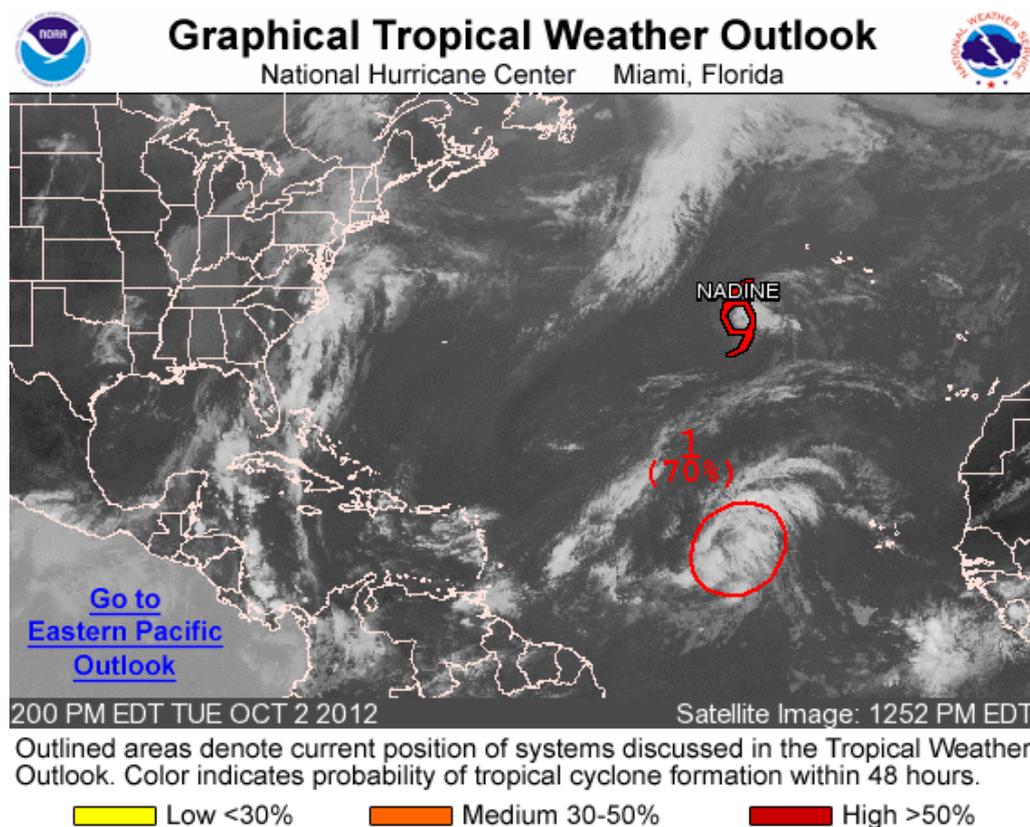
Figure 6.1: NHC Graphical Tropical Weather Outlook 2nd October 2012: an example of a graphical TWO issued by the NHC consisting of a satellite image containing symbols which indicate both regions of disturbed weather (circled area), and already formed tropical cyclones (red vortex symbol labelled "NADINE").

ity is stated in the accompanying text. There have also been two 1% probability forecasts issued for the 2012 season.

6.2   NHC 2012 tropical cyclone genesis forecast evaluation

The NHC's TC genesis forecasts represent probabilities of either the occurrence ($Y_i = 1$) or non-occurrence ($Y_i = 0$) of TC formation within 48 hours (i.e they are binary forecasts); hence, reliability diagrams are an appropriate tool to assess the quality of those forecasts. The NHC has been evaluating its TC

forecasts with reliability diagrams since 2007, but in a format that does not easily communicate the sampling error expected of observed frequencies of TC formation a reliable forecast system. The reliability of the 2012 TC forecasts is assessed here using the same reliability diagram format used in previous chapters (i.e. including consistency bars and on probability paper). Since the TC forecasts belong to discretised probability categories, the forecast values can be assumed to be fixed at each category. Forecast probability values of 0.5% have been assigned (only in this thesis) to the "near-zero" forecasts issued by the NHC so that the reliability of these forecasts can be quantified. The value of 0.5% has been selected because it represents the median of a continuous uniform distribution $\mathcal{U}(0,1)$ of probability values between 0% and 1%, where 1% is the next highest forecast probability category.

So, in the same notation as in Chapter 2, if X_i is a forecast value falling into category, or bin B_k and $I_k := \{i; X_i \in B_k\}$ denotes the set of indices i for which X_i falls into B_k then $X_i = r_k$ where

$$r_k = \frac{\sum_{i \in I_k} X_i}{\#I_k}, \tag{6.1}$$

is the bin average. Given that the forecast values are fixed at each bin, it might be tempting to construct consistency bars under the assumption that the observed frequencies f_k follow a binomial distribution with parameters I_k and r_k (see Section 1.6.4). Recall from Chapter 2 that these observed frequencies are expressed as

$$f_k = \frac{\sum_{i \in I_k} Y_i}{\#I_k}. \tag{6.2}$$

The consistency bars are then representative of sampling variations alone, and not additional uncertainty arising from varying I_k and r_k. Clearly, this method is also based on the assumption that the bin populations I_k are fixed, which is probably not justifiable in the case of the TC forecasts from year to year. In fact, the parameter I_k has a larger impact on the expected sampling uncertainty than

r_k [24], particularly where bin populations are low, as they are for the lowest and highest TC forecast probability categories. Hence, the *consistency resampling* method of Bröcker and Smith [24], employed in Chapters 2 and 3, has also been used here.

Reliability Diagram
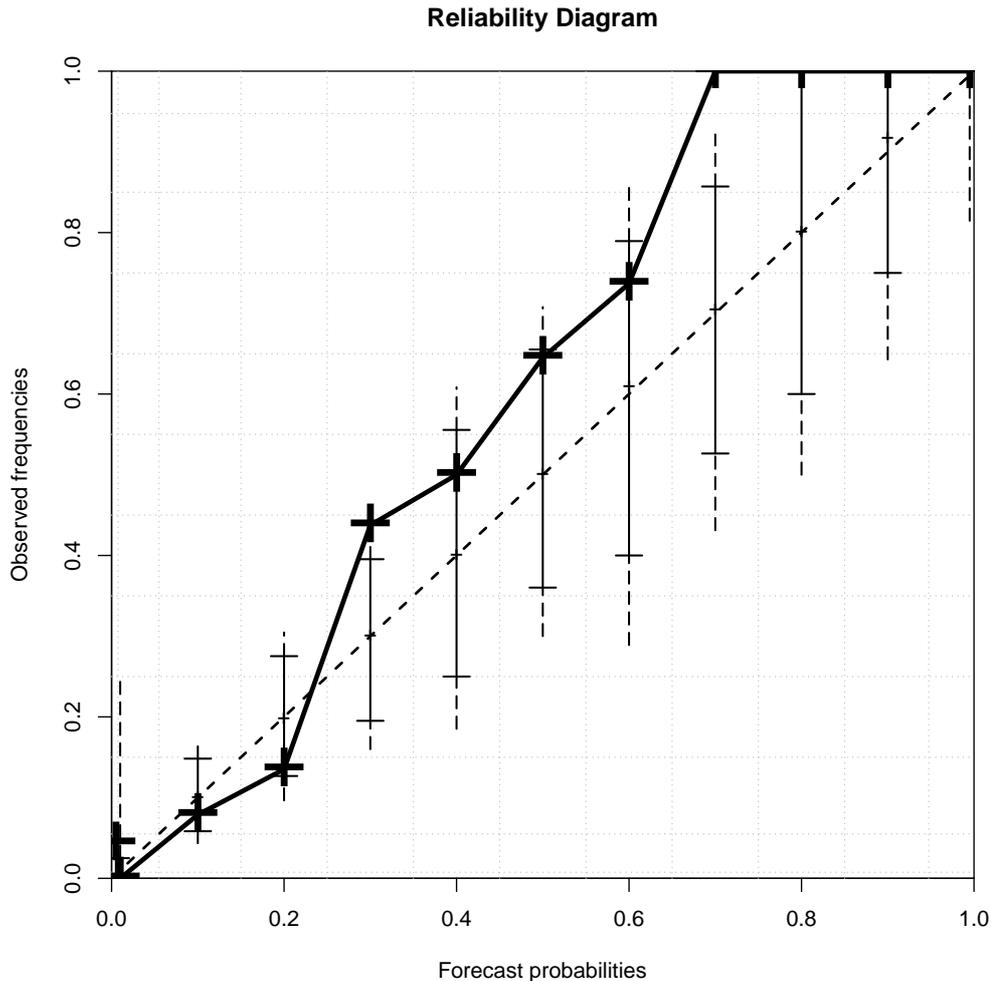
Figure 6.2: NHC 2012 TC forecast reliability: reliability diagram for the NHC's 2012 48-hr TC forecasts* with 5% - 95% (1% - 99% vertical dashed line) consistency bars. All but three forecast categories lie within the consistency bars, indicating that the forecast system is mostly reliable. The forecast probability bin boundaries (grey dotted lines) have been determined by taking the mid-points between each probability category value. *Sourced from NHC online Tropical Weather Outlooks.*

The reliability of the NHC's 2012 TC forecasts is conveyed by the reliability
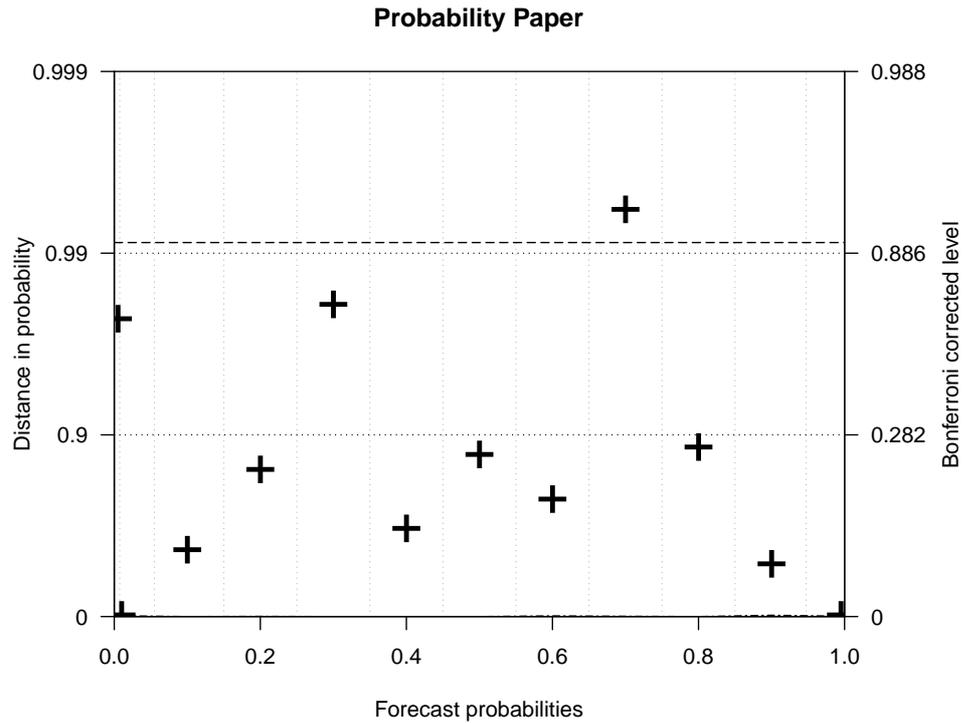
Figure 6.3: NHC 2012 TC forecast reliability: reliability diagram on probability paper for the NHC's 2012 48-hr TC forecasts* showing that all but three forecast categories are consistent with forecast reliability. The dash–dotted line denotes the exact position of the diagonal. The right-hand axis indicates the equivalent Bonferroni corrected levels i.e. for a reliable forecast, all of the points (12 categories) would be expected to fall within the 0.99 probability distance band with an 88.6% chance. In addition, the dashed lines indicate where the entire diagram would be expected to fall within with a 90% chance. The forecast probability bin boundaries (grey dotted lines) have been determined by taking the mid-points between each probability category value. *Sourced from NHC online Tropical Weather Outlooks.*

diagram and reliability diagram on probability paper in Figs. 6.2 and 6.3, respectively. The relevant forecast statistics are also tabulated in Table 6.1. Most of the observed frequencies lie inside the 5%-95% consistency bars in Fig. 6.2 and the 90% reliability band in Fig. 6.3, indicating a mostly reliable forecast system. The position of the reliability curve (solid line), however, suggests that NHC forecasters have overforecast (i.e. $f_k < r_k$) slightly for the lowest forecast probabilities (except for 0.5% probability forecasts), and have underforecast (i.e. $f_k > r_k$) to a larger degree for higher forecast probabilities, particularly for those higher than 0.6. This combination of an overforecast bias at lower probability bins and an underforecast bias at higher probability bins reflects a degree of under-confidence (see Wilks [217]). An interesting feature of the reliability diagram is exhibited at the 0.5% probability category where there is underforecasting due to two occurrences of TC development within 48 hours during the 2012 season. The distance between the observed relative frequency from the diagonal at the 0.5% bin compared to the 80% bin, for example, is smaller in Fig. 6.2 yet the distance on probability paper is considerably larger in Fig. 6.3 with the observed relative frequency lying outside the 5%-95% consistency range. This discrepancy in distance between the two diagrams is attributable to the differences in the values of I_k and r_k. Firstly, the sample size of the 0.5% bin ($\#I_1 = 46$) is larger than the 80% bin ($\#I_{10} = 14$) (see table 6.1), and secondly, the probability category 0.5% is more extreme (i.e. closer to 0 or 1), resulting in a more precise consistency bar (recall that the parameters I_k and r_k control the consistency bar width when employing the binomial consistency resampling approach above). The latter condition implies that, for a given sample size, there is greater sensitivity at the lowest probability bins to underforecasting and the highest probability bins to overforecasting. The difference between probability bins highlights the fact that consistency bars are necessary to reliably gauge the true extent to which a forecast system is calibrated from a reliability diagram.

Table 6.1: NHC 2012 TC forecast reliability diagram statistics

NHC Probability	Evaluation Probability	Observed relative frequency	Number in bin	Probability distance from diagonal
0	0.005	0.043	46	0.98
0.01	0.01	0	2	0
0.1	0.1	0.079	127	-0.55
0.2	0.2	0.135	74	-0.83
0.3	0.3	0.438	48	0.96
0.4	0.4	0.5	28	0.68
0.5	0.5	0.645	31	0.90
0.6	0.6	0.737	19	0.78
0.7	0.7	1.0	21	0.99
0.8	0.8	1.0	14	0.88
0.9	0.9	1.0	14	0.50
1.0	0.995	1.0	3	0

6.3 NHC 2012 tropical cyclone genesis forecast recalibration

The 2012 NHC TC genesis forecasts were shown to be some degree reliable in Section 6.2 using reliability diagrams with consistency bars, but can their reliability be improved using a simple recalibration scheme? As was shown in in Chapter 3, the largest improvements in recalibrated forecast skill appear to occur where the uncalibrated forecast skill is poorest (i.e. for small ensemble sizes and longer lead times), and/or where climatological probability of the event is closer to 0.5 (i.e. $\theta \rightarrow 0.5$).

The simple translation algorithm outlined in Section 2.4.1, although shown not to be the most effective of all the algorithms utilised in that chapter, is employed here to recalibrate the 2012 NHC TC genesis forecasts to assess the minimum achievable increase in forecast reliability. Forecast recalibration is carried out using two forms of cross-validation: one with the NHC's 2011 TC forecast-outcome dataset, and the other with leave-one-out cross-validation using the 2012 NHC TC genesis forecast-outcome dataset. So, for each probability

Reliability Diagram
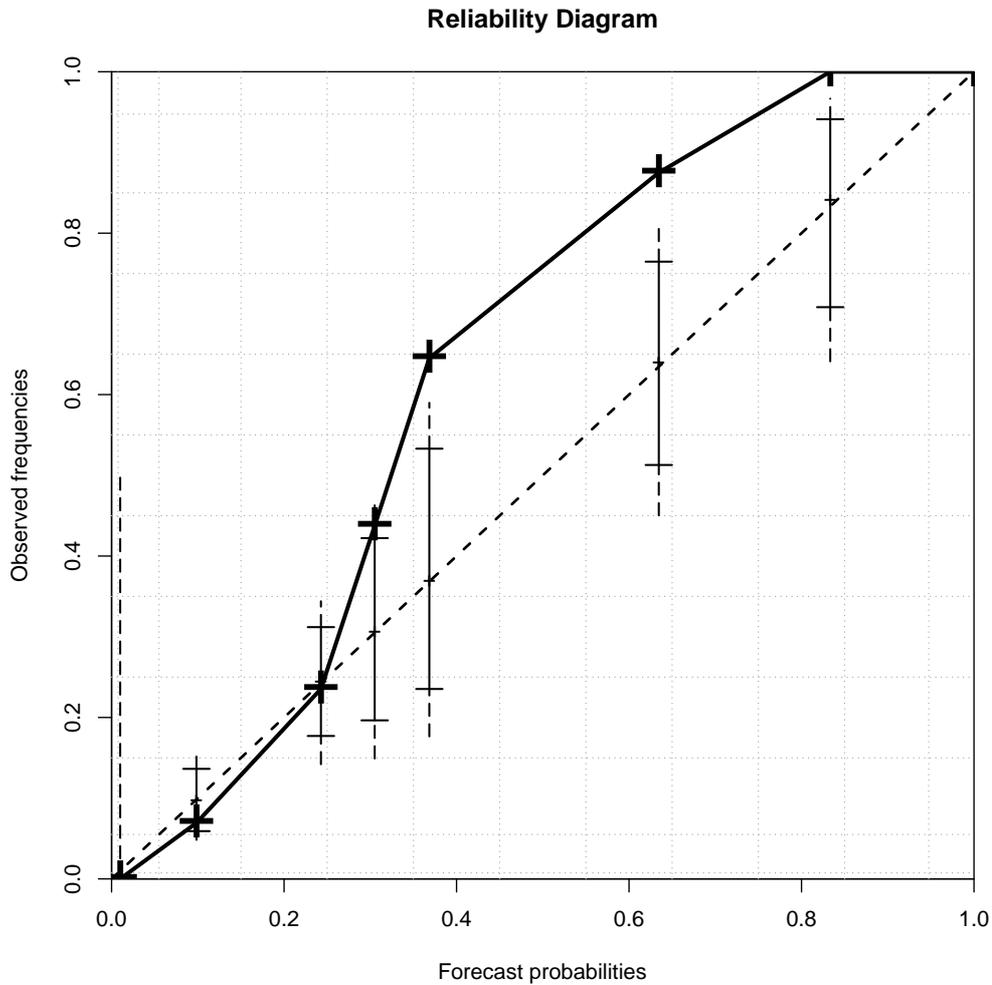
Figure 6.4: Recalibrated NHC 2012 TC forecast reliability: reliability diagram for the recalibrated NHC 2012 TC forecasts using 2011 forecasts as a training set with 5% - 95% (1% - 99% vertical dashed line) consistency bars. Forecast recalibration has resulted in a decrease of forecast reliability. The forecast probability bin boundaries (grey dotted lines) are identical to those on the original 2012 reliability diagram although the number of populated categories has decreased to 8. *Sourced from NHC online Tropical Weather Outlooks.*
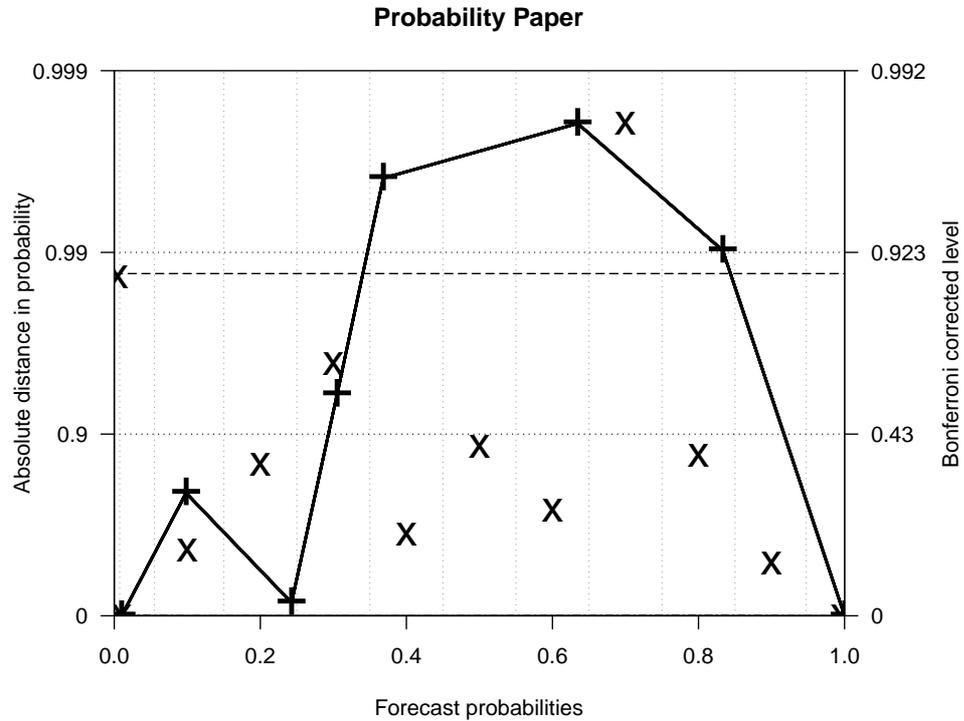
Figure 6.5: Recalibrated NHC 2012 TC forecast reliability: reliability diagram for the NHC 2012 TC forecasts recalibrated using 2011 forecasts as training set with 5% - 95% (1% - 99% vertical dashed line) consistency bars. Forecast recalibration has resulted in a decrease of forecast reliability since most recalibrated probability categories (pluses) have larger probability distances than raw forecast categories (crosses). For a reliable forecast, all of the points (8 categories) would be expected to fall within the 0.99 probability distance band with an 92.3% chance. The forecast probability bin boundaries (grey dotted lines) are identical to those on the original 2012 reliability diagram although the number of populated categories has decreased to 8. Refer to Fig. 6.3 for further details. *Sourced from NHC online Tropical Weather Outlooks.*

bin B_k in the evaluation set, the recalibrated probability p_i^{re}, $i \in I_k$ is equal to the observed frequency f_k^{train} corresponding to the same bin B_k in the training set. The recalibrated probability can be expressed as

$$p_i^{re} = f_k^{train}, \tag{6.3}$$

where r_k^{train} is the forecast probability category in the training set.

Figures 6.4 and 6.5 show the reliability diagrams of the recalibrated 2012 NHC forecasts evaluated truly out-of-sample using a training set of forecasts from the 2011 hurricane season. Like the raw 2012 forecasts, the recalibrated forecasts exhibit a significant underforecast bias at higher probability categories but to an even larger degree. Evaluated on their own, the 2011 forecasts demonstrate reliability to within $5\% - 95\%$ consistency (not shown), and so most of the 2012 forecast probabilities are only minimally adjusted after recalibration, suggesting why underforecasting is still evident at the higher categories in Fig. 6.4 (cf. Fig. 6.2). Note that the amount of populated forecast categories has decreased to 8 after recalibration. Figure 6.5 also displays the decrease in reliability with observed frequencies falling outside of the 90% Bonferroni threshold for 2 of the 8 forecast probability categories. Hence, recalibration has led to a decrease in the reliability of the 2012 TC forecasts. This deterioration of performance may, of course, be explainable by year-to-year variability in the quality of a forecast system, but also may be indicative of variability in the ocean-atmospheric conditions affecting the predictability of TC formation.

Figures 6.6 and 6.7 show the reliability diagrams resulting from recalibration of the 2012 TC forecasts with leave-one-out cross-validation. The reliability of the forecasts is significantly increased after recalibration, and are superior to those recalibrated with the 2011 TC forecast training set, with the exception of two probability categories, those with forecast averages $r_6 = 0.63$ and $r_8 = 0.78$. The simple translation recalibration algorithm with leave-one-out cross-validation benefits from the fact that the training set and evaluation set are almost identical, and so translating the forecast values in most bins is clearly effective. The reason for the two forecast categories with poor reliability is that there are two possible values for each recalibrated probability in each bin p_i^{re} where $i \in T_k$ is the collection of indices in the training set T_k for bin B_k. Since f_k^{train} can take two different values depending on the removed outcome (either

$Y_i = 0$ or $Y_i = 1$, respectively), since

$$f_k^{train} = \frac{\displaystyle\sum_{j \notin T_k} Y_j}{\#T_k - 1},$$

(6.4)

then so do the recalibrated forecast values p_i^{re}. Those recalibrated forecasts with corresponding outcome $Y_i = 0$ take a higher value than the forecasts with corresponding outcome $Y_i = 1$, and may be translated to separate vacant bins, resulting in $f_k^{re} = 1$ or $f_k^{re} = 0$. The margin between the two values increases with decrease in bin population $\#T_k$, increasing the likelihood of separation when binning. This result reflects the problems with reliability diagram forecast categorisation (see Section 3.4.1), but also the effect small that sample size can have. Of course, deploying the leave-one-out method in operational forecasting to recalibrate forecasts in real time is not practical given that the full training/evaluation dataset would not be complete until the end of the season. Instead, it might be used to retrospectively recalibrate a forecast system to be employed in the following season.

6.4 Time Until Event

An important characteristic of the NHC TC forecasts is that, even though they represent predictions of the formation of a tropical cyclone out to 48 hours ahead at the time of issuance, TC formation actually often occurs well within 48 hours. In fact, there is an inversely proportional empirical relationship between forecast probability values and Time Until Event (TUE). The association between forecast probability and TUE complicates both the interpretation of the overall reliability of the forecast system, and comparisons of the performance of different probability categories. Not only is there a bias towards smaller sample sizes at higher probability categories, but also a bias towards shorter TUE lengths at higher probability categories. Given that many of the forecasts of higher probability value are issued closer to the time of TC formation, one
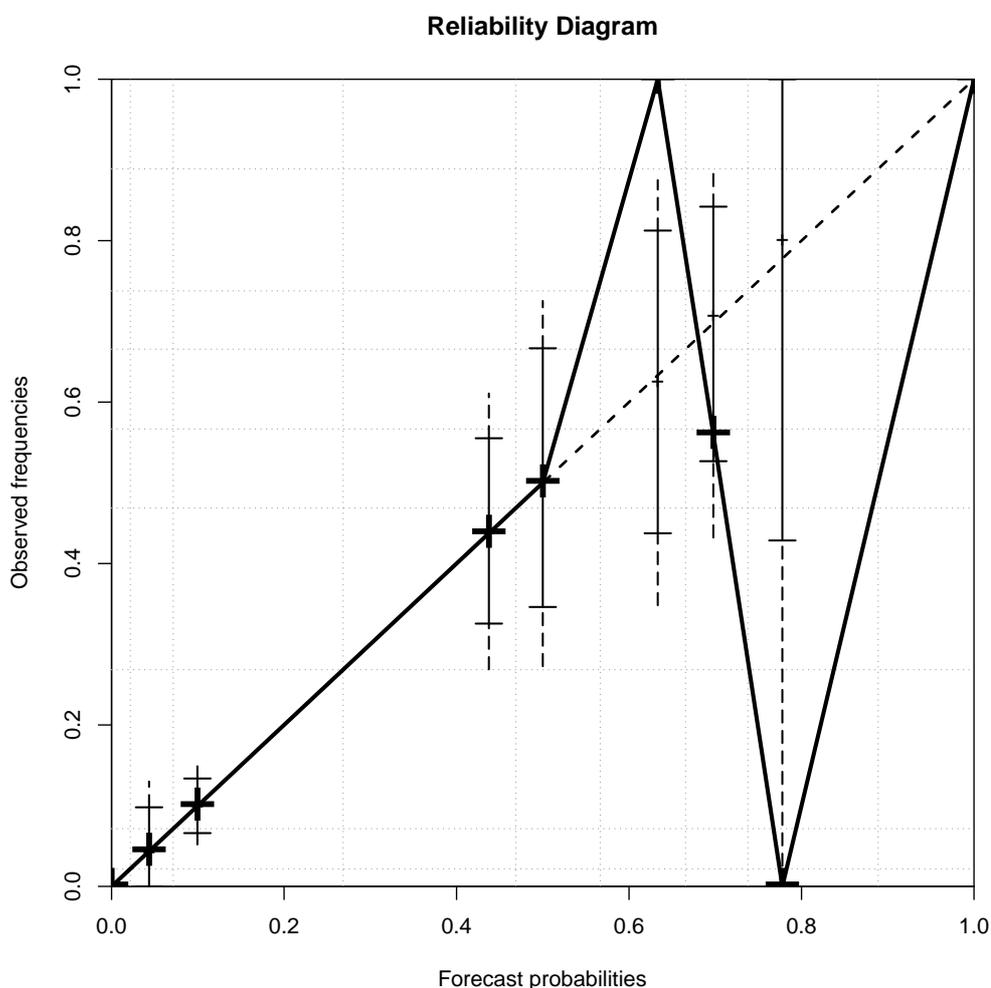
Figure 6.6: Recalibrated NHC 2012 TC forecast reliability: reliability diagram for the NHC 2012 TC forecasts recalibrated using leave-one-out cross-validation with 5% - 95% (1% - 99% vertical dashed line) consistency bars. Six of the nine recalibrated forecast probability categories lie on the diagonal indicating perfectly reliability while two others lie completely outside their corresponding consistency bars. The reliability curve shows that leave-one-out recalibration can both significantly improve and decrease reliability depending on the categorisation of the forecasts. The forecast probability bin boundaries (grey dotted lines) are identical to those on the original 2012 reliability diagram. *Sourced from NHC online Tropical Weather Outlooks.*

Probability Paper
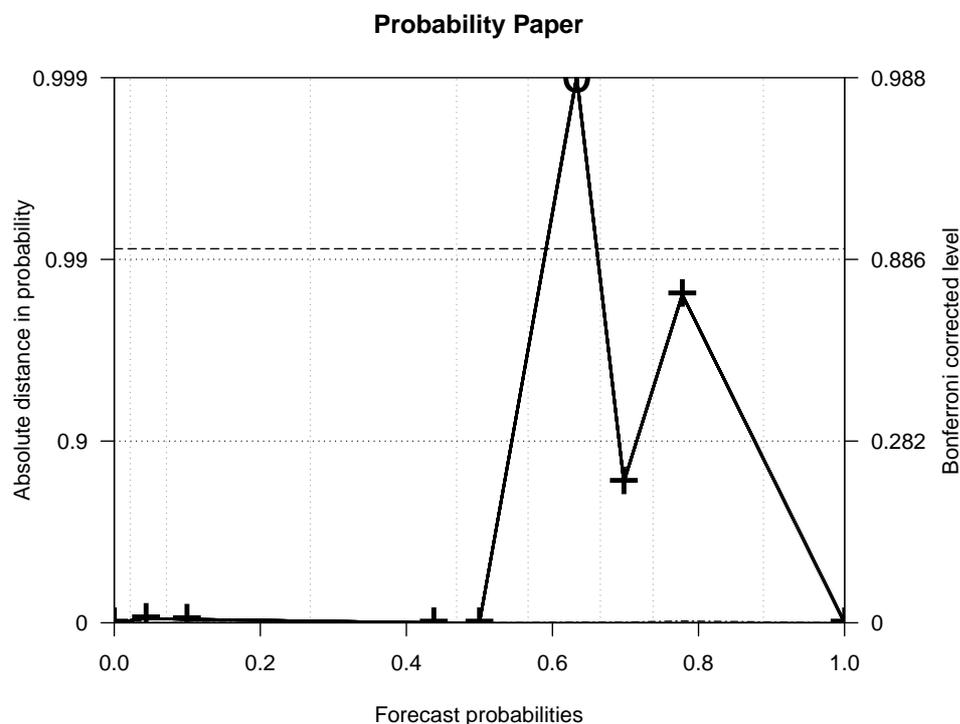
Figure 6.7: Recalibrated NHC 2012 TC forecast reliability: reliability diagram for the NHC 2012 TC forecasts recalibrated using leave-one-out cross-validation with 5% - 95% (1% - 99% vertical dashed line) consistency bars. Seven of the nine recalibrated probability categories (pluses) have smaller probability distances than raw forecast categories (crosses). The reliability curve shows that leave-one-out recalibration can both significantly improve and decrease reliability depending on the categorisation of the forecasts. All of the points (9 categories) would be expected to fall within the 0.99 probability distance band with an 91.4% chance. The forecast bin boundaries (grey dotted lines) are identical to those on the original 2012 reliability diagram. Refer to Fig. 6.3 for further details. *Sourced from NHC online Tropical Weather Outlooks.*

is left to conclude that there would also be a reliability bias at those higher probability categories. While sample size is accounted for by consistency bars, information about the TUE is not conveyed on reliability diagrams. Hence, in such forecasting scenarios, it is important to communicate the variability of TUE with forecast probability, unlike scenarios where forecasts have a fixed lead time, for example, what is the probability air temperature will be above

a given threshold at 24 hours lead time? A second companion diagram, or set of diagrams, to the reliability diagram is proposed here to communicate the distribution of TUE for each forecast probability category, and provide a more comprehensive picture of forecast reliability. By comparing the fractions of forecasts having different TUE lengths across forecast probability categories, it can be determined whether there might be a performance bias towards any particular forecast probability value r_k.
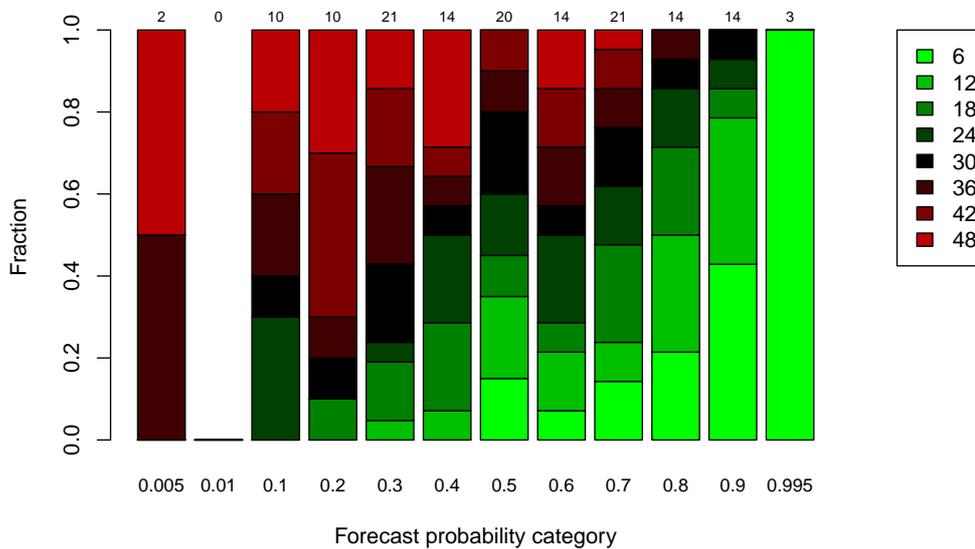
Figure 6.8: NHC 2012 TC forecast Time Until Event: fractions of verifying NHC 2012 TC forecasts* having different TUE lengths (in hours) for all probability categories. The coloured TUE categories denote the occurrence of TC formation between the time given and 6 hours previous to it. There is a clear pattern of larger fractions of shorter TUE with increasing forecast probability category. Total counts of verifying forecasts for each category are shown at the top of the bars. *Sourced from NHC online Tropical Weather Outlooks.*

Figure 6.8 shows the fractions of NHC 2012 forecasts which verify with a TC formation within 48 hours ($Y = 1$) at each probability category r_k. Given a TC formation event during the 2012 hurricane season, it is evident that there is a significant amount of variation in the proportions of TUE lengths, and
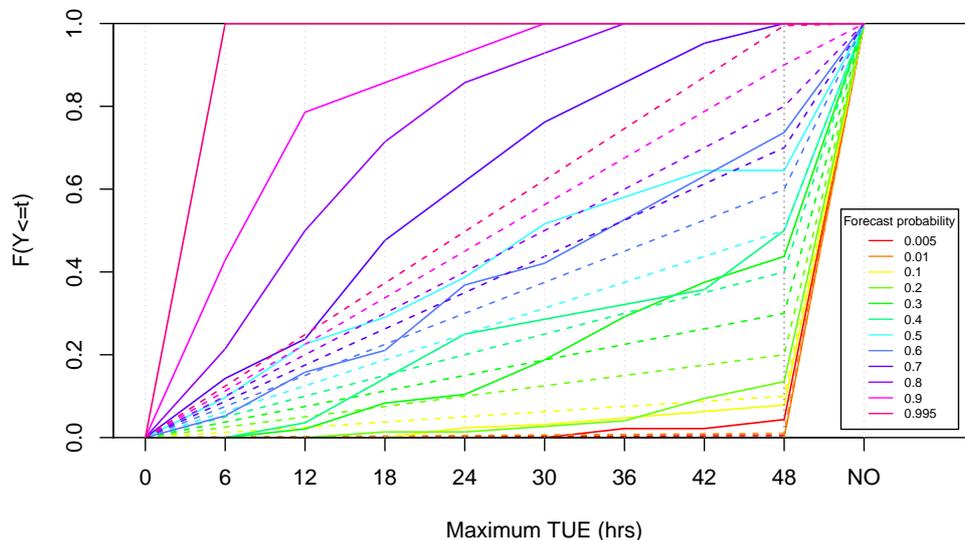
Figure 6.9: NHC 2012 TC forecast Time Until Event: CDFs of NHC 2012 TC forecast* TUE times (in hours) for each forecast probability category r_k (solid lines), and for a set of reliable forecasts ($f_k = r_k$) where the TUE times are computed with a discrete uniform distribution function (dashed lines). The higher probability curves lie well above the corresponding uniform distribution of reliable forecast TUE lengths. The TUE categories indicate the occurrence of a TC event between the time given and 6 hours previous to it, and "NO" indicates a non-occurrence of a TC within 48 hours. *Sourced from NHC online Tropical Weather Outlooks.*

that there is a pattern of shorter TUE with increasing forecast probability, suggesting there may be a reliability bias towards the higher bins. Caution should therefore be exercised when comparing the reliability of the different categories of the NHC 2012 forecasts. For example, approximately 40% of the 90% probability forecasts are verified within 6 hours whereas not a single 0.5% probability forecast is verified within 30 hours. Given this difference, the expectation would be for the forecasts in the 90% category to perform more reliably since they were issued nearer in time to the formation of a tropical cyclone. Figure 6.9 compares the cumulative distribution functions for the maximum TUE lengths of the actual forecasts in each probability category

231
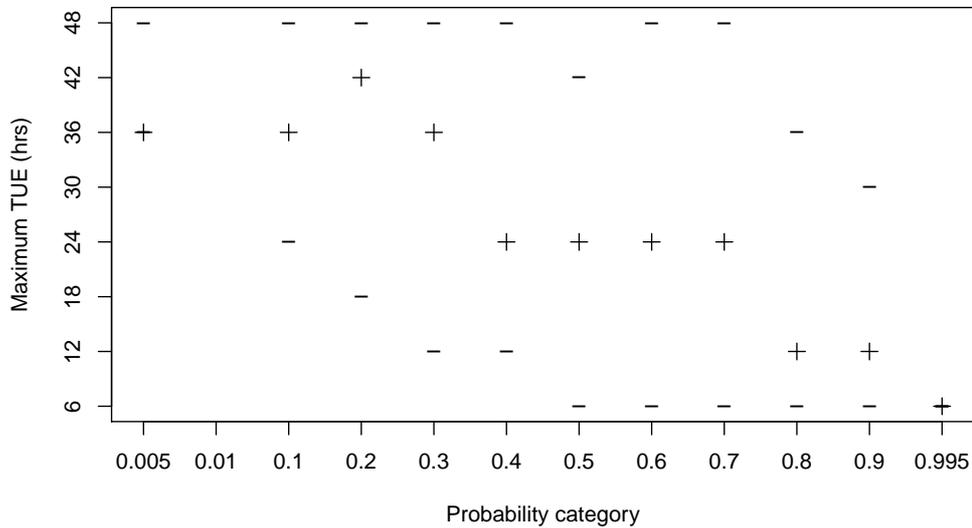
Figure 6.10: NHC 2012 TC forecast Time Until Event: Maximum, minimum (minuses) and median (pluses) of verifying NHC 2012 TC forecast* TUEs for each forecast probability category, r_k. The TUE time categories indicate the occurrence of a TC event between the time given and 6 hours previous to it. **Sourced from NHC online Tropical Weather Outlooks.*

with those for a set of reliable forecasts ($f_k = r_k$) for which the maximum TUE lengths are uniformly distributed. For example, a reliable forecast of probability category 50% with uniformly distributed maximum TUE would be expected to verify with an event occurrence within 48 hours 50% of the time, within 24 hours 25% of the time, etc. Instead, the NHC 2012 50% forecasts have a distribution skewed towards shorter maximum TUE lengths (compare the solid and dashed turquoise curves). In fact, all of the probability category curves above 40% lie entirely above their corresponding uniformly distributed maximum TUE curves, reflecting the reduced TUE lengths at those categories. Conversely, the lowest probability categories (0.5% − 20%) exhibit a bias towards longer TUE lengths, suggesting that achieving reliability is more difficult at those categories. Finally, Fig. 6.10 shows simpler versions of the distributions of

maximum TUE lengths for each probability category by displaying the median, minimum, and maximum values. The pattern of decreasing maximum TUE length with forecast probability category is again clearly evident. The reliability diagram statistics r_k and f_k are decomposed into two TUE ranges are listed in table 6.2. The values in the cells show the observed frequency f_k of each subset of forecasts with TUE lengths falling into either the range 0-24 hours or 24-48 hours. In both cases, the colour coding indicates where each observed frequency lies with respect to the consistency bars. Red coloured values fall outside the $1 - 99\%$ consistency interval while green coloured values fall within the $5\% - 95\%$ range, indicating forecast reliability. Orange values indicate the remaining $8\%ile$ range. The data in table 6.2 reveal a tendency for improved reliability of higher forecast probabilities at shorter TUE lengths, and improved reliability of lower forecast probabilities at longer TUE lengths (except for the lowest probability category 0.5%).

The underforecast bias at higher probability categories may be reflective of conditions being more favourable for tropical cyclone formation than expected, or of a physical phenomenon known as rapid intensification (RI) where the development of a tropical cyclone progresses rapidly over the last hours before its formation [93]. Investigation of such causes is beyond the scope of this thesis however.

Table 6.2: NHC 2012 TC forecast reliability diagram statistics by TUE

TUE	Forecast probability r_k											
	0.005	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0-24 hrs	0	0	0.024	0.014	0.104	0.25	0.387	0.368	0.619	0.857	0.929	1.0
24-48 hrs	0.044	0	0.055	0.122	0.333	0.25	0.258	0.368	0.381	0.143	0.071	0

Green $5\% - 95\%$; Orange $95 - 99\%$; Red $> 99\%$.

6.5   Forward view and conclusions

The performance of the NHC's 2012 short-term Atlantic basin tropical cyclone (TC) genesis forecasts has been assessed in this chapter using reliability diagrams. The adopted diagram format is based on that proposed by Bröcker and Smith [24] which includes $5\% - 95\%$ consistency bars. These consistency bars, quantified using a consistency bootstrap resampling technique, represent the sampling error expected of observed relative frequencies. The NHC's 2012 TC forecasts have been shown to have performed reliably for less extreme probability categories with some overforecast bias at the lower probability categories and underforecast bias at the higher probability categories. Subsequently, an assessment has been made whether recalibrating the 2012 forecasts with a simple translation algorithm could lead to improvement of their reliability. The recalibration procedure has been deployed out-of-sample using both the previous year's (2011) forecast-outcome dataset as training data, and via leave-one-out cross-validation.

Recalibration resulted in a decrease of forecast reliability with the previous year training data, and in improved reliability at most forecast probability categories with the leave-one-out approach. The decrease of reliability occurs as a result of year to year variability in both the quality of forecasts, and the predictability of TC formation. Hence, recalibration across years has been demonstrated not to be beneficial for, and in fact, degrades the reliability of the NHC's TC genesis forecasts in the cases considered. Of course, limited sample size of the training set also restricts the effectiveness of recalibration. Establishing robust conclusions on the matter, however, is beyond the scope of this thesis. The general improvement in reliability after leave-one-out cross-validation is a result of the similarity between the training and evaluation data, and would not likely be realistic in real-time forecast recalibration since the full training set would not be available until the end of the hurricane season.

The concept of "Time Until Event" and its relationship with the reliability

of the NHC 2012 TC forecasts has been also explored. A bias towards shorter TUE lengths has been exhibited at higher probability categories, indicating a reliability bias at those categories which is supported by the information contained in table 6.2. The 70% category is shown to be unreliable, however, with a substantial margin of underforecasting. Information on the impact of TUE on forecast reliability is not provided in conventional reliability diagrams. Hence, supplementary material such as table 6.2, and Figs. 6.8, 6.9, and 6.10 are recommended to accurately and robustly interpret the reliability of forecast systems in Time Until Event scenarios.

The insights gained from the research in this chapter are as follows:

1. the National Hurricane Center's short-term Atlantic basin tropical cyclone (TC) genesis forecasts from the 2012 hurricane season were generally reliable with a degree of overforecasting at the lower probability categories and underforecasting at the higher probability categories

2. increasing the reliability of the National Hurricane Center's short-term Atlantic basin tropical cyclone (TC) genesis forecasts after out-of-sample recalibration is difficult due to inter-annual variability in forecast distributions and predictability of tropical cyclone formation, and/or because of the limitations of recalibration algorithms

3. varying forecast "Time Until Event" complicates the interpretation of the reliability of forecast systems such as NHC short-term Atlantic basin tropical cyclone forecast system

The novel contributions of this chapter are:

- evaluation NHC 2012 short term TC genesis forecasts using reliability diagrams both with consistency bars and on probability paper to quantify forecast reliability

- out-of-sample recalibration of NHC 2012 short term TC genesis forecasts using a simple translation algorithm using the 2011 forecast-outcome dataset and leave-one-out cross-validation. In the first instance, recalibration failed to increase forecast reliability while the second approach was more effective given that the training and evaluation data are from the same hurricane season.

- evaluation of the relationship between NHC short term TC genesis forecast reliability and Time Until Event

- proposal of supplementary diagrams/tables to reliability diagrams which provide additional information about the effect of Time Until Event on forecast reliability where relevant

Chapter 7

Hurricane Count Modelling (long-term lead)

Statistical modelling studies of seasonal to decadal Atlantic basin tropical cyclone activity are diverse and numerous in the literature to date. Intensified research on long-term hurricane activity has been motivated by both the severe and increasing impacts [157, 148] caused by hurricane landfalls, and scientific interest in the physical mechanisms that control cyclone formation and development. Although there have been significant improvements in modelling techniques, there is a large degree of uncertainty in long-term projections [164], and the out-of-sample skill of seasonal predictions is still yet to be proven [208, 46, 156]. These limitations are due to uncertainty in the relationships between predictor variables and tropical cyclone (TC) activity, the difficulty in distinguishing between natural variability and long-term trend, and the relatively short length of a reliable historical record of tropical cyclone statistics [106, 76]. As a result, there has been a substantial amount of debate on the existence of long-term TC trends over the past century [102, 105, 103]. There does appear to be some evidence, however, that the frequency of the more intense TCs has increased since the 1970's [47, 52, 102, 210]. A more contentious issue is whether any detectable trends can be attributed to anthropogenic cli-

mate change [122, 78, 9]. In spite of the uncertainty surrounding long-term trends in hurricane activity, however, statistical hurricane modelling studies are worthwhile to improve our understanding of which physical mechanisms are important for the modulation of hurricane activity.

The relationship between long-term Atlantic basin hurricane behaviour and various environmental physical indices such as tropical Atlantic sea surface temperatures (SSTs), El Niño-Southern Oscillation (ENSO), and the Atlantic Multidecadal Oscillation (AMO) has been rigorously examined in the literature by means of regression models [64, 74, 46, 201]. A number of generalised linear model (GLM) and generalised additive model (GAM) techniques, based on those used in recent studies by Villarini et al. [207, 208] and Mestre and Hallegate [129], are employed in this chapter to model key categories of TC activity in the Atlantic basin. Whereas Villarini et al. [208] model CAT1-5 Atlantic basin hurricane and CAT1-5 U.S. landfall counts, and fractions of those hurricanes making landfall at the U.S. coast (i.e. CAT1-5 U.S. landfall fractions), the modelling of counts and fractions is extended here to include Atlantic basin named storms and CAT3-5 basin hurricanes, and the fraction of CAT1-5 Atlantic basin hurricanes intensifying into CAT3-5 basin hurricanes. Additional GLM and GAM modelling techniques which include both *polynomial* and *cubic spline regression smoothers* [72] are employed to examine both linear and nonlinear dependencies between response and predictor variables.

This chapter is structured as follows: Section 7.1 outlines the definitions of the GLMs and GAMs used to model hurricane counts and fractions, and the model fitting process which includes the use of quadratic polynomial and cubic spline regression smoothers. These smoothers allow for estimates of trends in a response variable that vary less than the response variable itself. Collinearity between predictor variables is also considered by modelling hurricanes with interaction terms [36]. This specific regression modelling framework, in conjunction with a unique combination of predictor variables, is a novel contribution to this thesis.

Section 7.2 describes the selection criteria which are utilised to select the most appropriate model fit. The Akaike Information Criterion (AIC; [5]) and Schwarz Bayesian Criterion (SBC; [176]), sometimes known as the Bayesian Information Criterion (BIC), are conventional model selection measures which rank the goodness-of-fit of a model according to a trade-off between model complexity and accuracy. Burnham and Anderson [27] recommend that the *corrected* Akaike Information Criterion (AIC_c) is used where the relative number of model parameters is large compared to data sample size. Given that size of the reliable historical hurricane count record is limited, the AIC_c is the preferred selection criterion here.

In addition to presenting the estimates of the model parameters along with their standard errors, as in Villarini et al. [208], confidence intervals for the parameter estimates constructed in Section 7.3. *Wald* confidence intervals are typically constructed for statistical inference of regression model parameters, but, given the limitations of small sample sizes of count data, inverted *score-test* and *likelihood-ratio* confidence intervals perform better so that actual error probabilities are close to their nominal levels [1]. Computation of these two inverted test confidence intervals is difficult, however, because they are dependent on the log-likelihood function which is not an explicit function of a regression model's parameters. A new 'sliding quadratic' root-finding algorithm for confidence interval construction based on a method proposed by Lang [110] is proposed in Section 7.3 as an alternative to constructing the inverted score-test (henceforth referred to as score) and inverted likelihood-ratio (LR) confidence intervals.

Generalised linear models and generalised additive models (GAM) models of count data are often subject to *overdispersion* [2] where the data have greater variability than expected by the model. Although previous tropical cyclone modelling studies have taken overdispersion into consideration [207], a test based on an auxiliary *ordinary least squares* regression is described in Section 7.4, and has been employed here for the first time to test for overdispersion in

tropical cyclone modelling.

Finally, the results of GLM and GAM modelling of tropical cyclone counts and fractions are presented in Section 7.5. The best-fit regression models in each hurricane category have been employed to produce predictions of the 2013 seasonal hurricane counts and U.S. landfall fractions in Chapter 8.

7.1 Modelling Atlantic basin and U.S. landfall hurricanes using GLMs and GAMs

Two types of regression model have been used to model hurricanes here. Firstly, a Poisson regression is employed to model annual counts of Atlantic basin named storms, Atlantic basin hurricanes, and U.S. landfalling hurricanes. Poisson regressions have emerged as the canonical method over the past couple of decades for modelling annual hurricane counts [46]. The sophistication of Poisson regression hurricane models has developed over time to incorporate a range of climate indices known to modulate regional hurricane activity [50, 48] as predictors, and to account for any non-linear dependencies of annual counts on these predictors. A relatively straightforward, although unique, combination of predictor variables is opted for here: year and global tropical mean and tropical Atlantic sea surface temperature anomalies. Both of the latter two environmental indices relate to the physical factors which modulate Atlantic basin hurricane activity. Tropical Atlantic sea surface temperatures are known to have a strong influence on hurricane activity because a warmer Atlantic Ocean supplies more available energy for cyclone formation [51, 203], and global tropical sea surface temperatures tend to control the atmospheric conditions such as windshear [64]. Tropical Atlantic sea surface temperature non-detrended anomaly data[1] is spatially averaged over a box 10°-25°N and 80°W-20°W, while the global

---

[1]sourced from the National Oceanic and Atmospheric Administrations (NOAA) Extended Reconstructed sea surface temperature (ERSSTv3b; [187])

tropical sea surface temperature anomalies are spatially averaged over a zonal band 30°S-30°N. Both datasets are temporally averaged over the period June to November, as in Villarini et al. [208].

Secondly, a logistic regression has been employed to model fractions of annual hurricane counts making landfall at the U.S. coastline [208], and the fractions of Atlantic basin hurricanes which develop into intense CAT3-5 hurricanes. Both response variables are regressed on the same predictor variables described above. Both sets of modelled hurricane categories are modelled over the period 1966-2012, which is the period of the reliable hurricane record available from the National Hurricane Center's (NHC) HURDAT database[2]. Figure 7.1 displays the time series of Atlantic basin named storms, CAT1-5 Atlantic basin hurricanes, and CAT1-5 U.S. landfalls.

Both the Poisson and logistic regressions fall into the class of generalised linear models [128]. A GLM is a linear regression technique applied to non Gaussian-dependent variables whose distribution belongs to the so called "exponential family" : Poisson, Gamma, Binomial, Gauss. GLMs can be used to determine the relative importance of various predictor variables on hurricane formation although they should be interpreted in conjunction with physical reasoning. Poisson regression and logistic regression models also fall into the class of generalised additive models (GAMs) which blend properties of GLMs with additive models to account for any non-linear dependence between the response variable and the predictor variables. Simple versions of both these models are deployed here by regressing various hurricane category annual counts on year, and tropical Atlantic and global tropical sea surface temperatures. Formal definitions of the GLMs and GAMs are now provided along with their mathematical notation.

GLMs are generalisations of ordinary linear regression models that allow for a non-normal distribution in the response variable based on the assumption that the predictor effects are linear in the parameters [30]. Let the linear predictor

---

[2]http://www.aoml.noaa.gov/hrd/data_sub/re_anal.html
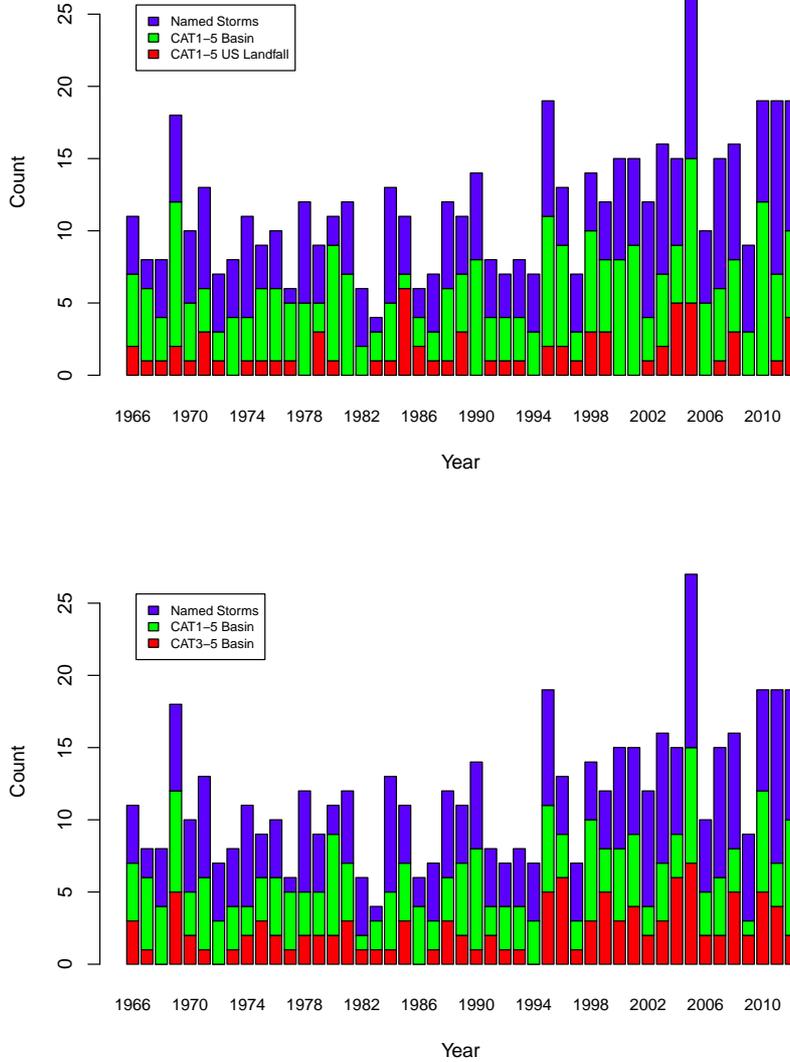
Figure 7.1: Time series of all annual Atlantic basin named storm counts from 1966-2012 with CAT1-5 basin hurricanes and CAT1-5 U.S. Landfalls shown as sub-categories (top), and CAT1-5 basin hurricanes and CAT3-5 Basin hurricanes shown as sub-categories (bottom).

be defined as

$$\eta = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{7.1}$$

where $\mathbf{x}_i^T = [x_{1i}, ..., x_{ki}]$ is the vector of predictors, and $\boldsymbol{\beta}$ is the $k \times 1$ parameter vector. The link function $\eta = \eta(\mu)$ relates the linear predictor to the mean μ of the distribution of the response variable.

A GAM is designed so that the linear predictor η is not restricted to a linear dependence on the predictors or *covariates*. The linear component $\mathbf{x}_i^T\boldsymbol{\beta}$ of the GLM model is substituted with an additive component of the form $\sum_{j=1}^{k} f_j(x_j)$ where $f_j(\cdot)$ are smooth univariate, or *basis*, functions, one for each covariate [72]. These basis functions define "transformed predictors" $f_j(x_j)$ which act additively on the response variable. The GAM now implies that the conditional mean is given by

$$g(E[Y_i|\mathbf{x}_i]) = \beta_0 + \sum_{j=1}^{k} f_j(x_{ij}), \tag{7.2}$$

where x_{ij} is the jth component of $\mathbf{x}_i^T$ and $g(\cdot)$ is prescribed by the type of model e.g. for a Poisson regression $g(\cdot) = \log(\cdot)$.

GLMs and GAMs are employed here to model both linear and nonlinear dependencies of annual storm counts and fractions on year, Atlantic basin sea surface temperature anomalies (SST_{Atl}), and global tropical sea surface temperature anomalies (SST_{trop}). A Poisson regression is used to model counts of Atlantic basin named storms, CAT1-5 and CAT3-5 basin hurricanes, and CAT1-5 U.S. landfalls while a logistic regression is used to model CAT1-5 U.S. landfalls and CAT3-5 basin hurricanes both as a fraction of the total number of annual CAT1-5 basin hurricanes. The computational regression analysis is carried out with the GAMLSS package available in the R statistical programming language [190].

7.1.1 Poisson regression model

The standard model used for modelling annual hurricane counts is the Poisson regression [46] although there is some suggestion that it may be a better fit for the intense hurricane (CAT3-5) category given the smaller counts [50]. Annual hurricane counts (Y_i) are modelled based on the assumption that the counts follow a Poisson distribution. If these counts are defined by Y_i in the ith year conditional on the vector $\mathbf{x}_i$ of predictors variables then Y_i is Poisson distributed

with density

$$f(Y_i = y|\mathbf{x}_i) = \frac{e^{\mu_i}\mu_i^y}{y!}, y = 0, 1, 2, ..., \tag{7.3}$$

and mean parameter

$$\mu_i = E[Y_i|\mathbf{x}_i] = \exp(\mathbf{f}(\mathbf{x}_i)^T\boldsymbol{\beta}), \tag{7.4}$$

in GAM model form. The functions $f_j(\cdot)$ are linear in a GLM model, in which case, the mean parameter definition is simplified to

$$\mu_i = E[Y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i^T\boldsymbol{\beta}). \tag{7.5}$$

The predictor is linked to the mean μ_i by the log link function

$$\mathbf{x}_i^T\boldsymbol{\beta} = \log(\mu_i). \tag{7.6}$$

The loglinear nature of the Poisson regression GLM means that interpreting the parameter estimates is not as straightforward as it is for a linear regression model, although they have the same basic structure. For a linear regression model, the regression coefficient, β_j, is interpreted as the estimated expected change in the response variable associated with a one unit change in the jth predictor variable, x_j, keeping all others fixed. For a Poisson regression model, the exponent of the coefficient $\exp(\beta_j)$ is the estimated expected multiplicative change in the response variable with a one unit change in the jth predictor variable x_j (or $f_j(x_j)$), keeping all others fixed. This means that the absolute magnitude of the effect is dependent on the value of the response variable. Both trends in storm activity over time and the dependency of storm counts on sea surface temperatures are modelled to determine their relative importance for modelling annual Atlantic basin hurricane counts.

7.1.2 Logistic regression model

A logistic regression is a standard model for binary response variables and can conveniently be employed to model fractions of an event occurring [2]. It has

been used here to model fractions of annual CAT1-5 Atlantic basin hurricanes that make landfall in the United States and/or intensify into CAT3-5 Atlantic basin hurricanes. The logistic regression model treats these fractions as binomial distributed. If n represents the number of basin hurricanes and Y_a a Poisson variable with mean μ_a represents the annual count rate of landfalls (and Y_b the annual counts of non-landfalling hurricanes with mean μ_b so that $n = Y_a + Y_b$) then the distribution of Y_a can be defined by

$$f(Y_a = y | n, \pi) = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \pi^y (1-\pi)^{(n-y)}, \tag{7.7}$$

where $\pi = \mu_a/(\mu_a + \mu_b)$. Hence, the mean and variance of Y_a/n are π and $\pi(1 - \pi)$. Storm fractions have been regressed on the same three predictor variables used in the Poisson regression model (i.e. year, SST_{Atl}, and SST_{trop}). These predictors are linked to the mean parameter π by the *logit* link function, given as

$$\mathbf{x}_i^T \boldsymbol{\beta} = \log \left(\frac{\pi_i}{1 - \pi_i} \right). \tag{7.8}$$

Like the Poisson regression model, the coefficients in the logistic regression model do not indicate a directly proportional relationship between the response variable and the predictor variables. Instead, they represent the change in the logit for each unit change in the predictor, so the relationship needs to be interpreted in terms of the odds ratio, that is

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\mathbf{f}(\mathbf{x}_i)^T \boldsymbol{\beta}), \tag{7.9}$$

where

$$\pi(x) = \frac{\exp(\mathbf{f}(\mathbf{x}_i)^T \boldsymbol{\beta})}{1 + \exp(\mathbf{f}(\mathbf{x}_i)^T \boldsymbol{\beta})}. \tag{7.10}$$

Equation (7.9) implies that $\exp(\beta_j)$ is the estimated expected multiplicative change in the odds of a U.S. landfall strike with a one unit change in the jth predictor variable x_j (or $f_j(x_j)$), keeping all others fixed.

7.1.3 Model fitting

To allow for a nonlinear relationship between hurricane counts and fractions and the predictor variables, extended GAM versions of the Poisson and logistic regression models are fitted by estimating each function $f_j(\cdot)$ by means of either a quadratic polynomial or a *cubic spline* regression smoother [72, 73]. Both these regression smoothers offer more flexibility than does a simple or multiple linear regression in the sense that changes in the response variable may be dependent on the value of the predictor variable.

Polynomial regression smoother

A popular and straightforward method for modelling nonlinear relationships between response and predictor variables is to fit a regression model with a polynomial regression smoother [121]. In its most parsimonious form, a polynomial regression model consists of a polynomial function of order p. For example, a model with predictor variable x incorporating a pth-order polynomial in x has

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p. \tag{7.11}$$

Cubic spline regression smoother

Cubic splines are arbitrary smooth polynomial functions which can be viewed as a link between conventional polynomials in the GLM framework and more modern methods of nonparametric smoothing such as scatterplot smoothers [33, 174]. They consist of piecewise defined cubic polynomials fitted over different regions of x, so, unlike the polynomial regression smoother, the coefficients of the spline function can vary over different regions of x. The cubic polynomials are joined at $\xi_m, m = 1, \ldots, \mathcal{K}$ points on the domain of x called *knots*, where the function values and first $p - 1$ derivatives are equal. The more knots that are used the more flexible the cubic spline is. Cubic splines can be represented as a linear combination of their natural B-spline, or *basis spline*, functions,

expressed as

$$f(x) = \sum_{m=1}^{\mathcal{K}} \alpha_m B_m(x), x \in [a, b] \tag{7.12}$$

where α_m are coefficients, and $B_m(x)$ are the piecewise cubic B-spline basis functions which are non-zero over a range of at least five distinct knots in the arbitrary domain $[a, b]$.

Cubic splines are considered the explicit, unique minimiser over all functions of the regulated residual sum of squares of the estimated model, given by

$$\min_f \sum_{i=1}^{N} [Y_i - f(x_i)]^2 + \lambda \int_a^b [f''(t)]^2 dt, \tag{7.13}$$

where λ is a fixed constant, and $a \leq x_1 \leq \ldots \leq x_N \leq b$. The term on the left-hand side evaluates the distance between the data and the predictor while the term on the right-hand side penalises curvature in the function.

The relative advantages and disadvantages of polynomial and cubic spline functions in GLMs and GAMs are now discussed. Fitting the polynomial smoother is more straightforward than for cubic splines, and the flexibility of the model can be controlled to some extent by specifying the order p of the polynomial. The selection of p is typically made using significance tests or conventional model selection criteria [121] although in this case only a quadratic polynomial is employed, hence $p = 2$. The polynomial model is also more parsimonious than a cubic spline fit when $p \leq 4$. On the other hand, polynomial regressions have undesirable non-local properties whereby a fitted value of the response variable at a given value of $x = x_i$ may depend strongly on other values which are some distance from x_i. Cubic splines do not suffer from this issue since they use local models [174]. Cubic splines also allow for a greater degree of flexibility than polynomial functions with fewer limitations on the functional form.

7.1.4 Interaction terms

The inclusion of more than one predictor variable in an regression analysis can have important ramifications for the interpretation of the fitted model. The relationship between a response variable y and a predictor variable x_1 can vary, depending on the value of a second predictor variable x_2. Collinearity between x_1 and x_2 implies that the causal relationship between the y and x_1 is moderated by x_2. Moderated relationships in regression models are sometimes referred to as "interaction effects" [36, 44]. The presence of interaction effects effectively means that the combined effect of two or more predictor variables on the response variable is not additive. To accommodate the impact of interaction between predictor variables in the hurricane regression models, a two-way interaction term is also introduced into the model. The interaction term takes the form of $\beta x_1 x_2$, so that the linear coefficient of the predictor variable x_1 changes smoothly according to the other predictor variable x_2. For example, a Poisson model with two predictors and a two-way interaction term can be expressed as

$$\mu_i = E[Y_i|\mathbf{x}_i] = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2). \qquad (7.14)$$

The two sea surface temperature predictor variables, tropical Atlantic sea surface temperature anomalies and global sea surface temperature anomalies, are positively correlated ($r = 0.78$) suggesting that there are interaction effects in the models which include them both. Burnham and Anderson [27] suggest not dropping a predictor unless the correlation coefficient is extremely high, and state that $|r| = 0.95$ is a reasonable cutoff value for dropping a covariate. The added two-way interaction term should account for any collinearity between the predictor variables.

7.2 Model selection

Model selection is performed here by means of a stepwise approach employing the Akaike Information Criterion (AIC) and the Schwarz Bayesian Criterion

(SBC) to rank models whilst managing the trade-off between model complexity and goodness of fit. These criteria are expressed as

$$AIC = -2\log(L) + 2d, \tag{7.15}$$

and

$$SBC = -2\log(L) + d\log(N), \tag{7.16}$$

where L denotes the maximum likelihood value for the estimated model, d is the number of parameters or degrees of freedom of the model, and N is the number of observations. The SBC penalises a model more where there is a larger number of parameters than the AIC when N is small, which is the case here since the modelled period is 1966-2012 (i.e. $N = 47$). The AIC with correction (AIC_c) is therefore used here which, like the SBC, penalises models with extra parameters [27]. The AIC_c is defined as

$$AIC_c = AIC + \frac{2d(d+1)}{N-d-1}. \tag{7.17}$$

Since AIC and SBC only measure the relative quality of the model fit (e.g. Hipel [75]), the model performance has been evaluated by analysing the model residuals, which should be i.i.d. Gaussian distributed if the model is a good fit. The normalized randomized quantile residuals [42] have been examined to assess the distribution of the residuals by computing the first four moments of their distribution (mean, variance, and coefficients of skewness and kurtosis), and their Filliben correlation coefficient [56]. In addition, quantile-quantile plots and worm plots have also been analysed as a visual reference of model goodness of fit (see appendix C).

7.3 Inference for regression coefficients

The standard approach to estimating confidence intervals (CI) for regression model parameters is to invert a two-sided significance test of $H_0 : \beta_j = \beta_0$

for the entire parameter space $\beta_j \in S(\mathscr{P})$. A $100(1-\alpha)\%$ confidence interval contains the set of β_0 values for which the test has p-value $\geq \alpha$. Three candidate test statistics for constructing confidence intervals for β_j are the inverted Wald, score, and likelihood-ratio tests. Each can be expressed in terms of the log-likelihood function $L(\beta_j)$ where the maximum likelihood estimate is $\hat{\beta}_j$. The Wald test statistic uses the *Fisher information* $\iota(\beta_j) = -E[\partial^2 L(\beta_j)/\partial \beta_j^2]$ and can be expressed as

$$[(\hat{\beta}_j - \beta_0)/SE(\hat{\beta}_j)]^2 = (\hat{\beta}_j - \beta_0)^2 \iota(\hat{\beta}_j), \tag{7.18}$$

where $\iota(\hat{\beta}_j)$ denotes $\iota(\beta_j)$ evaluated at $\hat{\beta}_j$. $SE(\hat{\beta}_j)$ is computed from the variance-covariance matrix for the regression coefficients, given as

$$\begin{aligned} SE(\hat{\beta}_j) &= \sqrt{Var(\hat{\beta}_j)} \\ &= \sqrt{\sigma^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1}}, \end{aligned} \tag{7.19}$$

where σ is the residual variance term and $\mathbf{X}$ is the $n \times k$ matrix of elements x_{ij}. The corresponding $100(1-\alpha)\%$ Wald CI for β_j is defined

$$\hat{\beta}_j \pm z_{\alpha/2} SE(\hat{\beta}_j), \tag{7.20}$$

where $z_{\alpha/2}$ denotes the $1-\alpha$ quantile of the standard normal distribution. The Wald test is an asymptotic approximation of the likelihood-ratio (LR) test using the Gaussian distribution [2]. The LR test statistic for a parameter from a single predictor regression model is given by

$$-2[L(\beta_0) - L(\hat{\beta})]. \tag{7.21}$$

The score test statistic is

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial \beta_0]^2}{-E[\partial^2 L(\beta)/\partial \beta_0^2]}, \tag{7.22}$$

where $u(\beta)$ is the score function, and the partial derivatives are evaluated at β_0.

The Wald confidence interval is most commonly employed in statistical software for its ease of use, but inversions of the score and likelihood-ratio tests provide better probability coverage where the sample size is small, or the parameter estimate is close to the lower and upper bounds of the parameter space [1]. The latter issue is usually irrelevant for inference of regression coefficients but in the case of hurricane counts and fractions the former issue is potentially problematic due to the limited availability of reliable data. The inverted score and likelihood-ratio tests provide probability coverage that is usually close to nominal levels. Interpreting the Wald interval is also difficult because of its dependency on the scale of measurement [1]. For example, when constructing a confidence interval for the coefficients of the Poisson regression, a Wald CI for e^{β_j} is not transformable from the exponentiated values of the Wald CI for β_j.

Inversion of the score and likelihood-ratio test statistics, which are both functions of the log-likelihood function $L(\beta_j)$, are difficult to perform, however, where the likelihood function is not an explicit function of the model parameters as is the case with regression coefficients [1]. Instead, a computational algorithm can be used to perform the test inversion by exhibiting all values of β_0 for which the p-value exceeds α in the test $H_0 : \beta_j = \beta_0$. The aim is to compute the confidence interval (CI) where

$$CI(\beta_j) = \{\beta_j \in S(\mathscr{P}) : z(\beta_j) \leq z_{\alpha/2}\} = [\hat{\beta}_j^L, \hat{\beta}_j^U], \qquad (7.23)$$

where the bounds of the interval $\hat{\beta}_j^L$ and $\hat{\beta}_j^U$ are the two roots of the test-inversion equation

$$z(\beta) = (\beta_j - \hat{\beta}_j)/SE(\hat{\beta}_j) = z_{\alpha/2}. \qquad (7.24)$$

An algorithm based on the 'sliding quadratic' root-finding algorithm devised by Lang [110] for computing the score and inverted likelihood-ratio test confidence intervals for contingency tables is appropriated for determining equivalent confidence intervals for regression coefficients. This algorithm is efficient and robust so that, when the root of the test-inversion equation is close to, or equal to, the boundary of $S(\mathscr{P})$, it will not fail unlike the bi-section and Newton-Raphson

methods [110]. The algorithm for finding the upper root $\hat{\beta}_j^U$ is detailed by Algorithm 2. The lower root $\hat{\beta}_j^L$ is computed in the same way as $\hat{\beta}_j^U$. The only essential differences are the initial values, given as $\beta_j^{(0)} = \hat{\beta} - \epsilon$ and $\beta_j^{(1)} = \hat{\beta} - 2\epsilon$, and the choice of linear equation root being the smallest root rather than the largest root.

7.4 Overdispersion

Both the Poisson and logistic regression models are potentially limited by the constraint that the mean of the response variable determines its variance. This constraint can result in a phenomenon called *overdispersion* [2, 180]. Overdispersion is encountered when the variance of observed count data is often larger than would be expected if the response variable were Poisson or binomially distributed. Such scenarios can arise where there is clustering, or *heterogeneity*, in a population which is not accounted for in the parameters of the Poisson and logistic regression models. For example, regressing hurricane counts on year alone may exclude dependence of the response variable on other important predictors each having a different mean for the response variable.

7.4.1 Poisson regression models

The assumption of independence, or *equidispersion*, of the observations is made for the Poisson regression model; where this assumption is baseless then the goodness of fit of the Poisson model may be compromised by overdispersion [29, 38]. To test for overdispersion of hurricane counts, the auxiliary *ordinary least squares* (OLS) regression approach described by Cameron and Trivedi [29] has been employed here. Once a Poisson regression model has been fitted using the standard GLM method (see Section 7.1.1), the predicted values $\hat{\mu}_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$ are used to perform an additional OLS regression, given by

$$z_i = \frac{(Y_i - \hat{\mu}_i)^2 - Y_i}{\hat{\mu}_i} = \alpha \hat{\mu}_i + \epsilon_i, \tag{7.25}$$

Algorithm 2: 'Sliding linear' root-finding algorithm

1: $\hat{\beta}_j$, $ase(\hat{\beta}_j)$ // *coefficient estimate and asymptotic standard error*

Ensure: $\beta_j^{min} \leq \beta_j^{opt} \leq \beta_j^{max}$ // *set bounds on β*

2: $\beta_j^{(0)} = \hat{\beta}_j + \epsilon$

3: $\beta_j^{(1)} = \hat{\beta}_j + 2\epsilon$

4: $z(\beta_j^{(0)}) = \frac{\beta_j^{(0)} - \hat{\beta}}{ase(\hat{\beta})}$

5: $z(\beta_j^{(1)}) = \frac{\beta_j^{(1)} - \hat{\beta}}{ase(\hat{\beta})}$

6: $a = \frac{(z(\beta_j^{(1)}) - z(\beta_j^{(0)}))}{(\beta_j^{(1)} - \beta_j^{(0)})}$

7: $b = z(\beta_j^{(0)}) - a\beta_j^{(0)}$ // *coefficients of linear equation that passes through points $(\beta_j^{(0)}, z(\beta_j^{(0)}))$ and $(\beta_j^{(1)}, z(\beta_j^{(1)}))$*

8: $c_1 = min\{z(\beta_j^{(1)}) + 0.5, z_{\alpha/2}\}$ // *$1 - \alpha$ is level of confidence*

9: $\beta_j^{(2)} = \frac{c_1 - b}{a}$

10: $z(\beta_j^{(2)}) = \frac{\beta_j^{(2)} - \hat{\beta}}{ase(\hat{\beta})}$

11: **for** $i = 2$ **to** N **do**

12:      $a = \frac{(z(\beta_j^{(i+1)}) - z(\beta_j^{(i)}))}{(\beta_j^{(i+1)} - \beta_j^{(i)})}$

13:      $b = z(\beta_j^{(i)}) - a\beta_j^{(i)}$

14:      $c_i = min\{z(\beta_j^{(i+1)}) + 0.5, z_{\alpha/2}\}$

15:      $\beta_j^{(i+1)} = \frac{c_i - b}{a}$

16:      $z(\beta_j^{(i+1)}) = \frac{\beta_j^{(i+1)} - \hat{\beta}}{ase(\hat{\beta})}$

17:      **if** $z(\beta_j^{(i+1)}) - z_{\alpha/2} \geq tol$ **then**

18:          $\hat{\beta}_j^U = \beta_j^{(i+1)}$

19:          *break* // breaks the loop

20:      **end if**

21: **end for**

22: **return** $\hat{\beta}_j^U$

---

where ϵ_i is an error term. The constant α is then tested under the null hypothesis that $\alpha = 0$.

7.4.2 Logistic regression models

The logistic regression, like the Poisson regression, is susceptible to overdispersion due to heterogeneity but also due to positive correlation in the underlying Bernoulli trials which determine the response [128, 180]. One commonly used method using a parametric model to test overdispersion in binomial data is to use a *beta-binomial* model. Although the beta-binomial distribution is not a member of the exponential family, it generalises to regression models in a straightforward manner. Let p_i and n_i be the parameters of a binomial distribution. Under the assumptions of the beta-binomial model, $Y_i \sim \mathcal{B}(n_i, p_i)$, and p_i is beta distributed with parameters (α_i, β_i). In addition, let $E(p_i) = \pi_i = \alpha_i/(\alpha_i + \beta_i)$ satisfy a logistic relationship, as in Eqn. (7.7), with predictors $\mathbf{x}_i$. In that case, Y_i follows a beta-binomial distribution where

$$E(Y_i) = m_i \pi_i, \qquad (7.26)$$

and

$$Var(Y_i) = \frac{m_i \alpha_i \beta_i [1 + (n_i - 1)(\alpha \beta_i + 1)^{-1}]}{(\alpha_i + \beta_i)^2} \qquad (7.27)$$
$$\equiv m_i \pi_i (1 - \pi_i)[1 + (n_i - 1)\psi_i],$$

where $\psi_i = (\alpha_i + \beta_i + 1)^{-1}$ is the scale parameter. A positive value of ψ_i indicates overdispersion.

7.5 Results of GLM and GAM modelling of Atlantic basin tropical cyclones

7.5.1 Poisson regressions for 1966-2012 storm counts

The results of modelling of annual counts of Atlantic basin named storms, CAT1-5 and CAT3-5 hurricane counts and CAT1-5 U.S. landfall hurricane counts over the period 1966-2012 using a Poisson regression model are now discussed. The response count data have been sourced from the HURDAT database[3], and time series of all storm categories from 1966-2012 are shown in Fig. 7.1. Each response variable is modelled so that the logarithm of the count rate μ is a function of a given combination of the three predictor variables (i.e. year, SST_{Atl}, and SST_{trop}). Three GLM or GAM versions of the regression model are considered, one to model linear dependence of the response variable on the predictors, given by

$$E[Y_i] = \mu_i = \exp(constant), \tag{7.28}$$

and two to model non-linear dependence with polynomial and cubic spline functions, given by

$$E[Y_i] = \mu_i = \exp(f_{polyn}(i)), \tag{7.29}$$

and

$$E[Y_i] = \mu_i = \exp(f_{spline}(i)). \tag{7.30}$$

The parameter estimates and measures of model fit for the "best-fit" models of annual hurricane count rates ranked according to the AIC model selection criterion are listed in Table 7.1. SST_{Atl} and SST_{trop} are both retained as significant predictors for all four storm categories, and in all cases the modelled

---

[3]http://www.aoml.noaa.gov/hrd/data_sub/re_anal.html

relation between the logarithm of the count rate μ and these two predictors is linear. There is a clear positive relationship between SST_{Atl} and all storm count rates, and a negative relationship between SST_{trop} and all storm count rates.

The score CIs for the regression coefficients of both the SST_{Atl} and SST_{trop} predictors indicate that the signs of these relationships are reliable at the 95% confidence level except for the relationship between SST_{trop} and CAT1-5 U.S. landfalls which has the least precise CI. This result is consistent with several studies [100, 204, 208] which suggest that sea surface temperatures in the Atlantic basin relative to global tropical SSTs are an important factor in modulating hurricane activity. There is also a linear relation between the logarithm of the rate μ of Atlantic basin named storms and year, although this is not significant. Year has been found to be significant when acting as sole predictor but its effect on all storm count response variables is minimal compared to the other two predictors.

None of the dependencies of the storm count rates are modelled via the regression smoothers in these best-fit models. The AIC_c values for the models which do include nonlinear dependencies (i.e. quadratic polynomial and cubic spline fits) are comparable to those of the linear models, however, suggesting that they are penalised more for the higher degrees of freedom. This marginality in the model selection may be reflective of the short duration of the modelled time period. The SBC model selection criterion penalises models more for increased degrees of freedom, but in this case the relative ranking of the models is the same given the use of AIC_c. The inclusion of interaction terms is evidently detrimental to the model fit as they increase the degrees of freedom, and are shown to be insignificant in all model fits.

Multidecadal variability is more pronounced in named storms and hurricane counts than the U.S. landfall counts (see Fig. 7.1). The counts are smaller for the U.S. landfall time series making trend detection or cycle detection more difficult [34, 207]. The (un)detectability of trends in storm counts over time is reflected in the p-values where storm counts are regressed on "year" only.

For example, year is retained as an important predictor for named storms (p-value= 4.3×10^{-5}), but not so for CAT1-5 U.S. landfalls (p-value= 0.26).

The model fit diagnostics shown in Table 7.1, and plotted in appendix C, indicate that all four best-fit models are adequately able to reproduce the annual counts of all four storm categories over the period 1966-2012. The normalized (randomised) quantile residuals and worm plots for the basin named storm model exhibit the best approximation of a normal distribution which is perhaps reflective of the largest degree of multidecadal variability in that category.

Finally, tests for overdispersion using an auxiliary OLS regression explained in Section 7.4 show that the null hypothesis of equidispersion is rejected only for the best-fit Poisson regression model for CAT1-5 U.S. landfall counts (p-value= 0.03). Hence, only this model is subject to overdispersion, but, given that the result is not highly significant, it is not necessarily a poor fit. Overdispersion of the CAT1-5 U.S. landfall counts may be indicative of statistical dependence in the data [2], but an investigation of this phenomenon is suggested as a subject for further research.

7.5.2 Logistic regressions for 1966-2012 storm fractions

A logistic regression has been used to model CAT3-5 Atlantic basin hurricanes and CAT1-5 U.S. Landfall hurricanes as a fraction of the total basin hurricane annual counts for the period 1966-2012. Fractions of storm counts for these two categories are regressed on year, SST_{Atl}, and SST_{trop}. Each response variable is modelled so that the logarithm of the odds ratio is a function of a given combination of the three predictor variables. Three GLM or GAM versions of the regression model are considered, one to model linear dependence of the response variable on the predictors, given by

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(constant), \tag{7.31}$$

and two to model non-linear dependence with polynomial and cubic spline func-

Table 7.1: Poisson regression models of Atlantic basin storms 1966-2012

	Basin named storms	CAT1-5 basin hurricanes	CAT3-5 basin hurricanes	CAT1-5 U.S. landfalls
Intercept	2.01***	1.67***	0.61***	0.23
	(0.13)	(0.07)	(0.12)	(0.15)
Year	0.01*	-	-	-
Standard error	(0.01)	-	-	-
Score 95% CI	(0,0.02)	-	-	-
SST_{Atl}	0.97***	1.53***	1.95***	1.50**
Standard error	(0.22)	(0.29)	(0.46)	(0.59)
Score 95% CI	(0.52,1.42)	(0.96,2.10)	(1.03,2.86)	(0.32,2.68)
SST_{trop}	-1.37***	-1.74***	-1.80***	-1.80*
Standard error	(0.41)	(0.51)	(0.83)	(1.04)
Score 95% CI	(-1.78,-0.95)	(-2.76,-0.72)	(-3.47,-0.13)	(-3.89,0.30)
Deg. of Freedom for the fit	4	3	3	3
Mean (residuals)	0.03	0.04	0.02	-0.07
Variance (residuals)	0.75	0.54	0.82	1.26
Skewness (residuals)	0.37	0.43	-0.05	-0.12
Kurtosis (residuals)	3.04	3.0	3.76	3.06
Filliben (residuals)	0.99	0.99	0.98	0.99
AIC_c	244.54	201.10	161.02	152.47
SBC	250.99	206.70	166.0	157.46

$*p < 0.10; **p < 0.05; ***p < 0.01$ (two-tailed).

Standard errors are given in parentheses. Score CIs are determined using a 'sliding linear' root-finding algorithm. The plot is produced using R statistical software (R Development Core Team, 2008) using the freely available Generalized Additive Models for Location Scale and Shape (GAMLSS) package [190].

tions, given by

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(f_{polyn}(i)), \tag{7.32}$$

and

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(f_{spline}(i)). \tag{7.33}$$

Interpreting the effect of the predictors is more complicated with a logistic regression since the coefficients represent a change in the logit function for each
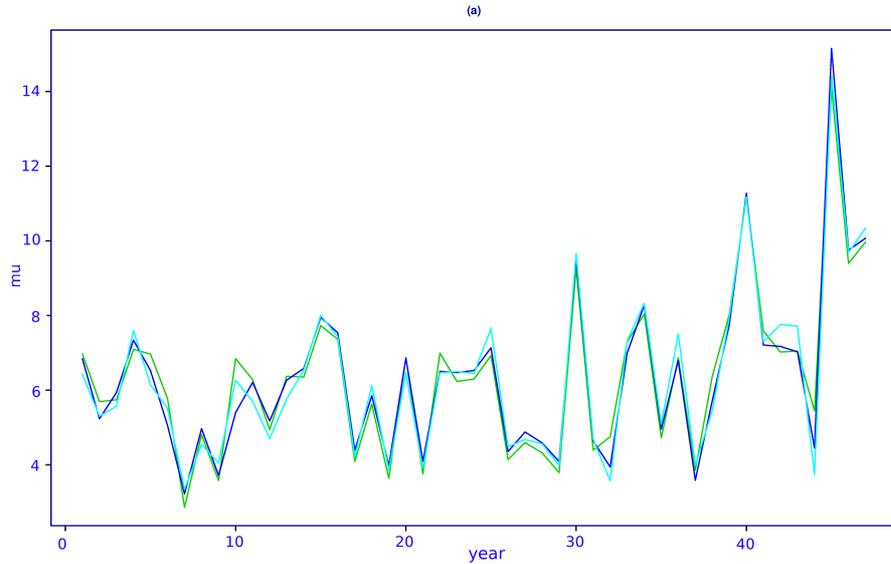
Figure 7.2: Modelling Atlantic basin CAT1-5 basin hurricanes: fitted values of the rate of CAT1-5 Atlantic basin hurricane annual counts μ regressed on SST_{Atl} and SST_{trop} from 1966-2012 with linear (green line), quadratic polynomial (dark blue), and cubic spline (light blue) Poisson regression models. The linear fit corresponds to the best-fit model in the second column of Table 7.1 with $AIC_c = 201.1$.

unit change in the predictor, not the response variable itself (i.e. the fraction of CAT1-5 U.S. landfalls or CAT3-5 basin hurricanes). The parameter estimates and measures of model fit for the best-fit models of annual hurricane count fractions ranked according to the AIC model selection criterion are listed in Table 7.2.

Unlike the Poisson regression models of storm counts, there are no clear important predictors for both CAT3-5 Atlantic basin hurricane fractions and CAT1-5 U.S. landfall fractions. Year is the only significant predictor retained by the best-fit model for the former category where the modelled relation between $log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ and the predictor is positive and linear. The high precision of the score CI for the regression coefficient shows that the positive relationship is reliable at the 95% confidence level. The best-fit model for the latter category only retains SST_{Atl} as a predictor, but it is not significant. In that case,

the modelled relation between $\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ and the predictor is negative and linear, but the score CI for the regression coefficient straddles both negative and positive values reflecting the lack of significance of SST_{Atl} as a predictor.

Substantially high statistical confidence in the predictor of the CAT3-5 Atlantic basin hurricane fraction model compared to the CAT1-5 U.S. landfall fraction model reflects the higher counts and larger multidecadal variability of CAT3-5 Atlantic basin hurricanes during the years 1966-2012 [34, 202]. The results of this logistic regression modelling exercise are somewhat divergent from those of Villarini et al. [208] who found that SST_{trop} is an important predictor for CAT1-5 U.S. landfalls. The period of their modelling study was 18782008, however, providing a much larger sample size, although the reliability of the data over the earlier part of that period is questionable [106].

Like the best-fit Poisson regression models in the previous section, none of the modelled dependencies of the storm count rates are nonlinear in the best-fit logistic regression models, but the AIC_c and SBC values are similar for all three of the linear, quadratic polynomial and cubic spline fits. The linear models are penalised less by AIC_c and SBC for having less degrees of freedom so that they tend to be ranked as the best-fit models. As before, the short duration of the modelled time period (i.e. small storm count sample size $N = 47$) results in similarity of the values of AIC_c and SBC.

Inclusion of interaction terms has a similar impact on model selection as was the case for the Poisson regression models. The additional model parameters increase the degrees of freedom, but do not result in relative improvement of model fit. Moreover, they are again shown to be insignificant in all model fits. Model fit diagnostics are shown in Table 7.2, and plotted in appendix C, and indicate that both best-fit models are adequately able to reproduce the annual counts of both storm count fraction categories over the period 1966-2012.

Testing for overdispersion reveals that the beta-binomial model (see Section 7.4) is a slightly better fit than the CAT1-5 U.S. landfall fraction logistic regression model. The AIC_c values are 144.45 and 144.87, respectively. This result

Table 7.2: Logistic regression models of Atlantic basin storms 1966-2012

	CAT3-5 basin hurricanes fractions	CAT1-5 U.S. landfall fractions†
Intercept	-0.97***	-1.16***
	(0.26)	(0.18)
Year	0.02**	-
Standard error	(0.01)	-
Score-test 95% CI	(0,0.04)	-
SST_{Atl}	-	-0.14
Standard error	-	(0.47)
Score-test 95% CI		(-1.07,0.79)
SST_{trop}	-	-
Standard error	-	-
Score-test 95% CI	-	-
Deg. of Freedom for the fit	2	3
Mean (residuals)	-0.04	0.05
Variance (residuals)	0.59	0.78
Skewness (residuals)	0.04	0.60
Kurtosis (residuals)	2.43	3.73
Filliben (residuals)	0.99	0.98
AIC	133.89	144.45
SBC	137.31	149.45

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$ (two-tailed).

†The CAT1-5 U.S. landfall best-fit model is the beta-binomial version.

Standard errors are given in parentheses. Score-test CIs are determined using a 'sliding linear' root-finding algorithm. The plot is produced using R statistical software (R Development Core Team, 2008) using the freely available Generalized Additive Models for Location Scale and Shape (GAMLSS) package [190].

indicates that there is overdispersion present in the CAT1-5 U.S. landfall fraction data, but its effect on the validity of the logistic regression model would be minimal. The parameter values for the beta-binomial fit are shown in Table 7.2

rather than the logistic regression model. The equivalent test for overdispersion of the CAT3-5 Atlantic basin hurricane fraction logistic regression model shows no overdispersion is present.

7.6 Conclusions

GLM and GAM regression models have been employed to describe annual Atlantic basin tropical cyclone counts and fractions over the period 1966-2012 in this chapter. Four different categories of annual counts have been modelled (i.e. basin named storms, CAT1-5 and CAT3-5 basin hurricanes, and CAT1-5 U.S. landfalls using a Poisson regression while fractions of CAT1-5 Atlantic basin hurricanes which make landfall over the U.S., or develop into CAT3-5 hurricanes have been modelled using a logistic regression. There are three predictor variables: year, tropical Atlantic SST anomalies (SST_{atl}), and global tropical SST anomalies (SST_{trop}) have been included in the models. The latter two have often been cited in the literature as playing an important role in the modulation of Atlantic basin tropical cyclone activity. A novel combination of GLM and GAM techniques which includes regression smoothers to model both linear and nonlinear dependencies of hurricane response variables on the three predictor variables has been incorporated into the models. In addition, collinearity between the predictor variables has been accounted for by the inclusion of interaction terms in the models.

An innovative approach to constructing inverted score and likelihood-ratio test confidence intervals for regression coefficients using a 'sliding linear' root-finding algorithm has been proposed and executed. These CIs provide better probability coverage that is closer to the nominal level than the conventional Wald CI where the sample sizes are small, but are difficult to construct because likelihood function is not an explicit function of the model parameters. The 'sliding linear' root-finding algorithm is an efficient and robust method which be used for finding the lower and upper bounds of the inverted score and likelihood-

ratio tests.

The results of the Poisson regression model fits has revealed that SST_{atl} and SST_{trop} are important predictors for explaining annual TC counts for all four categories. The signs of the regression coefficients of these two predictors have been found to be in agreement with those estimated by Villarini et al. [207, 208] albeit over a different modelled time period. The positive relationship between tropical cyclone counts and SST_{atl} is consistent with scientific understanding of the physical influences on TC formation while the negative relationship between tropical cyclone counts and SST_{trop} supports recent findings on relative sea surface temperatures [100, 204]. Year is retained as a significant predictor of basin named storms, but its effect is not as strong as SST_{atl} and SST_{trop}. Although the relative importance of the three predictor variables has been assessed, it is not entirely clear whether the dependency of Atlantic basin tropical cyclone counts and fractions on the predictors over the period 1966-2012 is linear or nonlinear. Given the relatively short duration of the modelled period, there is a relatively small variation in the values of AIC_c and SBC for most model fits. The relative parsimony of the linear models (as low as 3 degrees of freedom) compared to the models which include quadratic polynomial and cubic spline regression smoothers means that they tend to be penalised less, yet are able to reproduce the variability exhibited by the count data over the last 47 years reasonably well.

The modelling of CAT3-5 basin hurricane and CAT1-5 U.S. landfall fractions using a logistic regression has resulted in less significant fits of the three predictor variables. This is likely to be attributable to the fact these are sub-categories of total Atlantic basin counts, and therefore contain lower counts (i.e. less data), resulting in less power in significance tests. Year is retained as a statistically significant predictor of fractions of CAT1-5 basin hurricanes developing into CAT3-5 basin hurricanes while the model which includes SST_{atl} as the sole predictor for CAT1-5 U.S. landfalls is selected as the best fit by AIC_c and SBC. The relationship between CAT1-5 U.S. landfalls and SST_{atl} appears

to be insignificant, however. The lack of strong influence of the three predictors on CAT3-5 basin hurricane and CAT1-5 U.S. landfall fractions may be indicative of the small counts of these two categories over a relatively short modelled time period. Analysis of both the Poisson and logistic regression model fit diagnostics has shown that the residuals of all best-fit models exhibit reasonable approximations of a normal distribution while tests for overdispersion have demonstrated that there are no serious deficiencies in the models. The beta-binomial model for CAT1-5 U.S. landfall fractions has a slightly better fit than the logistic model according to AIC_c. The best-fit Poisson and logistic regression models in each hurricane category are employed to produce predictions of the 2013 seasonal hurricane counts and U.S. landfall fractions in Chapter 8.

The following novel contributions or innovations in this chapter are:

- development of GLM and GAM models of annual hurricane counts and fractions using Poisson and logistic regression models with polynomial functions and cubic splines employing a unique set of predictor variables;

- determination of score and inverted likelihood-ratio confidence intervals for regression model coefficients using an innovative 'sliding linear' root-finding algorithm;

- application and interpretation of tests for overdispersion of tropical cyclone count data for Poisson and logistic regression models.

Chapter 8

Forecasting the 2013 Atlantic basin Hurricane Season

The investigations presented in Chapters 3-7 have motivated and illustrated a proposed statistical framework for best-practice forecast construction, recalibration and evaluation in the setting of testbed dynamical and stochastic target systems, and real-world forecasting. Discussion focussed on more specific statistical aspects of applied hurricane forecasting in Chapter 5 in the context of several predictands and forecast lead times which are important to relevant decision-makers. The challenges posed by small-count data, and the slow collection of annual forecast evaluation data were highlighted, along with suggested approaches to address these challenges. A number of novel statistical forecast systems designed to exploit the limited information contained in a relatively short historical hurricane record were then introduced and evaluated. These forecast systems are simple to construct, and easy to implement, making them potentially useful as benchmark hurricane forecast models.

This final chapter brings together the forecast construction and evaluation methods featured within the statistical framework to be tested in a real-world hurricane forecasting case-study. A real-time outlook for the 2013 Atlantic basin hurricane season is presented, and then evaluated using the outcomes of

the 2013 hurricane season. The purpose of the out-of-sample forecast evaluation procedure in this chapter is to assess the potential skill of each statistical forecast system as a benchmark model, and compare their performance with other operational forecast systems.

The various statistical forecast systems introduced and evaluated in Chapter 5 are reviewed in Section 8.1, and then implemented in forecasting mode to construct predictions of total counts of named storms, CAT1-5 hurricanes, and CAT1-5 US landfalls occurring during the 2013 season. Subsequently, the performance of these forecasts is assessed with various forecast evaluation measures discussed in earlier chapters, and compared with equivalent predictions issued by other forecasting organisations (i.e. operational forecast centres, academic institutes, etc).

The novel (re)analysis of the National Hurricane Center's (NHC) 48 hour tropical cyclone (TC) genesis forecasts, presented in Chapter 6, is extended to the 2013 hurricane season in Sections 8.2 and 8.3. The assessment of the reliability of the NHC's 48-hour TC genesis forecasts for the 2012 hurricane season before and after recalibration is repeated for the 2013 season in Section 8.2. Next, the relationship between the reliability and "Time Until Event" (TUE) of the genesis forecasts is examined in Section 8.3. The TUE diagrams proposed in Chapter 6 as supplementary to reliability diagrams are presented to complete the interpretation of the reliability of the NHC's 2013 TC genesis forecasts. All of the analyses above and predictions for the 2013 hurricane season are new contributions.

The conjectures and methodologies for hurricane forecast construction, recalibration, and evaluation discussed in this thesis have been formalised since before the 2013 hurricane season commenced. Furthermore, statistical analyses of 2013 seasonal data and predictions produced from the forecast systems presented thus far have been made in real-time. No other analyses have been made.

8.1 Statistical forecast systems

Probabilistic predictions of three different TC count categories for the 2013 hurricane season are produced here: Atlantic basin named storms, CAT1-5 Atlantic basin hurricanes, and CAT1-5 U.S. landfalls. The skill of these three predictions is measured with the ignorance score defined with respect to the climatological reference forecast, and also compared with that of other operational forecasts. The 2013 hurricane season had officially come to a close at the time of writing (November 2013), and so the official counts for the season are available for the forecast evaluation procedure. An interesting point to note at this point is that the 2013 hurricane season ended with storm numbers well below the predictions of many operational forecasting centres[1]. The hurricane count data, used in the construction and evaluation of the 2013 predictions, is sourced from the HURDAT database[2].

8.1.1 Synoptic conditioning forecast system

Accurate pre-season predictions of the ENSO phase are widely considered to be key for constructing skilful statistical hurricane forecasts (see Gray [64] and Camargo et al. [28] and Section 5.1.3). The synoptic conditioning (SC) forecast system outlined in Section 5.2 is deployed here to produce probabilistic forecasts of the 2013 by conditioning historical storm counts on the ENSO phase [3] during the peak of each season. A Poisson process is used to model seasonal storm counts where the mean parameter λ is determined by the historical storm average during El Niño episodes and non-El Niño episodes. Hence, annual storm prediction Y_t is defined by

$$Y_t \sim \begin{cases} Pois(\lambda_A), & \text{if } \phi_t = A. \\ Pois(\lambda_B), & \text{if } \phi_t = B. \end{cases} \tag{8.1}$$

---

[1]http://hurricane.atmos.colostate.edu/forecasts/2013/nov2013/nov2013.pdf
[2]http://www.aoml.noaa.gov/hrd/data_sub/re_anal.html
[3]data sourced from http://www.cpc.ncep.noaa.gov/products/precip/CWlink/MJO/enso.shtml#current

where λ_A and λ_B are the mean storm counts in El Niño and non-El Niño years, respectively. Recall that the underlying assumption of the SC forecast system in Section 5.2 is that storm counts are distributed according to one of two probability distributions P_A or P_B dependent on the ENSO phase.
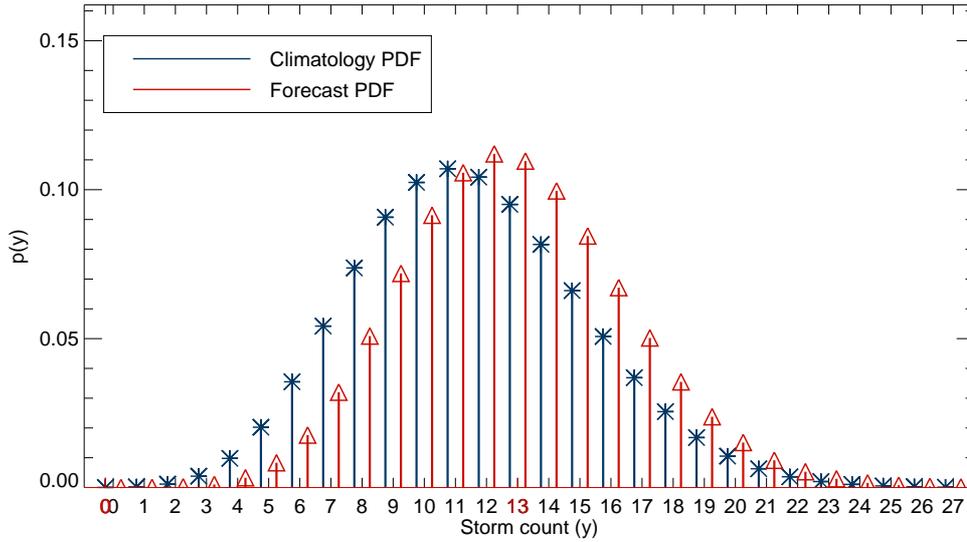
Figure 8.1: Synoptic conditioning forecast for 2013: SC forecast (red) and climatological forecast (blue) PDFs for Atlantic basin named storms in 2013. The synoptic conditioning technique utilises information on the annual August-October ENSO phases. There were 13 named storms in 2013 (axis label coloured red) which the SC forecast PDF has assigned larger probability mass to than the climatological PDF, and hence, has achieved superior skill $IGN = -0.28$.

The unconditional climatological forecast employed to measure the skill of the SC forecast system is, as in Section 5.2, defined by a weighted convex linear combination of the P_A and P_B, that is

$$P_{clim} = \alpha P_A + \beta P_B. \tag{8.2}$$

The values of α and β are updated to include the ONI data in 2013 (i.e. $\alpha = 0.33$ and $\beta = 0.67$). Figure 8.1 shows the forecast PDF for named storm counts for the 2013 season along with the climatological forecast PDF. The forecast

skill results of the 2013 forecasts of named storms, CAT1-5 basin hurricanes, and CAT1-5 U.S. landfalls are listed in Table 8.1.

8.1.2 Conditional analogue forecast system

A straightforward and computationally inexpensive empirical statistical forecasting scheme based on temporal analogue searching was presented in Section 5.4. Two analogue methods, *single analogue* and *series analogue*, have been designed to exploit self-repeating patterns of hurricane counts in the historical time series. In short, forecasts of future hurricane outcomes are conditioned by finding where analogues of current outcomes occurred in the past, and constructing distributions of the *images* (i.e. successive outcomes) of these analogues. Hence, the forecast system is referred to as *conditional analogue forecasting.*

The effectiveness of the conditional analogue (CA) forecast system was tested in a testbed hurricane system environment, and it was shown that both analogue methods demonstrated superior skill to both a Bayesian forecast model and a climatological model. The CA forecast system was calibrated with three training sets of size $N = 2^{12}$, but, when producing forecasts of 2013 hurricane outcomes, may be disadvantaged by the relatively short historical storm datasets with which to calibrate. Two model parameters need to be optimised to produce predictions in forecasting mode. Firstly, the parameter κ controlling the top-hat probability mass weights on each hurricane outcome y. Secondly, the blending parameter α determining the balance of weight between the kernel dressed forecast and climatological probability masses p and p_{clim}, respectively. As explained in Section 5.4.1, the parameter optimisation step is executed out-of-sample, and employs some cost function such as ignorance. Whereas parameter optimisation utilised training sets of forecast-outcome pairs in training mode in the previous chapter, it is executed here using k-fold cross-validation [69] (see 1.8). Given the limited size of the reliable historical hurricane record, the leave-one-out $k = N$ method is the most appropriate [155]. Hence, the

optimised parameters are given by

$$(\hat{\kappa}, \hat{\alpha}) := \arg\min_{\kappa, \alpha} -\frac{1}{N} \sum_{i \notin T}^{N} \alpha \times p_{\kappa, T}(Y_i) + (1 - \alpha) \times p_{clim}(Y_i), \qquad (8.3)$$

where T denotes the leave-one-out training set storm counts from 1966-2013. The CA single analogue forecast PDF showing the forecast PDF of named storm counts for the 2013 season is shown in Fig. 8.2 along with the climatological forecast PDF.
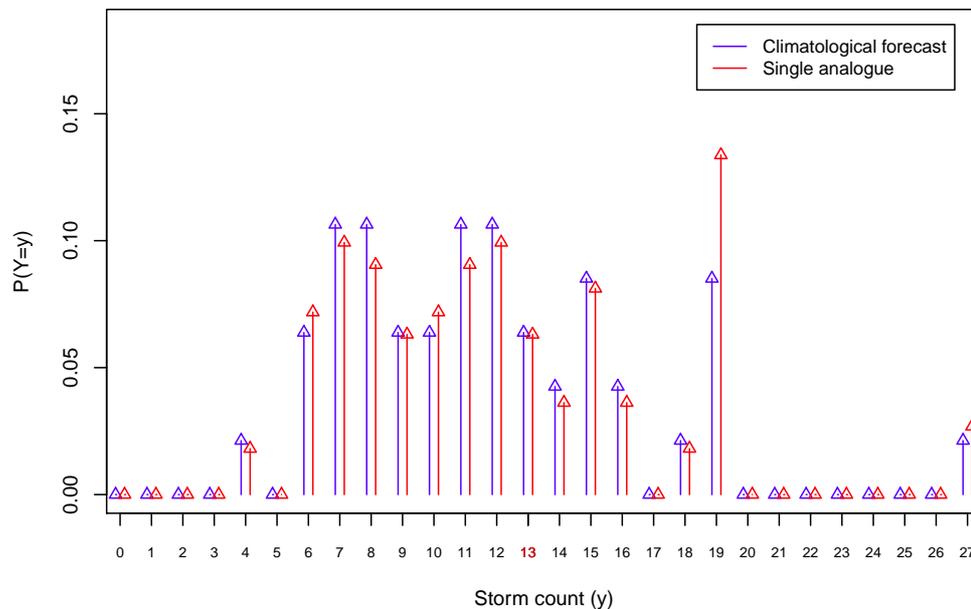
Figure 8.2: Conditional analogue forecast for 2013: single CA forecast (red) and climatological forecast (blue) PDFs for Atlantic basin named storms in 2013. There were 13 named storms in 2013 (axis label coloured red) which the CA forecast has assigned larger probability mass to than the climatological PDF, and hence, achieves superior skill $IGN = -0.40$.

The skill of the 2013 single and series analogue forecasts of basin named storms, CAT1-5 basin hurricanes, and CAT1-5 U.S. landfalls are listed in Table 8.1. The CA forecast system evidently demonstrates skill where forecasting basin named storms and CAT1-5 basin hurricanes with the series analogue

method performing slightly better than the single analogue method. The series analogue method has the advantage that, if series analogues are found elsewhere in the time series, they contain more information than a single analogue so that there is more information utilised in the forecast. If the CA forecast system shows skill, there may be some indication of periodic behaviour in hurricane activity over the Atlantic basin. Robust testing of this idea is not possible with the limited size of the hurricane record, however, and is beyond the scope of this thesis.

8.1.3 Hurricane regression models

The Poisson regression models fitted to annual tropical cyclone counts over the period 1966-2012 which were described in Chapter 7 are deployed here to provide predictions of basin named storms, CAT1-5 basin hurricanes, and CAT1-5 U.S. landfalls for the 2013 season. Model selection has been performed for these three categories using the corrected Akaike Information Criterion (AIC_c). Given the lower degrees of freedom of the linear GLM versions of the models compared to the GAM models which includes regression smoothers, and the limited duration of the reliable hurricane record, the linear models have been preferred in each case. The predictor variables which were found to be important were tropical Atlantic sea surface temperature (SST) anomalies SST_{Atl} and global tropical SST anomalies SST_{trop}. Year as a predictor variable has also been retained only for the model of Atlantic basin named storms. So, in that case, the logarithm of the mean count rate of annual basin named storms according to the best-fit linear model is given by

$$\mu_i = E[Y_i | year, SST_{Atl}, SST_{trop}] \qquad (8.4)$$
$$= exp(\beta_0 + year\beta_1 + SST_{Atl}\beta_2 + SST_{trop}\beta_3),$$

where the regression coefficients take the values $\beta_0 = 2.01$, $\beta_1 = 0.01$, $\beta_2 = 0.97$, and $\beta_3 = -1.37$. A probabilistic forecast for the 2013 season can be produced

using this fitted count rate parameter, given by

$$f(Y_i = k|\mathbf{x}_i) = \frac{e^{\mu_i}\mu_i^k}{k!}, k = 0, 1, 2, ..., \tag{8.5}$$

where k is the storm count. The Poisson regression forecast PDF showing the predictive distribution of named storm counts for the 2013 season is shown in Fig. 8.3 along with the climatological forecast PDF.
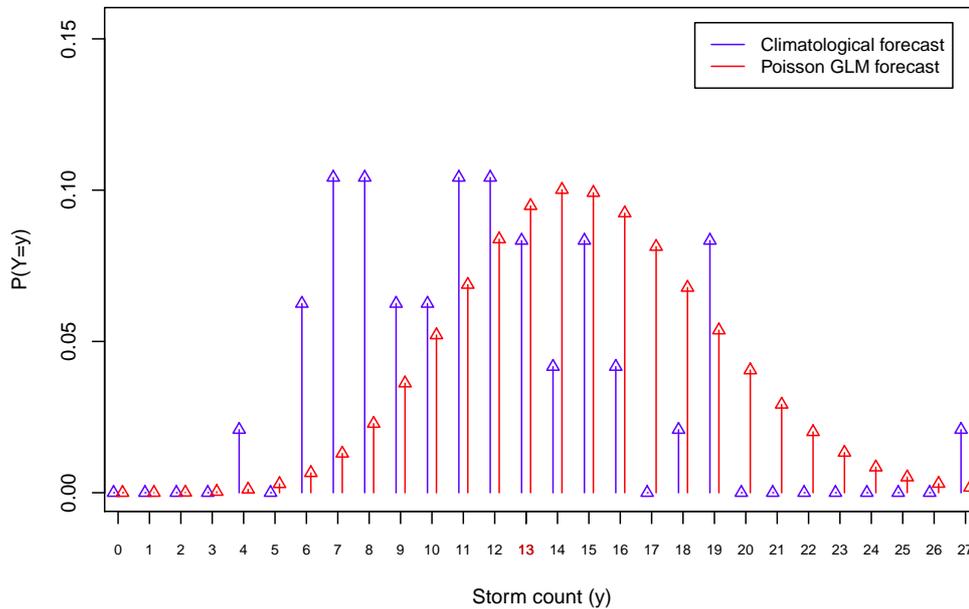
Figure 8.3: Poisson GLM forecast for 2013: Poisson GLM forecast (red) and climatological forecast (blue) PDFs for Atlantic basin named storms in 2013. The regression coefficients of the model are: $\beta_0 = 2.01$, $\beta_1 = 0.01$ (year), $\beta_2 = 0.97$ (SST_{Atl}), and $\beta_3 = -1.37$ (SST_{trop}). There were 13 named storms in 2013 (axis label coloured red) which the Poisson GLM forecast has assigned larger probability mass to than the climatological PDF, and hence, achieves superior skill $IGN = -0.16$.

The skill of the 2013 Poisson GLM forecasts of basin named storms, CAT1-5 basin hurricanes, and CAT1-5 U.S. landfalls are listed in Table 8.1. The Poisson GLM forecast is more skilful than the climatological forecast where forecasting basin named storms and CAT1-5 U.S. landfalls, but less so for the prediction of CAT1-5 basin hurricanes.

8.1.4   Review of skill of 2013 hurricane forecasts

One of the key requirements of robust forecast evaluation is sufficient forecast-outcome pair sample size. This requirement has been been highlighted throughout this thesis (see Chapters 1 through 5). Arriving at robust conclusions about the skill of the predictions of the 2013 hurricane season produced by the forecast systems introduced in this thesis is not possible. Nevertheless, it of interest to assess whether these forecast systems have at least some capability of producing an accurate hurricane forecast.

Table 8.1: 2013 hurricane forecast skill (IGN)

	Storm Category (number of storms in 2013)					
	Named storms (13)		CAT1-5 basin hurricanes (2)		CAT1-5 U.S. landfalls (0)	
Forecast system	Parameter values	IGN	Parameter values	IGN	Parameter values	IGN
SC	$\delta = \epsilon = 1$	-0.28	$\delta = \epsilon = 1$	0.84	$\delta = \epsilon = 1$	0.41
CA single analogue	$\kappa = 0.99,$ $\alpha = 0.15$	0.02	$\kappa = 0.99,$ $\alpha = 0.21$	-0.17	$\kappa = 0.93,$ $\alpha = 1.0$	0.62
CA series analogue	$\kappa = 0.99,$ $\alpha = 0.22$	-0.28	$\kappa = 0.99,$ $\alpha = 0.42$	-1.25	$\kappa = 0.99,$ $\alpha = 1.0$	0.95
Poisson GLM	$\beta_0 = 2.01,$ $\beta_1 = 0.01$ $\beta_2 = 0.97,$ $\beta_3 = -1.37$	-0.16	$\beta_0 = 1.67,$ $\beta_2 = 1.53,$ $\beta_3 = -1.74$	0.13	$\beta_0 = 0.23,$ $\beta_2 = 1.50,$ $\beta_3 = -1.80$	-0.28

The statistical forecasts of the 2013 Atlantic basin hurricane season are now compared with operational forecasts of the 2013 Atlantic basin hurricane season as a brief (but not statistically significant) assessment of their performance as benchmark models. At the beginning of the 2013 hurricane season, many operational forecast organisations anticipated an active season [65, 143, 198, 147] due to favourable environmental conditions such as anomalously warm tropical Atlantic SSTs, and an expected cool-neutral (non-El Niño) ENSO phase. The expectation of above-long term average activity persisted along with these conditions until the mid-season predictions were issued in August 2013. The season,

which closed on 30th November, proved to be one of quietest on record, however, owing to anomalous mid-tropospheric conditions, unfavourable for hurricane formation[4]. Such unexpected outcomes can lead to large statistical forecast error if the statistical relationship between hurricane activity and predictor variables is incomplete, or changing over time [43]. In theory, dynamical models should not be susceptible to the same kind of forecast error due to anomalous hurricane behaviour. The UK Met Office TC dynamical model forecasts is also included for comparison.

Table 8.2 below lists sets of predictions of the 2013 hurricane season from four high profile forecasting organisations along with predictions from the statistical forecast systems introduced in this thesis. All of the predictions are point forecasts (with uncertainty intervals where available), hence, they are compared with the medians of the probabilistic forecasts presented above. Comparison of the forecasts is intended to be cursory, and not an assessment of skill. Clearly, virtually all of the hurricane predictions are higher than the actual 2013 hurricane season outcomes. The median forecasts produced from the statistical forecast systems presented in this thesis have performed comparatively well, at least in the named storm category. The predictions of these forecast systems are all within 2 counts of the observed outcome of 13 named storms. In the other categories they are less accurate but are comparable with the operational forecasts. To reiterate, any quantitative evaluation of forecast skill would not be statistically significant here, a much larger set of out-of-sample evaluations is necessary to prove forecast skill (as discussed in Chapter 5). Still, the relatively accurate predictions produced from the thesis statistical forecast systems indicates that they may at least provide useful benchmark forecast models, particularly the single CA method.

---

[4]http://hurricane.atmos.colostate.edu/Forecasts/

Table 8.2: 2013 statistical hurricane forecasts (operational/thesis

		Storm Category (number of storms in 2013)		
Forecasting	Model	Named	CAT1-5	CAT1-5
centre	type	storms	basin hurricanes	U.S. landfalls
Colorado State University (CSU)	Statistical	18	8	-
National Oceanic and Atmospheric Administration (NOAA)	Statistical	13-19	6-9	-
Tropical Storm Risk (TSR)	Statistical	$14.8 \pm 2.9^*$	$6.9 \pm 1.8^*$	$1.8 \pm 1.5^*$
UK Met Office (UKMO)	Dynamical	$14 \pm 4^{**}$	$9 \pm 5^{**}$	-
Observed outcome		13	2	0
Thesis forecast system				
SC (median)	Statistical	12	7	1
CA single analogue (median)	Statistical	11	6	1
CA series analogue (median)	Statistical	13	7	2

*1 forecast error standard deviation **range represents 70% probability

8.2 NHC 2013 48-hour tropical cyclone genesis forecast reliability and recalibration

The reliability of the National Hurricane Center's 2012 48-hour TC genesis forecasts before and after recalibration was examined in Chapter 6. While the raw forecasts were found to be reliable at less extreme probability categories with some under-forecast bias at higher probability categories, out-of-sample recalibration using the 2011 forecasts as the training set degraded the performance of the forecasts, and increased the margin of under-forecasting. Leave-one-out cross-validation recalibration resulted in improved reliability of the 2012 forecasts.

The evaluation of the NHC 48-hour TC genesis forecasts is extended here
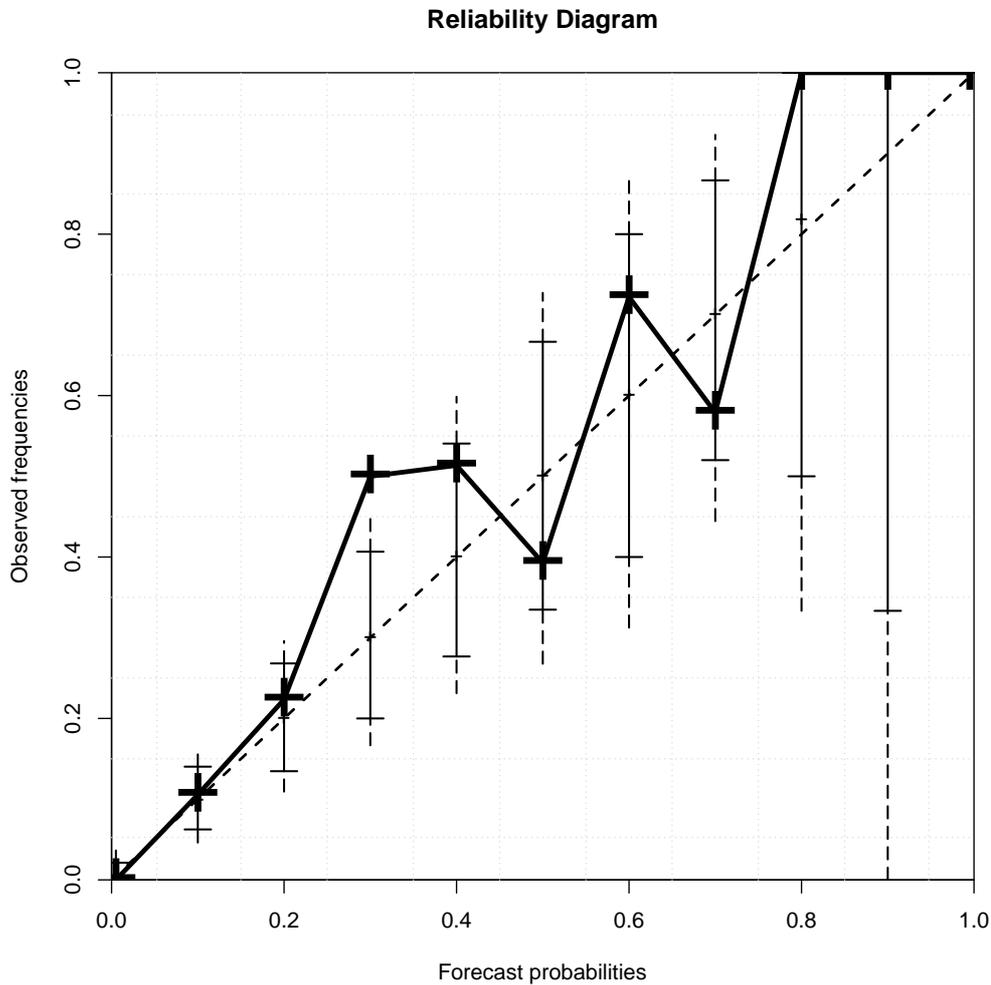
Figure 8.4: NHC 2013 TC forecast reliability: reliability diagram for the NHC's 2013 48-hr TC forecasts* with 5% - 95% (1% - 99% vertical dashed line) consistency bars. Forecast categories 80% and 90% have consistency bars with wide intervals and medians which lie off the diagonal because of small bin populations. The forecast probability bin boundaries (grey dotted lines) have been determined by taking the mid-points between each probability category value. **Sourced from NHC online Tropical Weather Outlooks.*

to the 2013 hurricane season. Again, the reliability of the forecasts is assessed pre- and post-recalibration. Recalibration is implemented using the 2012 forecasts as the training set, and the reliability of the recalibrated 2013 forecasts is compared with the recalibrated 2012 forecasts (evaluated in Section 6.3) to determine whether out-of-sample recalibration can be beneficial for short-term
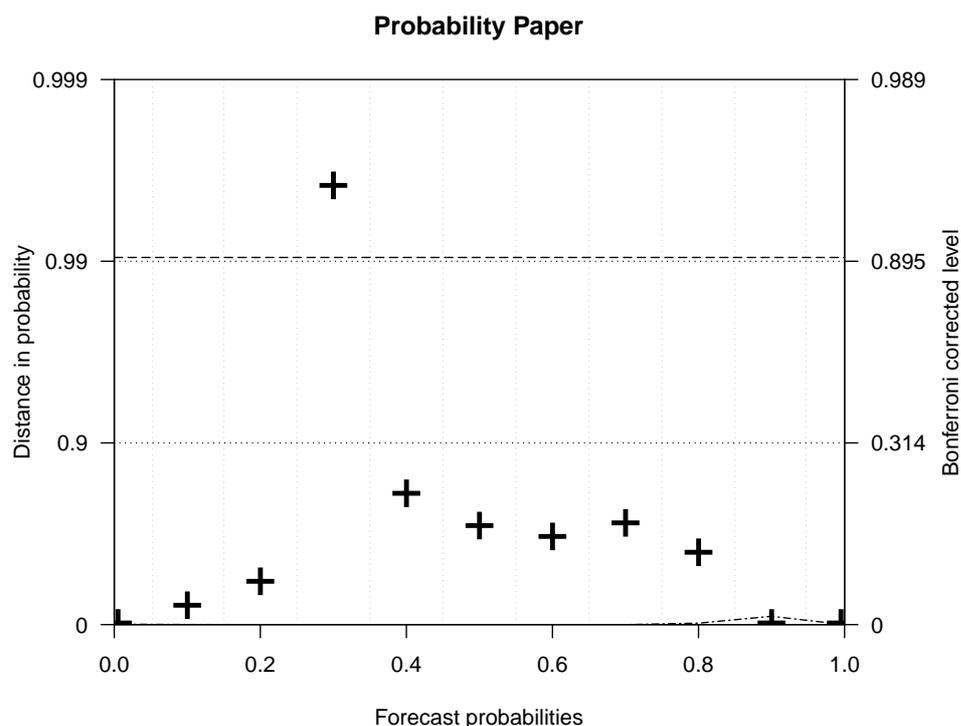
Figure 8.5: NHC 2013 TC forecast reliability: reliability diagram on probability paper for the NHC's 2013 48-hr TC forecasts*. The consistency bar median of forecast categories 0.8 and 0.9 lie off the diagonal because of small sample sizes. The dash–dotted line denotes the exact position of the diagonal. The right-hand axis indicates the equivalent Bonferroni corrected levels i.e. for a reliable forecast, all of the points (11 categories) would be expected to fall within the 0.99 probability distance band with an 89.5% chance. If it were not for the 0.3 probability category, the forecast could be considered reliable. In addition, the dashed lines indicate where the entire diagram would be expected to fall within with a 90% chance. The forecast probability bin boundaries (grey dotted lines) have been determined by taking the mid-points between each probability category value. *Sourced from NHC online Tropical Weather Outlooks.*

TC forecasts.

Figures 8.4 and 8.5 illustrate the reliability of the NHC 2013 48-hour TC genesis forecasts. The forecast system would be considered reliable but for one forecast category falling outside the 5% - 95% consistency bars. The performance has also improved on the 2012 hurricane season (see Fig. 6.2) and is less under-confident, although there is still some indication of under-forecast bias at the highest probability categories. Since the 2013 forecasts are evidently more reliable with more probability categories falling within the 5% - 95% consistency bars than the 2012 forecasts, it is of interest to determine whether forecast recalibration can be any more beneficial for the 2013 forecasts than it was for the 2012 forecasts in Section 6.3. The improvement after forecast recalibration was shown to be reduced in Chapter 3 if pre-recalibration forecast skill was already high since there is a maximum level skill possible for binary forecasts (see Section 3.2).

The simple translation method outlined in Section 2.4.1 is again employed here to recalibrate the 2013 NHC TC genesis forecasts out-of-sample using the 2012 TC forecast-outcome dataset as training data. Figures 8.6 and 8.7 show the results of forecast recalibration using the 2012 forecasts as the training set. Only three of the seven forecast categories now fall within the 5% - 95% consistency bars indicating a decrease of the reliability of the recalibrated 2013 forecasts. The degradation in forecast performance replicates the result of the 2012 forecasts recalibrated with the 2011 forecasts as training data. Particularly poor, is the highest recalibrated forecast probability category ($r_k = 0.999$) and the forecast category with $r_k = 0.649$ which both suffer from significant over-forecasting and lie well beyond the lower limit of the 1% - 99% consistency bars. The lack of reliability of these two forecast categories reflects the under-forecast bias demonstrated by the 2012 forecasts. Recalibration has appears to resulted in over-compensation at the higher categories such that some of the recalibrated forecast values are too high.

Reliability Diagram
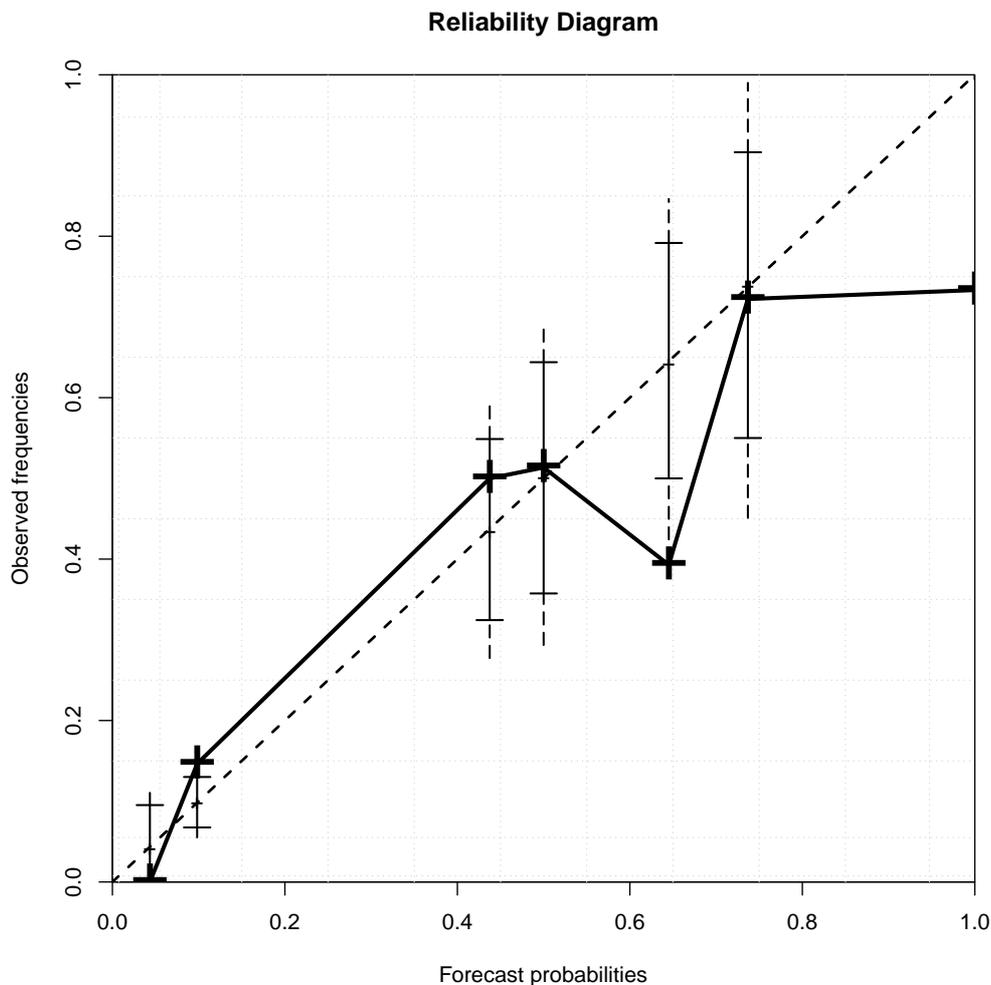
Figure 8.6: Recalibrated NHC 2013 TC forecast reliability: reliability diagram for the recalibrated NHC 2013 TC forecasts using 2012 forecast-outcome set as training data with 5% - 95% (1% - 99% vertical dashed line) consistency bars (the highest category $r_7 = 0.999$ has a consistency bar with zero width). The forecast probability bin boundaries (grey dotted lines) are identical to those on the original 2013 reliability diagram although the number of populated categories has decreased to 7. Forecast recalibration has resulted in a decrease of forecast reliability (c.f. Fig. 8.4). *Sourced from NHC online Tropical Weather Outlooks.*
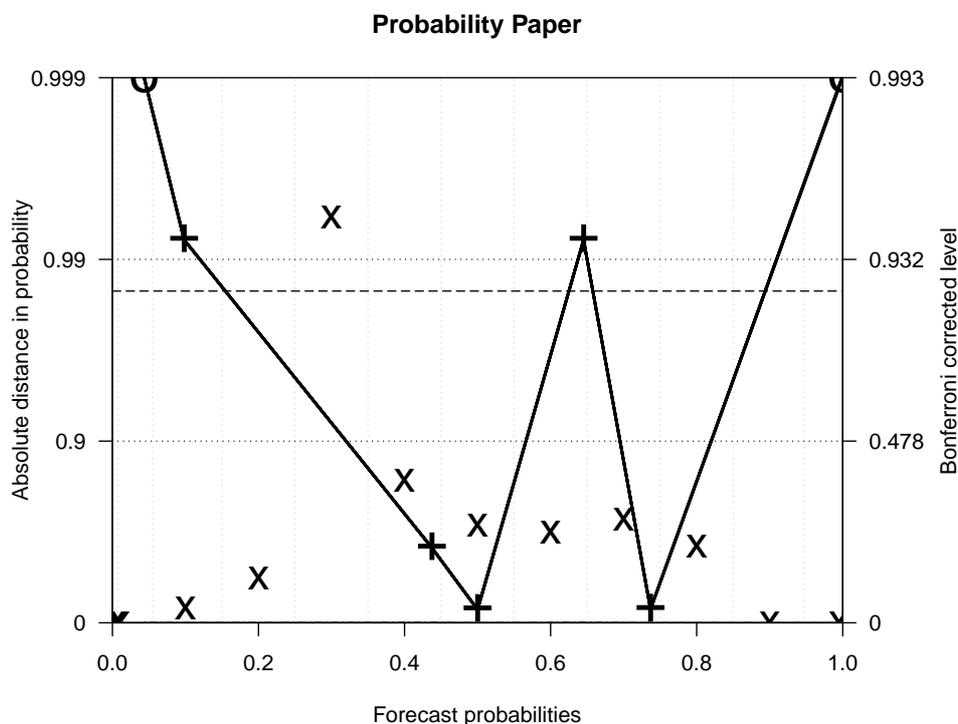
Figure 8.7: Recalibrated NHC 2013 TC forecast reliability: reliability diagram for the NHC 2013 TC forecasts recalibrated using the 2011 forecast-outcome set as training data with 5% - 95% (1% - 99% vertical dashed line) consistency bars. Forecast recalibration has resulted in a decrease of forecast reliability since most recalibrated probability categories (pluses) have larger probability distances than raw forecast categories (crosses). The forecast probability bin boundaries (grey dotted lines) are identical to those on the original 2013 reliability diagram although the number of populated categories has decreased to 7. See Fig. 8.5 for further details. *Sourced from NHC online Tropical Weather Outlooks.*

8.3 NHC 2013 tropical cyclone forecast Time Until Event

This final section investigates the forecast Time Until Event (TUE) profile of the NHC's 2013 48-hour TC raw forecasts, thereby completing the examination of their reliability. Some indication of an inversely proportional relationship between forecast probability and TUE was exhibited by the 2012 forecasts (see
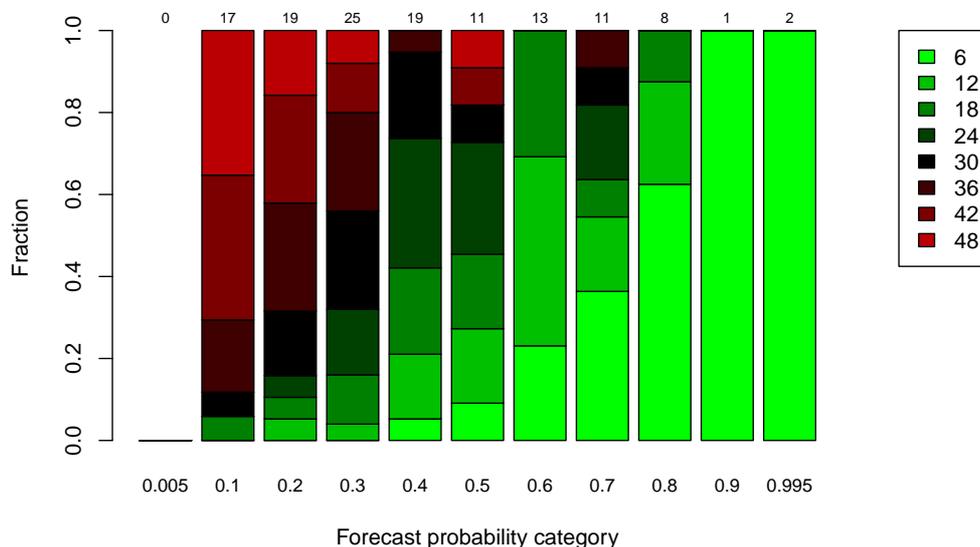
Figure 8.8: NHC 2013 TC forecast Time Until Event: fractions of verifying NHC 2013 TC forecasts* having different TUE lengths (in hours) for all probability categories. The coloured TUE categories denote the occurrence of TC formation between the time given and 6 hours previous to it. There is a clear pattern of larger fractions of shorter TUE with increasing forecast probability category. Total counts of verifying forecasts for each category are shown at the top of the bars. *Sourced from NHC online Tropical Weather Outlooks.*

Section 6.4) demonstrating a reliability bias towards higher forecast probabilities. By decomposing the forecasts by TUE lengths, it was shown that there is indeed some bias towards reliability where higher probability categories have shorter TUE lengths and lower probability categories have longer TUE lengths. The reliability bias would otherwise be masked when reading from a reliability diagram only so several supplementary diagrams were introduced. These diagrams are employed again here to examine whether the 2013 forecasts are subject to the same reliability bias. Figure 8.8 displays the fractions of NHC 2012 forecasts which verify with a TC formation within 48 hours ($Y = 1$) at each probability category r_k. Like the corresponding diagram for the 2012 forecasts, there is significant variation in the proportions of TUE lengths with a
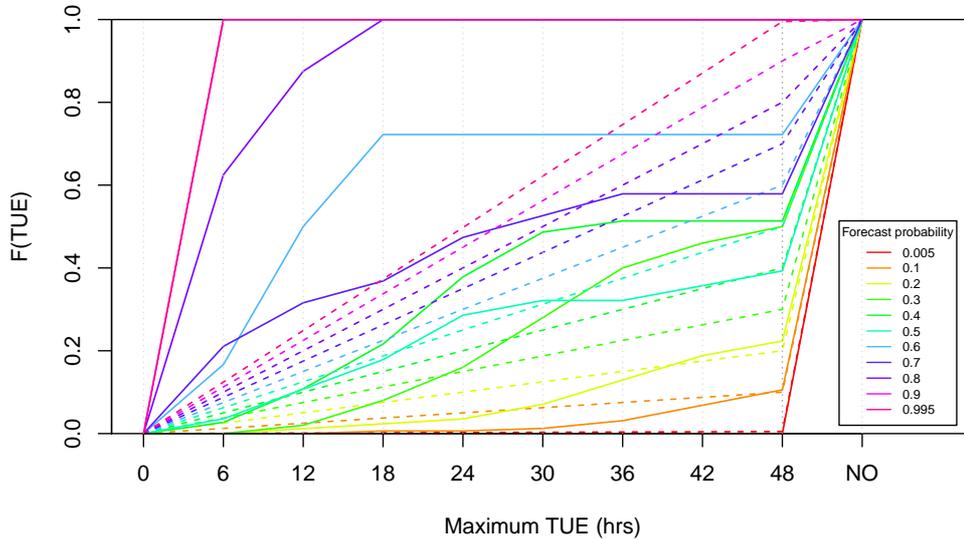
Figure 8.9: NHC 2013 TC forecast Time Until Event: CDFs of NHC 2013 TC forecast* TUE times (in hours) for each forecast probability category r_k (solid lines), and for a set of reliable forecasts ($f_k = r_k$) where the TUE times are computed with a discrete uniform distribution function (dashed lines). The higher probability curves lie well above the corresponding uniform distribution of reliable forecast TUE lengths. The TUE categories indicate the occurrence of a TC event between the time given and 6 hours previous to it, and an "NO" indicates a non-occurrence of a TC within 48 hours. *Sourced from NHC online Tropical Weather Outlooks.*

tendency for shorter lengths with increasing forecast probability. Figure 8.9 compares the cumulative distribution functions for the maximum TUE times of the actual forecasts in each probability category with those for a set of reliable forecasts ($f_k = r_k$) for which the maximum TUE lengths are uniformly distributed. As in Fig. 6.9 in Section 6.4, the CDF curves at the highest probability categories lie above the corresponding uniform CDF curves, demonstrating higher empirical probabilities at shorter TUE lengths. The lower probability categories of the 2013 forecasts do not appear to exhibit the same bias towards longer TUE lengths, however, as did the 2012 forecasts. To confirm whether a reliability bias does exist, the reliability diagram statistics r_k and f_k are listed

in Table 8.3 according to TUE (see Section 6.4 for the equivalent 2012 table). The statistics indicate a similar pattern of improved reliability of higher forecast probabilities at shorter TUE lengths, and improved reliability of lower forecast probabilities at longer TUE lengths to the 2012 forecast reliability statistics.

Table 8.3: NHC 2012 TC forecast reliability diagram statistics by TUE

TUE	Forecast probability r_k										
	0.005	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0-24 hrs	0	0.006	0.035	0.16	0.378	0.286	0.722	0.474	1.0	1.0	1.0
24-48 hrs	0	0.099	0.188	0.34	0.135	0.107	0	0.105	0	0	0

Green $5 - 95\%$; Orange $95 - 99\%$; Red $> 99\%$.

8.4 Forward view and conclusions

The statistical framework for hurricane forecast construction, evaluation, and recalibration proposed in Chapter 5 has been tested in a real-world hurricane forecasting scenario in this chapter. Probabilistic forecasts of storm counts during the 2013 Atlantic basin hurricane season have been constructed from the SC and CA forecast systems in Section 8.1, and evaluated with a proper scoring rule. The SC forecast and CA forecast systems achieved superior skill to the climatological forecast when predicting the total number of basin named storms for the 2013 season. In addition, the CA forecast system performed better in the predictions of CAT1-5 basin hurricanes using both the single and series CA methods. Both systems failed to outperform the climatological forecast, however, where predictions were made for the number of CAT1-5 U.S. landfalls. The superior performance of the two forecast systems' predictions of basin named storms compared to those of CAT1-5 U.S. landfalls may be reflective of the larger quantity of available observational data of the former category. The limitations of U.S. landfall predictions have already been discussed in Section 5.3. A cursory comparison of the 2013 seasonal hurricane predictions produced from the SC forecast and CA forecast systems, and those issued by several global

283

operational forecasting centres was also provided. All of the forecasts tended to over-estimate counts of all three categories: named basin storms, CAT1-5 basin hurricanes, and CAT1-5 U.S. landfalls. The inaccuracy of the statistical forecasts of the 2013 hurricane season outcomes serves as an example of where statistical modelling can perform poorly if relationships between predictands and predictors are incompletely understood, or are changing over time. The SC and CA models performed comparatively well, however, and may prove useful at least as benchmark models for future predictions of hurricane counts.

The latter part of the chapter focussed on the performance of the NHC's operational 2013 short-term TC forecasts in Section 8.2, evaluated using reliability diagrams. The 2013 TC forecasts demonstrated good reliability overall with 10 out of 11 forecast categories falling within the 5% - 95% consistency bars, and superior reliability to the equivalent operational forecasts from the 2012 hurricane season (see Section 6.2). The 2013 TC forecasts have also been recalibrated using a simple translation method, and the 2012 forecasts as the training set, and then re-evaluated with reliability diagrams. Recalibration resulted in a decrease of forecast reliability (as it did where recalibrating the 2012 TC forecasts with the 2011 training set in Section 6.3) suggesting that the predictability of TC formation, and hence, reliability of the TC forecast system, varies from year to year. Limited sample size of the training set has also most likely restricted the effectiveness of recalibration.

To present a more robust interpretation of the reliability of the NHC's 2013 TC forecasts, the relationship between forecast reliability and "Time Until Event" was investigated by analysing the profile of forecast TUE lengths on the diagrams and table proposed in Chapter 5. The added dimension of forecast TUE provides a more realistic interpretation of the reliability of each forecast probability category. Like the 2012 forecasts, there is a bias towards shorter TUE lengths at higher forecast probability categories indicating a potential for a reliability bias at those higher categories. Categorising the reliability statistics by TUE uncovers a more accurate picture of forecast reliability which is

dependent on TUE.

Included in this chapter are the following novel contributions or innovations:

- deployment of statistical hurricane forecast systems introduced in Chapter 5 to produce and evaluate predictions of the 2013 hurricane season

- comparison of statistical hurricane forecast systems introduced in Chapter 5 with existing operational seasonal hurricane forecasts

- evaluation and recalibration of the National Hurricane Center's 2013 48-hour TC genesis forecasts using reliability diagrams with consistency bars, and out-of-sample recalibration

- analysis of "Time Until Event" of the National Hurricane Center's 2013 48-hr TC genesis forecasts to provide a more complete illustration of forecast reliability

Appendices

Appendix A

Dynamical Systems

A.1   The Lorenz63 System

The Lorenz63 system [118] is a three dimensional dynamical system defined by
a set of three ordinary differential equations (with respect to time) given as

$$\dot{x} = -\sigma x + \sigma y \tag{A.1}$$

$$\dot{y} = -xz + rx - y \tag{A.2}$$

$$\dot{z} = xy - bz, \tag{A.3}$$

where σ is the Prandtl number, r is the Rayleigh number, and b is the system
parameter. The standard parameter values are: $\sigma = 10$, $r = 28$, and $b = 8/3$
[188], and the initial conditions are set to $\{x_0 = 0, y_0 = -0.01, z_0 = 9\}$. Nu-
merical solutions are obtained using a fourth order Runge-Kutta time stepping
scheme [160], with a time step of $h = 10^{-2}$.

A.2   Logistic Map

The logistic map is considered one of the most simple of chaotic nonlinear
dynamical systems given that it is one-dimensional and involves a single control
parameter. Exact solutions exist for the state variable, and the system can be

easily graphically visualised. The trajectory of the state variable is given by

$$x_{i+1} = ax_i(1 - x_i), \tag{A.4}$$

where x_{i+1} is the system's state at time $i + 1$, and a is the control parameter. The values of x_n are constrained so that $0 \leq x_n \leq 1$.

A.3   Toy hurricane system

A stochastic toy system is used to simulate annual Atlantic basin hurricane counts in several sections in this thesis. The mean number of storms follows a cycle of T_p years, while the number of storms in any given year is a random variable denoted Y. The annual storm counts are generated according to a stochastic Poisson process given as

$$Y_t \sim Pois(\lambda(t)), \tag{A.5}$$

where Y_t is the number of hurricanes in a given year t. The time-dependent mean parameter λ is determined by a sinusoidal function given by

$$\lambda(t) = A \cdot sin\left(\frac{2\pi t}{T_p}\right) + C, \tag{A.6}$$

where A are constants representing the amplitude and offset. The parameter values are typically set so that the simulated storm counts are similar to those that are observed in the Atlantic basin [61, 31]. These values correspond to $A = 2.5$, $C = 6.0$, $T_p = 60$ for CAT1-5 Atlantic basin hurricanes.

Appendix B

Forecast evaluation statistics of binary forecasts of Lorenz63

B.1 Datasets

The full set of numerical results of the binary forecast evaluation experiments in Chapters 2 and 3 are presented in this appendix. The target system is the three-dimensional Lorenz63 nonlinear dynamical system, formally defined in Appendix A.1. All probabilistic binary forecasts are constructed to predict the location of the x state variable lying above or below a given threshold x_θ. To generate system states of x, the Lorenz63 system is integrated using a fourth order Runge-Kutta time stepping scheme [160], with a time step of $h = 10^{-2}$.

Sequences of observed system states are generated by sampling at a given rate f_s using the model Ψ with additional observational noise. The noise level is set to 5% of the standard deviation of the climatological distribution of true states of x in all experiments. The size of the entire dataset of forecast-outcome pairs is $N = 2^{10}$, which is equally divided into the training and evaluation subsets. The former is used to recalibrate the forecasts while the latter is used to evaluate the recalibrated forecasts. Each non-overlapping sequence consists of sampled states up to the maximum forecast lead time of $\tau = 25.6$s. For

example, in a single non-overlapping sequence with $\tau = 25.6$s (measured in "Lorenz" seconds [118]) there are $25.6 * 5 = 128$ sampled states from which the initial conditions and evaluation outcomes are determined for each forecast lead time. These initial conditions and outcomes are then combined to form the observation time series - one for each forecast lead time.

Table B.1: Lorenz63 datasets

Parameter	
size of dataset (training + evaluation)	2^{10}
sampling rate (f_s)	5
observational noise level as percentage of climatological range of x (NL)	5%
climatological standard deviation of x (σ)	0.37
lead time range (τ)	$\{0.2, 0.4, 0.8, 1.6, 3.2,$ $6.4, 9.2, 12.8, 18.2, 25.6\}$*
ensemble size (N_{ens})	$\{4, 8, 16, 32, 64,$ $128, 256, 512, 1024\}$

*in Lorenz63 seconds [118].

B.2   Forecasts

Corresponding binary forecasts are produced for each of the outcomes described in Section B.1 above. The model Ψ is initialised with the initial conditions sampled at time $t = 0$. The resulting binary forecast is determined from the ensemble of size N_{ens} at lead time τ depending on the climatological distribution quantile θ using each of the density construction methods described in Section 2.2. Each forecast evaluation experiment is defined by a given set of the forecast-parameters; N_{ens}, τ, and θ. The numerical values of all relevant sampling

parameters and forecast-parameters used to produce the datasets are listed in Table B.1.

All forecast evaluation results listed in Sections B.3 and B.4 show the ignorance scores (rounded to 2 decimal places) of the forecasts relative to the climatological reference forecast p_{clim}. The best score is highlighted in green if it is strictly the minimum value or yellow if it is the joint minimum value.

B.3   Forecast evaluation results under PMS

	Forecast PDF Construction Method: Naive Counted								
				Ensemble Size N_{ens}					
	4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ	$\theta = 0.5$								
0.2	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.4	-0.97	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.8	Inf	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97
1.6	Inf	-0.95	-0.95	-0.95	-0.95	-0.95	-0.96	-0.95	-0.95
3.2	Inf	Inf	-0.91	-0.90	-0.90	-0.91	-0.90	-0.90	-0.90
6.4	Inf	-0.75	-0.76	-0.77	-0.77	-0.77	-0.77	-0.78	-0.78
9.2	Inf	Inf	-0.55	-0.56	-0.56	-0.57	-0.57	-0.57	-0.57
12.8	Inf	Inf	-0.16	-0.21	-0.22	-0.22	-0.22	-0.22	-0.22
18.2	Inf	Inf	0.03	-0.02	-0.04	-0.05	-0.05	-0.05	-0.05
25.6	Inf	Inf	0.03	0.01	0	0	-0.01	-0.01	-0.01
	$\theta = 0.9$								
0.2	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.4	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.8	-0.47	-0.47	-0.47	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
1.6	Inf	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45
3.2	Inf	Inf	-0.41	-0.41	-0.42	-0.42	-0.42	-0.42	-0.42
6.4	Inf	Inf	Inf	Inf	-0.37	-0.37	-0.37	-0.37	-0.37
9.2	Inf	Inf	Inf	Inf	-0.3	-0.3	-0.3	-0.3	-0.3
12.8	Inf	Inf	Inf	Inf	Inf	-0.22	-0.22	-0.22	-0.22
18.2	Inf	Inf	Inf	Inf	Inf	-0.12	-0.12	-0.12	-0.12
25.6	Inf	Inf	Inf	Inf	Inf	-0.05	-0.06	-0.05	-0.05
	$\theta = 0.99$								
0.2	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.4	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.8	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
1.6	Inf	-0.05	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
3.2	Inf	Inf	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
6.4	Inf	Inf	Inf	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
9.2	Inf	Inf	Inf	Inf	-0.02	-0.03	-0.02	-0.02	-0.02
12.8	Inf	Inf	Inf	Inf	-0.02	-0.01	-0.02	-0.02	-0.02
18.2	Inf	Inf	Inf	Inf	Inf	-0.02	-0.02	-0.02	-0.02
25.6	Inf	Inf	Inf	Inf	Inf	Inf	Inf	-0.01	-0.01

APPENDIX B: FORECAST EVALUATION STATISTICS OF LORENZ63 BINARY FORECASTS

	Forecast PDF Construction Method: Adjusted Counted								
	Ensemble Size N_{ens}								
	4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ	$\theta = 0.5$								
0.2	-0.83	-0.90	-0.94	-0.96	-0.97	-0.98	-0.98	-0.98	-0.98
0.4	-0.83	-0.90	-0.94	-0.96	-0.97	-0.97	-0.98	-0.98	-0.98
0.8	-0.82	-0.89	-0.93	-0.95	-0.96	-0.97	-0.97	-0.97	-0.97
1.6	-0.80	-0.87	-0.91	-0.93	-0.94	-0.95	-0.95	-0.95	-0.95
3.2	-0.75	-0.82	-0.87	-0.88	-0.89	-0.90	-0.90	-0.90	-0.90
6.4	-0.63	-0.69	-0.73	-0.75	-0.76	-0.77	-0.77	-0.77	-0.77
9.2	-0.41	-0.47	-0.53	-0.54	-0.56	-0.57	-0.57	-0.57	-0.57
12.8	-0.06	-0.11	-0.16	-0.20	-0.22	-0.22	-0.22	-0.22	-0.22
18.2	0.11	0.07	0.02	-0.03	-0.04	-0.04	-0.05	-0.05	-0.05
25.6	0.14	0.08	0.03	0.01	0	0	-0.01	-0.01	-0.01
	$\theta = 0.9$								
0.2	-0.42	-0.45	-0.46	-0.47	-0.48	-0.48	-0.48	-0.48	-0.48
0.4	-0.42	-0.45	-0.46	-0.47	-0.48	-0.48	-0.48	-0.48	-0.48
0.8	-0.42	-0.45	-0.46	-0.47	-0.47	-0.47	-0.48	-0.48	-0.48
1.6	-0.40	-0.42	-0.44	-0.44	-0.45	-0.45	-0.45	-0.45	-0.45
3.2	-0.35	-0.39	-0.40	-0.41	-0.41	-0.41	-0.41	-0.42	-0.42
6.4	-0.31	-0.34	-0.35	-0.36	-0.37	-0.37	-0.37	-0.37	-0.37
9.2	-0.23	-0.27	-0.28	-0.29	-0.30	-0.30	-0.30	-0.30	-0.30
12.8	-0.13	-0.16	-0.19	-0.19	-0.21	-0.21	-0.22	-0.22	-0.22
18.2	-0.02	-0.04	-0.09	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11
25.6	0.06	0.02	0	-0.03	-0.04	-0.05	-0.06	-0.05	-0.05
	$\theta = 0.99$								
0.2	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.4	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.8	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
1.6	-0.04	-0.05	-0.05	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
3.2	-0.03	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
6.4	-0.01	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
9.2	0	0	0	-0.01	-0.02	-0.03	-0.02	-0.02	-0.02
12.8	0.01	0.01	0.02	0	-0.02	-0.01	-0.02	-0.02	-0.02
18.2	0.01	0	0	0	0	-0.02	-0.02	-0.02	-0.02
25.6	0.01	0.02	0.02	0.01	0	0.01	0	-0.01	-0.01

Forecast PDF Construction Method: Bayesian									
				Ensemble Size N_{ens}					
	4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ	$\theta = 0.5$								
0.2	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99
0.4	-0.98	-0.99	-0.98	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99
0.8	-0.96	-0.97	-0.97	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
1.6	-0.91	-0.93	-0.95	-0.96	-0.96	-0.97	-0.97	-0.97	-0.97
3.2	-0.81	-0.85	-0.90	-0.91	-0.91	-0.91	-0.91	-0.91	-0.91
6.4	-0.58	-0.67	-0.76	-0.80	-0.81	-0.81	-0.82	-0.82	-0.82
9.2	-0.21	-0.34	-0.47	-0.54	-0.57	-0.58	-0.58	-0.59	-0.59
12.8	0.14	0.04	-0.07	-0.16	-0.22	-0.22	-0.22	-0.22	-0.22
18.2	0.23	0.13	0.05	-0.01	-0.03	-0.04	-0.04	-0.05	-0.04
25.6	0.14	0.08	0.03	0.01	0	0	-0.01	-0.01	-0.01
	$\theta = 0.9$								
0.2	-0.04	-0.26	-0.35	-0.39	-0.42	-0.44	-0.45	-0.45	-0.46
0.4	0.1	-0.16	-0.27	-0.32	-0.38	-0.4	-0.42	-0.42	-0.43
0.8	0.26	-0.02	-0.15	-0.21	-0.27	-0.3	-0.32	-0.33	-0.34
1.6	0.43	0.16	0.02	-0.04	-0.1	-0.12	-0.13	-0.13	-0.13
3.2	0.58	0.35	0.19	0.11	0.06	0.02	0.02	0.02	0.02
6.4	0.68	0.51	0.38	0.31	0.26	0.22	0.21	0.22	0.21
9.2	0.74	0.62	0.51	0.46	0.42	0.39	0.38	0.39	0.39
12.8	0.69	0.62	0.52	0.47	0.45	0.43	0.42	0.43	0.42
18.2	0.53	0.48	0.41	0.37	0.36	0.34	0.33	0.34	0.34
25.6	0.25	0.21	0.19	0.17	0.16	0.15	0.14	0.14	0.14
	$\theta = 0.99$								
0.2	0.27	0.22	0.21	0.13	0.08	0.06	0.04	0.02	0.01
0.4	0.29	0.24	0.23	0.16	0.12	0.1	0.09	0.07	0.06
0.8	0.3	0.26	0.26	0.2	0.16	0.14	0.13	0.12	0.12
1.6	0.31	0.29	0.29	0.23	0.19	0.18	0.17	0.17	0.17
3.2	0.31	0.29	0.29	0.24	0.2	0.19	0.18	0.18	0.18
6.4	0.28	0.26	0.28	0.22	0.19	0.18	0.17	0.16	0.16
9.2	0.24	0.23	0.24	0.2	0.16	0.16	0.15	0.14	0.14
12.8	0.18	0.18	0.19	0.15	0.13	0.13	0.12	0.11	0.11
18.2	0.11	0.11	0.11	0.1	0.09	0.09	0.08	0.07	0.07
25.6	0.05	0.06	0.06	0.05	0.04	0.05	0.04	0.03	0.03

APPENDIX B: FORECAST EVALUATION STATISTICS OF LORENZ63 BINARY FORECASTS

				Forecast PDF Construction Method: Kernel dressed and blended					
				Ensemble Size N_{ens}					
	4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ	$\theta = 0.5$								
0.2	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.4	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.8	-0.96	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97
1.6	-0.92	-0.94	-0.94	-0.95	-0.95	-0.95	-0.95	-0.95	-0.95
3.2	-0.84	-0.86	-0.89	-0.89	-0.90	-0.90	-0.90	-0.90	-0.90
6.4	-0.67	-0.71	-0.73	-0.75	-0.75	-0.76	-0.77	-0.77	-0.77
9.2	-0.41	-0.46	-0.51	-0.53	-0.54	-0.55	-0.56	-0.56	-0.57
12.8	-0.08	-0.11	-0.15	-0.19	-0.21	-0.21	-0.21	-0.22	-0.22
18.2	0.03	0.03	0.01	-0.03	-0.04	-0.04	-0.05	-0.05	-0.05
25.6	0.02	0.02	0.01	0.01	0	0	-0.01	-0.01	-0.01
	$\theta = 0.9$								
0.2	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.4	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.8	-0.47	-0.47	-0.47	-0.48	--0.48	-0.48	-0.48	-0.48	-0.48
1.6	-0.44	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45
3.2	-0.38	-0.4	-0.41	-0.41	-0.41	-0.41	-0.42	-0.42	-0.42
6.4	-0.33	-0.35	-0.36	-0.36	-0.37	-0.37	-0.37	-0.37	-0.37
9.2	-0.23	-0.26	-0.27	-0.28	-0.29	-0.29	-0.3	-0.3	-0.3
12.8	-0.13	-0.17	-0.18	-0.19	-0.2	-0.21	-0.21	-0.21	-0.22
18.2	-0.05	-0.06	-0.09	-0.1	-0.11	-0.12	-0.12	-0.12	-0.12
25.6	-0.01	-0.03	-0.03	-0.05	-0.04	-0.05	-0.05	-0.05	-0.05
	$\theta = 0.99$								
0.2	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.4	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.8	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
1.6	-0.05	-0.05	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
3.2	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
6.4	-0.01	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
9.2	0	0	0	-0.01	-0.02	-0.02	-0.02	-0.02	-0.02
12.8	0.01	0.01	0.01	-0.01	-0.02	-0.01	-0.01	-0.02	-0.02
18.2	0	0	0	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02
25.6	0	0	0	0	0	0	0	-0.01	-0.01

B.4 Forecast evaluation results under IMS

				Forecast PDF Construction Method: Naive Counted					
				Ensemble Size N_{ens}					
	4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ	$\theta = 0.5$								
0.2	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.4	-0.97	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.8	Inf	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97
1.6	Inf	-0.95	-0.95	-0.95	-0.96	-0.96	-0.96	-0.96	-0.96
3.2	Inf	Inf	-0.91	-0.90	-0.90	-0.91	-0.90	-0.90	-0.90
6.4	Inf	Inf	-0.76	-0.77	-0.77	-0.77	-0.77	-0.78	-0.78
9.2	Inf	Inf	-0.55	-0.56	-0.56	-0.57	-0.57	-0.57	-0.57
12.8	Inf	Inf	Inf	-0.20	-0.21	-0.21	-0.22	-0.22	-0.22
18.2	Inf	Inf	0	-0.04	-0.05	-0.05	-0.05	-0.05	-0.05
25.6	Inf	Inf	0.06	0.02	0	-0.01	-0.01	-0.01	-0.01
	$\theta = 0.9$								
0.2	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.4	Inf	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.8	Inf	Inf	Inf	Inf	Inf	-0.48	-0.48	-0.48	-0.48
1.6	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45
3.2	Inf	Inf	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41
6.4	Inf	Inf	-0.37	-0.37	-0.37	-0.37	-0.37	-0.37	-0.37
9.2	Inf	Inf	Inf	-0.29	-0.3	-0.3	-0.3	-0.3	-0.3
12.8	Inf	Inf	Inf	Inf	Inf	-0.22	-0.22	-0.22	-0.22
18.2	Inf	Inf	Inf	Inf	Inf	Inf	-0.12	-0.12	-0.12
25.6	Inf	Inf	Inf	Inf	Inf	-0.05	-0.05	-0.05	-0.05
	$\theta = 0.99$								
0.2	Inf	Inf	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.4	Inf	Inf	Inf	Inf	-0.06	-0.06	-0.06	-0.06	-0.06
0.8	Inf	Inf	Inf	-0.06	-0.05	-0.05	-0.05	-0.05	-0.05
1.6	Inf	Inf	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
3.2	Inf	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
6.4	Inf	Inf	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
9.2	Inf	Inf	Inf	Inf	-0.02	-0.02	-0.02	-0.02	-0.02
12.8	Inf	Inf	-0.02	-0.02	-0.02	-0.01	-0.01	-0.02	-0.02
18.2	Inf	Inf	Inf	Inf	Inf	-0.02	-0.02	-0.02	-0.01
25.6	Inf	Inf	Inf	Inf	Inf	Inf	0	0	0

Forecast PDF Construction Method: Adjusted Counted								
Ensemble Size N_{ens}								
4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ $\theta = 0.5$								
0.2 -0.83	-0.90	-0.94	-0.96	-0.97	-0.98	-0.98	-0.98	-0.98
0.4 -0.83	-0.90	-0.94	-0.96	-0.97	-0.97	-0.98	-0.98	-0.98
0.8 -0.82	-0.89	-0.93	-0.95	-0.96	-0.97	-0.97	-0.97	-0.97
1.6 -0.80	-0.87	-0.91	-0.93	-0.95	-0.95	-0.95	-0.95	-0.96
3.2 -0.75	-0.83	-0.87	-0.88	-0.89	-0.90	-0.90	-0.90	-0.90
6.4 -0.63	-0.69	-0.73	-0.75	-0.76	-0.77	-0.77	-0.77	-0.77
9.2 -0.40	-0.47	-0.53	-0.55	-0.56	-0.57	-0.57	-0.57	-0.57
12.8 -0.06	-0.1	-0.16	-0.20	-0.21	-0.21	-0.22	-0.22	-0.22
18.2 0.08	0.03	-0.01	-0.01	-0.04	-0.05	-0.05	-0.05	-0.05
25.6 0.15	0.08	0.06	0.02	0	-0.01	-0.01	-0.01	-0.01
$\theta = 0.9$								
0.2 -0.42	-0.45	-0.46	-0.47	-0.47	-0.48	-0.48	-0.48	-0.48
0.4 -0.42	-0.45	-0.46	-0.47	-0.47	-0.47	-0.47	-0.48	-0.48
0.8 -0.41	-0.44	-0.45	-0.46	-0.46	-0.46	-0.47	-0.47	-0.47
1.6 -0.40	-0.43	-0.44	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45
3.2 -0.35	-0.39	-0.40	-0.40	-0.41	-0.41	-0.41	-0.41	-0.41
6.4 -0.31	-0.34	-0.36	-0.36	-0.37	-0.37	-0.37	-0.37	-0.37
9.2 -0.22	-0.27	-0.28	-0.29	-0.30	-0.30	-0.30	-0.30	-0.30
12.8 -0.14	-0.18	-0.21	-0.21	-0.21	-0.22	-0.22	-0.22	-0.22
18.2 -0.03	-0.06	-0.09	-0.11	-0.11	-0.12	-0.12	-0.12	-0.12
25.6 0.04	0.02	0	-0.01	-0.03	-0.05	-0.05	-0.05	-0.05
$\theta = 0.99$								
0.2 -0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.4 0.05	-0.05	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.8 -0.05	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
1.6 -0.04	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05
3.2 -0.03	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
6.4 -0.02	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
9.2 0	0	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.02
12.8 0	-0.01	-0.02	-0.02	-0.01	-0.01	-0.02	-0.01	-0.02
18.2 0.02	0.01	0.01	0.01	-0.02	-0.01	-0.01	-0.01	-0.01
25.6 0.02	0.02	0.02	0.01	0.01	0	0	0	0

	Forecast PDF Construction Method: Bayesian								
				Ensemble Size N_{ens}					
	4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ	$\theta = 0.5$								
0.2	-0.98	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99	-0.99
0.4	-0.97	-0.98	-0.98	-0.98	-0.99	-0.99	-0.99	-0.99	-0.99
0.8	-0.96	-0.98	-0.98	-0.98	-0.98	-0.99	-0.99	-0.99	-0.99
1.6	-0.90	-0.94	-0.95	-0.96	-0.97	-0.97	-0.97	-0.97	-0.97
3.2	-0.78	-0.85	-0.89	-0.90	-0.91	-0.91	-0.91	-0.91	-0.91
6.4	-0.56	-0.67	-0.75	-0.79	-0.80	-0.81	-0.81	-0.82	-0.82
9.2	-0.20	-0.35	-0.47	-0.54	-0.57	-0.58	-0.58	-0.59	-0.59
12.8	0.13	0.02	-0.08	-0.17	-0.20	-0.21	-0.21	-0.21	-0.21
18.2	0.22	0.1	0.05	-0.02	-0.03	-0.04	-0.04	-0.04	-0.04
25.6	0.15	0.08	0.06	0.02	0	-0.01	-0.01	-0.01	-0.01
	$\theta = 0.9$								
0.2	-0.12	-0.31	-0.39	-0.42	-0.44	-0.45	-0.46	-0.46	-0.46
0.4	0	-0.23	-0.33	-0.36	-0.4	-0.42	-0.43	-0.44	0.44
0.8	0.14	-0.11	-0.23	-0.26	-0.32	-0.35	-0.36	-0.38	-0.38
1.6	0.31	0.05	-0.07	-0.11	-0.17	-0.19	-0.2	-0.2	-0.21
3.2	0.48	0.24	0.1	0.05	-0.01	-0.04	-0.05	-0.05	-0.06
6.4	0.58	0.41	0.29	0.24	0.19	0.15	0.14	0.14	0.14
9.2	0.66	0.54	0.45	0.41	0.37	0.34	0.33	0.33	0.33
12.8	0.61	0.56	0.48	0.44	0.41	0.4	0.39	0.39	0.38
18.2	0.48	0.44	0.4	0.37	0.34	0.33	0.31	0.32	0.31
25.6	0.22	0.21	0.18	0.17	0.16	0.14	0.13	0.13	0.13
	$\theta = 0.99$								
0.2	0.2	0.15	0.1	0.07	0.02	0	-0.02	-0.03	-0.04
0.4	0.22	0.18	0.13	0.11	0.07	0.05	0.03	0.02	0.01
0.8	0.23	0.2	0.16	0.15	0.11	0.1	0.09	0.08	0.08
1.6	0.25	0.22	0.19	0.18	0.15	0.14	0.14	0.14	0.14
3.2	0.24	0.23	0.2	0.19	0.17	0.16	0.15	0.15	0.15
6.4	0.23	0.22	0.19	0.18	0.15	0.15	0.14	0.14	0.14
9.2	0.2	0.19	0.17	0.16	0.14	0.13	0.13	0.12	0.12
12.8	0.16	0.15	0.13	0.13	0.11	0.1	0.1	0.1	0.1
18.2	0.12	0.12	0.11	0.1	0.08	0.07	0.07	0.07	0.06
25.6	0.05	0.06	0.05	0.04	0.04	0.04	0.03	0.03	0.03

Forecast PDF Construction Method: Kernel dressed and blended								
Ensemble Size N_{ens}								
4	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
τ — $\theta = 0.5$								
0.2 -0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.4 -0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98
0.8 -0.96	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97	-0.97
1.6 -0.93	-0.94	-0.95	-0.95	-0.96	-0.96	-0.96	-0.96	-0.96
3.2 -0.85	-0.87	-0.89	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
6.4 -0.68	-0.71	-0.73	-0.75	-0.76	-0.76	-0.77	-0.77	-0.77
9.2 -0.41	-0.47	-0.52	-0.54	-0.55	-0.56	-0.57	-0.57	-0.57
12.8 -0.08	-0.11	-0.15	-0.19	-0.21	-0.21	-0.21	-0.22	-0.22
18.2 0.01	0	-0.01	--0.04	-0.05	-0.05	-0.05	-0.05	-0.05
25.6 0.02	0.02	0.03	0.01	0	-0.01	-0.01	--0.01	-0.01
$\theta = 0.9$								
0.2 -0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.4 -0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
0.8 -0.46	-0.46	-0.46	-0.46	--0.46	-0.47	-0.47	-0.47	-0.47
1.6 -0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45
3.2 -0.38	-0.4	-0.40	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41
6.4 -0.32	-0.35	-0.36	-0.36	-0.37	-0.37	-0.37	-0.37	-0.37
9.2 -0.23	-0.27	-0.28	-0.29	-0.29	-0.30	-0.30	-0.3	-0.3
12.8 -0.14	-0.17	-0.20	-0.21	-0.21	-0.22	-0.22	-0.22	-0.22
18.2 -0.05	-0.07	-0.09	-0.11	-0.12	-0.12	-0.12	-0.12	-0.12
25.6 -0.02	-0.02	-0.03	-0.03	-0.04	-0.05	-0.05	-0.05	-0.05
$\theta = 0.99$								
0.2 -0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.4 -0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
0.8 -0.06	-0.05	-0.05	-0.06	-0.05	-0.05	-0.05	-0.05	-0.05
1.6 -0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.06	-0.06
3.2 -0.03	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
6.4 -0.02	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
9.2 0	0	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.02
12.8 0	-0.01	-0.02	-0.02	-0.02	0.02	-0.01	-0.02	-0.02
18.2 0.01	0	0	0	-0.02	-0.01	-0.01	-0.01	-0.01
25.6 0	0.01	0.01	0	0	0	0	0	-0

Appendix C

Hurricane Regression Modelling Diagnostics

All diagnostic output is based on the normalised quantile residuals of the fitted regression models in Chapter 7. This type of residual analysis ensures that, irrespective of the distribution of the response variable, the true residual values r have a standard normal distribution based on the assumption that the model is an adequate fit. Normality tests are well established in statistical practice so the analysis of the normalised quantile residuals is a convenient check for model adequacy. The following notation follows Dunn and Smyth [42].

Let $y_1, \ldots, y_N$ denote the response outcomes which are assumed to be independent and distributed according to a distribution $\mathcal{P}(\mu_i, \theta)$ where $\mu_i = E[y]$ and θ is the parameter vector of the regression model. The response variable is assumed to depend on the vector of predictors $\mathbf{x}_i$, and the $k \times 1$ parameter vector $\boldsymbol{\beta}$. Also, let $F(y; \mu, \theta)$ represent the cumulative distribution function of $\mathcal{P}(\mu_i, \theta)$. In the case where F is continuous, then the $F(y; \mu_i, \theta)$ are uniformly distributed on the unit interval so that the quantile residuals r are given by

$$r_{q,i} = \Phi[F(y; \hat{\mu}_i, \hat{\theta})], \tag{C.1}$$

where Φ is the standard normal distribution function. Ignoring sampling uncertainty in the $\hat{\mu}_i$ and $\hat{\theta}$, all $r_{q,i}$ have an asymptotic standard normal distribution

as long as $\boldsymbol{\beta}$ and θ are consistently estimated.

Given the case that y and F are discrete, as is the case with storm counts in the regression modelling exercise in Chapter 7, a more general definition of the quantile residuals is necessary. In this case, the normalised quantile residuals are *randomised*. Let $a_i = lim_{y \to y_i} F(y; \mu_i, \theta)$ and $b_i = F(y_i; \mu_i, \theta)$. Now let the randomised quantile residual for yi be given by

$$r_{q,i} = \Phi^{-1}(u_i), \tag{C.2}$$

where u_i is a random variable on the interval $(a_i, b_i]$. The formulation of the normalised quantile residuals in Eqn. (C.2) ensures that all $r_{q,i}$ are standard normal distributed, taking into account sampling uncertainty in the $\hat{\mu}_i$ and $\hat{\theta}$. See Dunn and Smyth [42] for further details.

C.1 Regression diagnostics plots

All of the plots below have been produced with R statistical software (R Development Core Team, 2008) using the freely available Generalized Additive Models for Location Scale and Shape (GAMLSS) package [190].
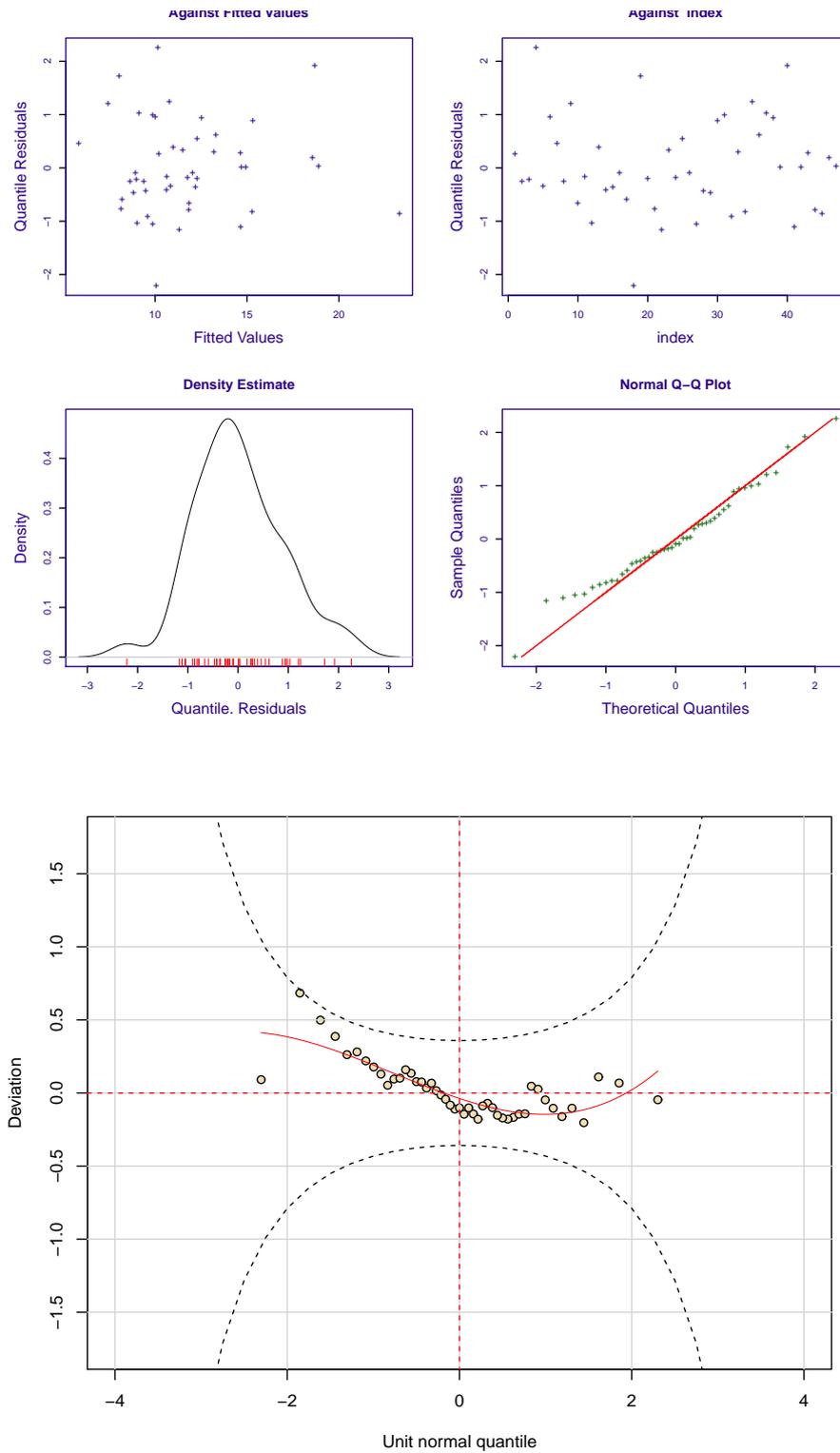
Figure C.1: Diagnostics plots and worm plot for Poisson model of Atlantic basin named storm counts regressed on year, SST_{Atl} and SST_{trop} from 1966-2012.
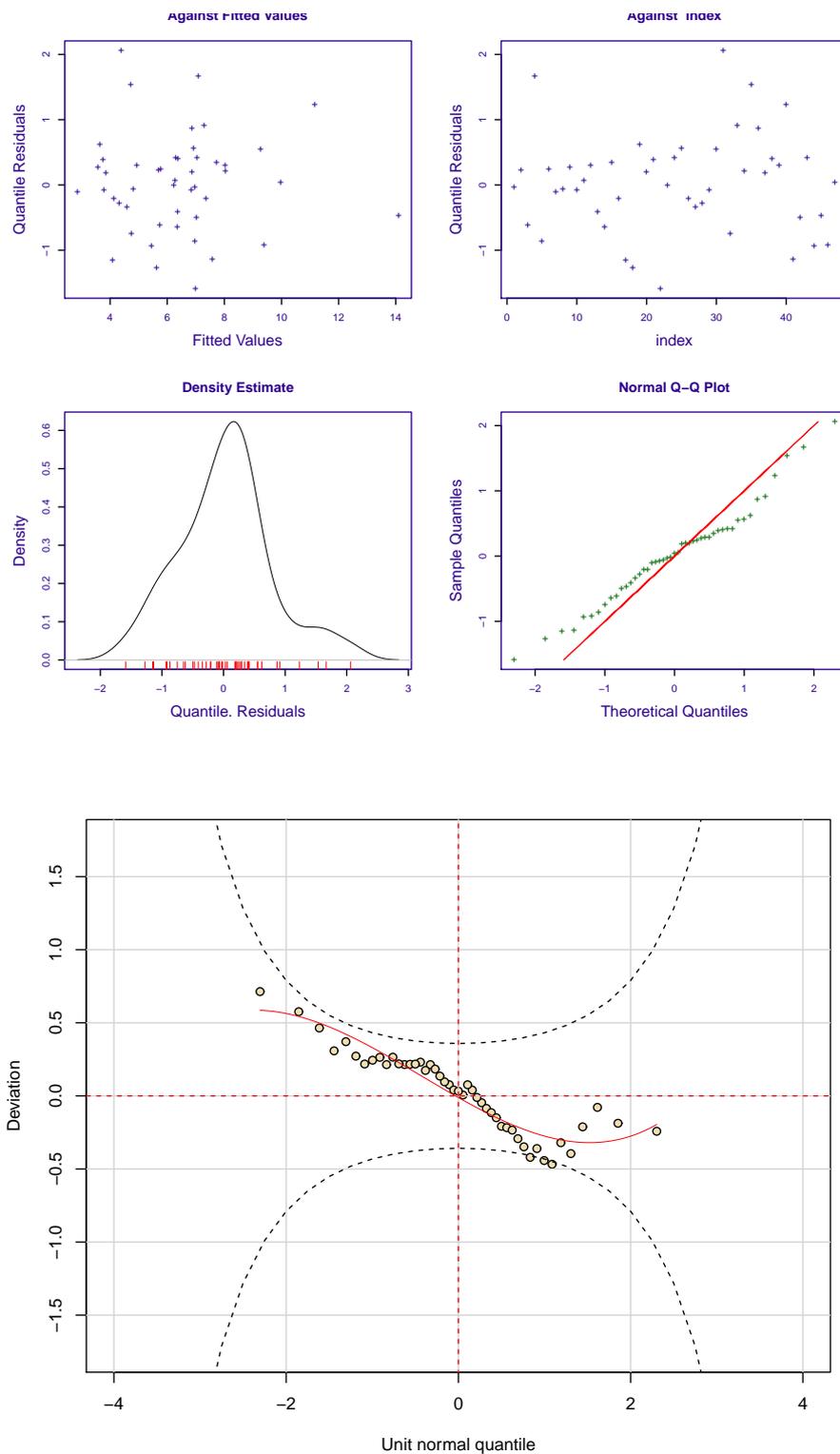
Figure C.2: Diagnostics plots and worm plot for Poisson model of Atlantic basin CAT1-5 hurricane counts regressed on SST_{Atl} and SST_{trop} from 1966-2012.
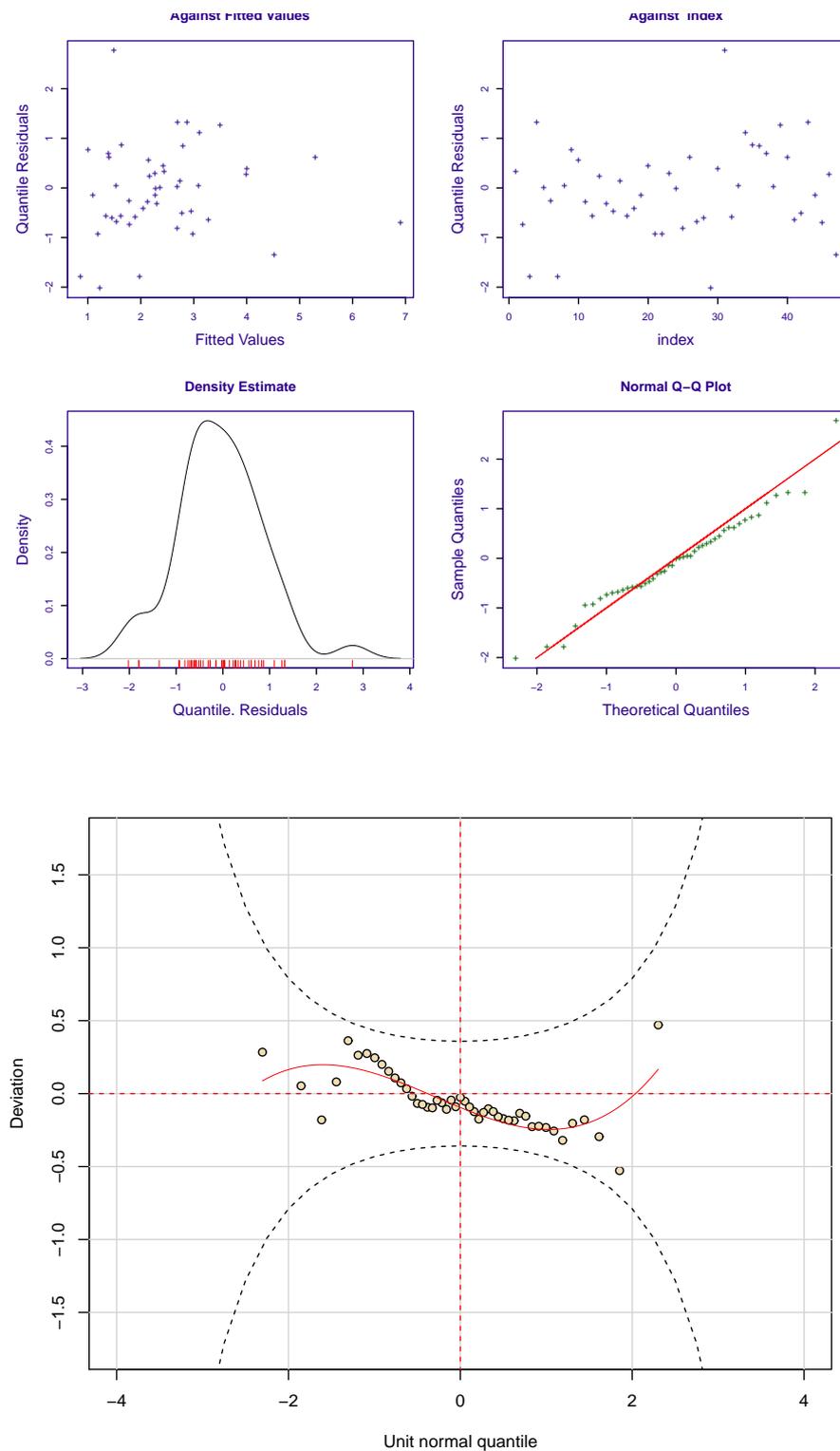
Figure C.3: Diagnostics plots and worm plot for Poisson model of Atlantic basin CAT3-5 hurricane counts regressed on SST_{Atl} and SST_{trop} from 1966-2012.
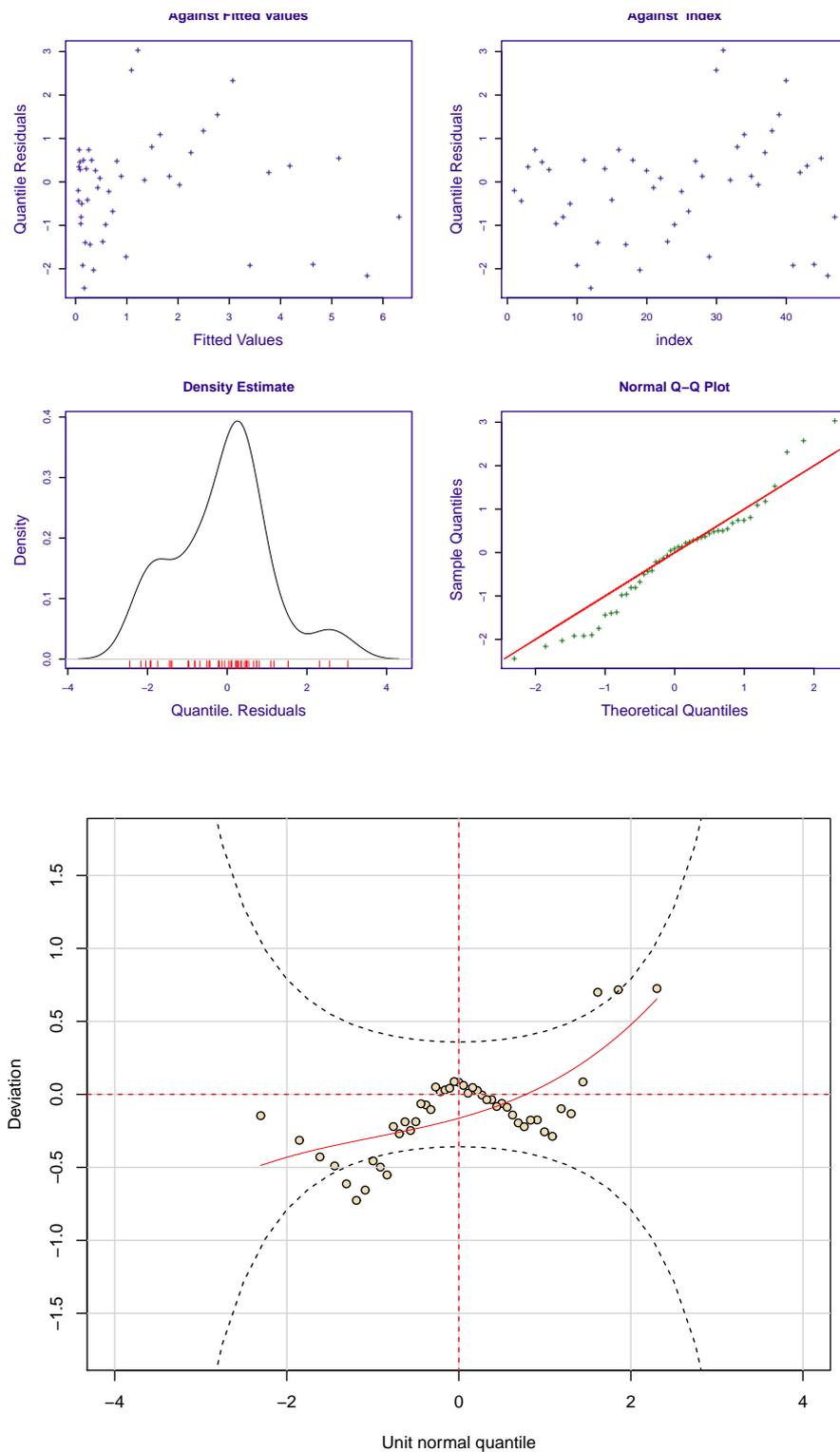
Figure C.4: Diagnostics plots and worm plot for Poisson model of Atlantic CAT1-5 US landfall counts regressed on SST_{Atl} and SST_{trop} from 1966-2012.
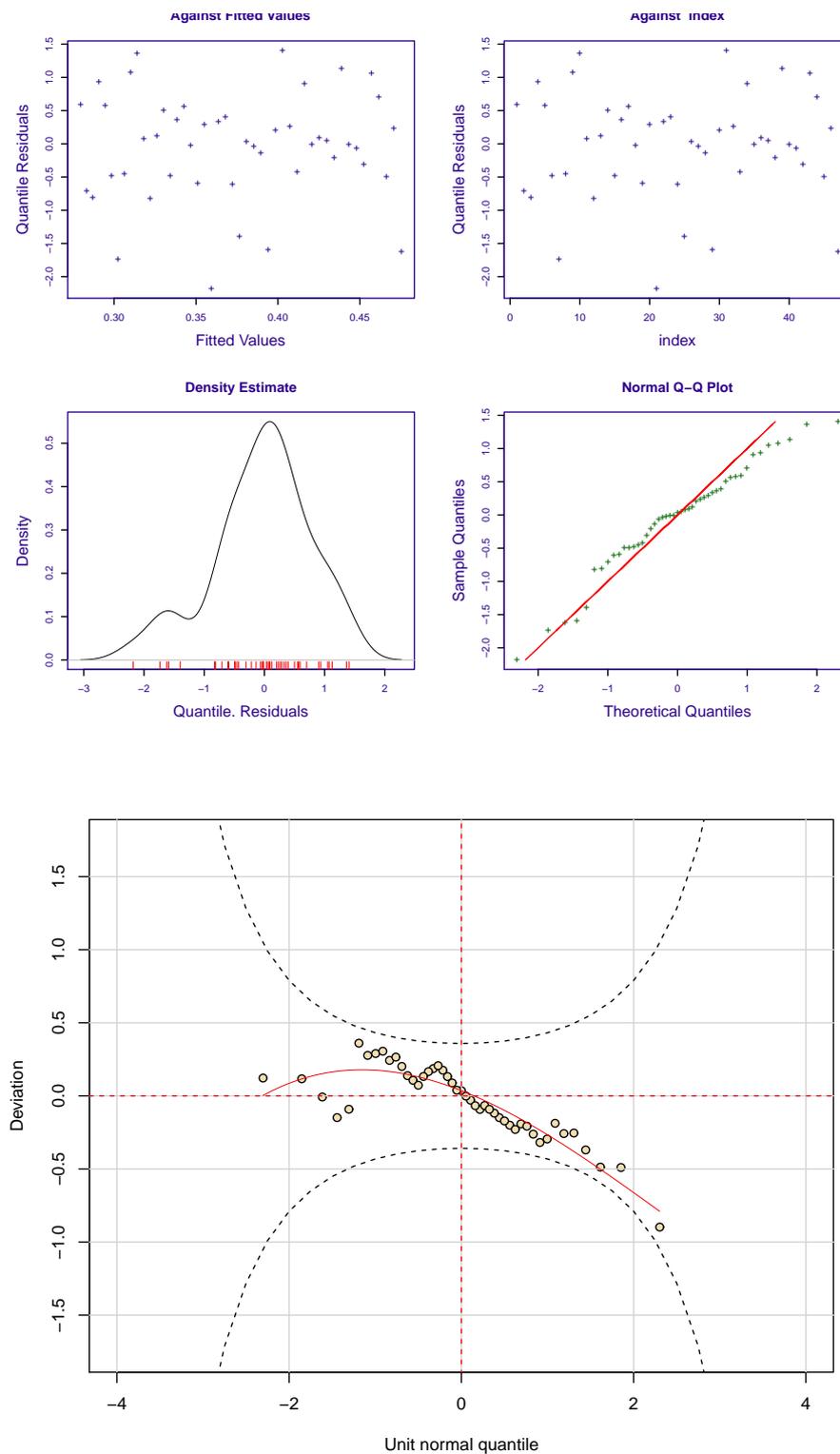
Figure C.5: Diagnostics plots and worm plot for logistic model of Atlantic basin CAT3-5 hurricane count fractions regressed on year from 1966-2012.
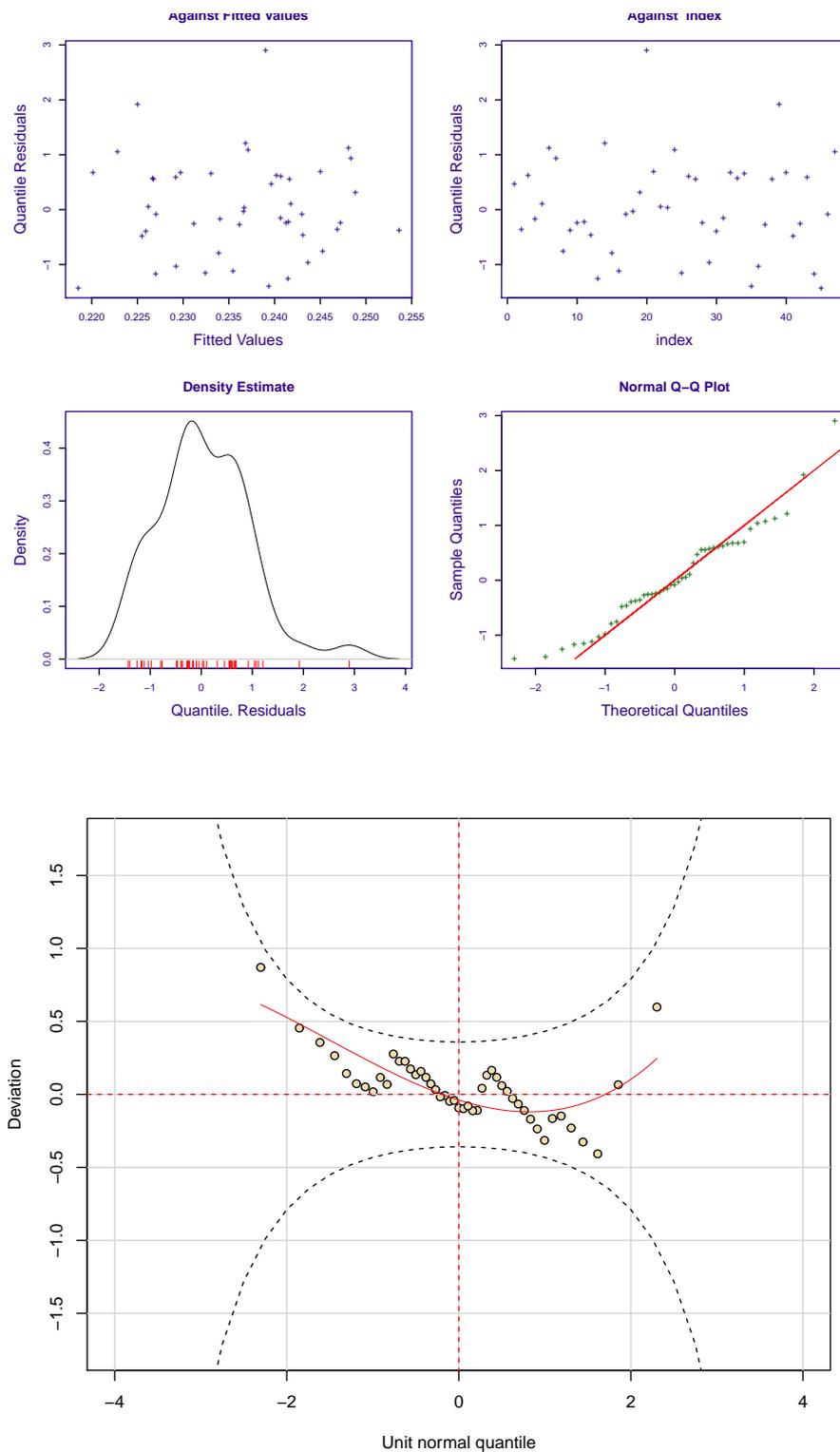
Figure C.6: Diagnostics plots and worm plot for logistic model of Atlantic CAT1-5 US landfall count fractions regressed on SST_{Atl} from 1966-2012.

Appendix D

Glossary

Bibliography

[1] A. Agresti. Score and pseudo-score confidence intervals for categorical data analysis. *The American Statistician.*

[2] A. Agresti. *An Introduction to Categorical Data Analysis.* John Wiley and Sons, 2nd edition, 2007.

[3] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, May 1998.

[4] J. Aitchison and C. G. G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420, December 1976.

[5] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[6] J. S. Armstrong. *Principles of Forecasting: A handbook for researchers and practitioners.* Kluwer Academic Publishers, 2001.

[7] F. Atger. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review*, 131(8):1509–1523, August 2003.

[8] F. Atger. Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 130(597):627–646, January 2004.

[9] M. A. Bender, T. R. Knutson, R. E. Tuleya, J. J. Sirutis, G. A. Vecchi, S. T. Garner, and I. M. Held. Modeled impact of anthropogenic warming on the frequency of intense atlantic hurricanes. *Science*, 327(5964):454–458, January 2010.

[10] R. E. Bergen and R. P. Harnack. Long-range temperature prediction using a simple analog approach. *Monthly Weather Review*, 110(8):1083–1099, August 1982.

[11] J. L. Beven. The boguscane - a serious problem with the ncep medium-range forecast model in the tropics. Preprints, 23rd Conf. on Hurricanes and Tropical Meteorology, Dallas, TX, Amer. Meteor. Soc., 845-848, (1999).

[12] R. Binter. *Applied Probabilistic Forecasting*. PhD thesis, London School of Economics and Political Science, 2012.

[13] E. S. Blake and C. W. Landsea. The deadliest, costliest, and most intense united states tropical cyclones from 1851 to 2006 (and other frequently requested hurricane facts). Tech Memo. NWS TPC-5, 2011.

[14] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[15] M. C. Bove, J. B. Elsner, C. W. Landsea, X. Niu, and J. J. OBrien. Effect of el niño on u.s. landfalling hurricanes, revisited. *Bulletin of the American Meteorological Society*, 79(11):2477–2482, November 1998.

[16] A. Bowman. A comparative study of some kernel-based nonparametric density estimators. *JSCS*, 21(3-4):313–327, March 1985.

[17] A. A. Bradley, S. S. Scwartz, and H. Tempei. Sampling uncertainty and confidence intervals for the brier score and brier skill score. *Weather and Forecasting*, 23(5):992–1006, October 2008.

[18] A. A. Bradley, H. Tempei, and S. S. Scwartz. Distributions-oriented verification of probability forecasts for small data samples. *Weather and Forecasting*, 18(5):903–917, October 2003.

[19] G. W. Brier. Verification of forecaster's confidence and the use of probability statements in weather forecasting. Research Paper No. 16, U.S. Weather Bureau, Washington D.C., (1944).

[20] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, January 1950.

[21] G. W. Brier and R. A. Allen. *Verification of weather forecasts.* Boston: American Meteorological Society, 1951.

[22] E. Britton, P. Fisher, and J. Whitley. The inflation report projections: understanding the fan chart. *Bank of England Quarterly Bulletin*, 38(1):30–37, February 1998.

[23] J. Bröcker. Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, 136(11):4488–4502, November 2008.

[24] J. Bröcker and L. A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, 2007.

[25] J. Bröcker and L. A. Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

[26] J. Bröcker and L. A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A*, 60(4):663–678, 2008.

[27] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer-Verlag, 2nd edition, 2002.

[28] S. J. Camargo, A. B. Barnston, P. J. Klotzbach, and C. W. Landsea. Seasonal tropical cyclone forecasts. *WMO Bulletin*, 56(4):297–309, October 2007.

[29] A. C. Cameron and P. K. Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, 46(3):347–364, December 1990.

[30] A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 1998.

[31] P. Chylek and G. Lesins. Multidecadal variability of atlantic hurricane activity: 1851-2007. *Geophysical Research Letters*, 113(D22):D22106, November 2008.

[32] R. Clark and M. Déque. Conditional probability seasonal predictions of precipitation. *Quarterly Journal of the Royal Meteorological Society*, 129(587):179–193, December 2003.

[33] M. J. a. Costa. *Penalized Spline Models and Applications*. PhD thesis, University of Warwick, 2008.

[34] K. Coughlin, E. Bellone, T. Laepple, S. Jewson, and K. Nzerem. A relationship between all atlantic hurricanes and those that make landfall in the usa. *Quarterly Journal of the Royal Meteorological Society*, 135(639):371–379, 2009.

[35] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[36] D. R. Cox. Interaction. 52(1):1–24, April 1984.

[37] P. S. Dailey, G. Zuba, G. Ljung, I. M. Dima, and J. Guin. On the relationship between north atlantic sea surface temperatures and u.s.

hurricane landfall risk. *Journal of Applied Meteorology and Climatology*, 48(1):111–129, January 2009.

[38] C. Dean and J. F. Lawless. Tests for detecting overdispersion in poisson regression models. 84(406):467–472, June 1989.

[39] M. Demaria, M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan. Further improvements to the statistical hurricane intensity prediction scheme (ships). *Weather and Forecasting*, 20(4):531–543, August 2005.

[40] H. Du. *Combining Statistical Methods with Dynamical Insight to Improve Nonlinear Estimation*. PhD thesis, London School of Economics and Political Science, 2009.

[41] H. Du and L. A. Smith. Parameter estimation through ignorance. *Physical Review E*, 86(1):016213, July 2012.

[42] P. K. Dunn and G. K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:236–244, 1996.

[43] J. B. Elsner and A. Birol Kara. *Hurricanes of the North Atlantic*. Oxford, 1999.

[44] J. B. Elsner and T. Jagger. A hierarchical bayesian approach to seasonal hurricane modeling. *Journal of Climate*, 17(14):2813–2827, July 2004.

[45] J. B. Elsner and T. Jagger. Comparison of hindcasts anticipating the 2004 florida hurricane season. *Weather and Forecasting*, 21(2):182–192, April 2006.

[46] J. B. Elsner and T. Jagger. Prediction models for annual u.s. hurricane counts. *Journal of Climate*, 19(12):2935–2952, June 2006.

[47] J. B. Elsner, J. P. Kossin, and T. H. Jagger. The increasing intensity of the strongest tropical cyclones. *Nature*, 455:92–95, September 2008.

[48] J. B. Elsner and T. H. Murnane, R. J. Jagger. Forecasting u.s. hurricanes 6 months in advance. *Geophysical Research Letters*, 33(10):L10704, May 2006.

[49] J. B. Elsner, X. Niu, and A. A. Tsonis. Multi-year prediction model of north atlantic hurricane activity.

[50] J. B. Elsner and C. P. Schmertmann. Improving extended-range seasonal predictions of intense atlantic hurricane activity. *Journal of Climate*, 8(3):345–351, September 1993.

[51] K. Emanuel. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, 436:686–688, August 2005.

[52] K. Emanuel. Environmental factors affecting tropical cyclone power dissipation. *Journal of Climate*, 20(22):5497–5509, November 2007.

[53] K. Emanuel. The hurricane-climate connection. *Bulletin of the American Meteorological Society*, 89:153–175, May 2008.

[54] K. Emanuel, F. Fondriest, and J. Kossin. Potential economic value of seasonal hurricane forecasts. *Weather, Climate, and Society*, 4(2):110–117, April 2012.

[55] C. A. T. Ferro. Comparing probabilistic forecasting systems with the brier score. *Weather and Forecasting*, 22(5):1076–1088, October 2007.

[56] J. J. Filliben. The probability plot correlation coefficient test for normality. *Techometrics*, 17(1):111–117, February 1975.

[57] C. for Ocean-Atmospheric Prediction Studies. Florida State University.

[58] A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2):1134–1140, February 1986.

[59] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, 69(2):243–268, April 2007.

[60] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. 102(477):359–378, January 2007.

[61] S. B. Goldenberg, C. W. Landsea, A. M. Mestas-Nunez, and W. M. Gray. The recent increase in atlantic hurricane activity: Causes and implications. *Science*, 293(474):474–479, July 2001.

[62] S. B. Goldenberg and L. J. Shapiro. Physical mechanisms for the association of el niño and west african rainfall with atlantic major hurricane activity. *Journal of Climate*, 9(6):1169–1187, June 1996.

[63] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society*, 14(1):107–114, 1952.

[64] W. M. Gray. Atlantic seasonal hurricane frequency. part i: El niño and 30mb quasi-biennial oscillation influences. *Monthly Weather Review*, 112(9):1649–1668, September 1984.

[65] W. M. Gray and P. J. Klotzbach. The tropical meteorology project. Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA, (1984-2013).

[66] R. Hagedorn, F. J. Doblas-Reyes, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting - i. basic concept. *Tellus A*, 57A(3):219–233, April 2005.

[67] R. Hagedorn and L. A. Smith. Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*, 16(2):143–155, October 2009.

[68] A. B. Hagen, Strahan-Sakoskie, and C. Luckett. A reanalysis of the 1944-53 atlantic hurricane seasons the first decade of aircraft reconnaissance. *Journal of Climate*, 25(13):4441–4460, July 2012.

[69] P. Hall, J. S. Marron, and B. U. Park. Smoothed cross-validation. *Probability Theory and Related Fields*, 92(1):1–20, 1992.

[70] T. Hall, H. E. Brooks, and C. A. Doswell III. Precipitation forecasting using a neural network. *Weather and Forecasting*, 14(3):338–345, June 1999.

[71] D. E. Hanley, M. A. Bourassa, J. J. O'Brien, S. R. Smith, and R. Spade, Elizabeth. A quantitative evaluation of enso indices. *Journal of Climate*, 16(8):1249–1258, April 2003.

[72] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman And Hall, 1990.

[73] T. J. Hastie and R. J. Tibshirani. *The Elements of Statistical Learning.* Springer, 2nd edition edition, 2009.

[74] J. P. Hess, J. B. Elsner, and N. E. Laseur. Improving seasonal hurricane predictions for the atlantic basin. *Weather and Forecasting*, 10(2):425–432, June 1995.

[75] K. W. Hipel. Geophysical model discrimination using the akaike information criterion. *IEEE Transactions on Automatic Control*, 26(2):358–378, April 1981.

[76] G. J. Holland and P. J. Webster. Heightened tropical cyclone activity in the north atlantic; natural variability or climate trend? *Phil. Trans. R. Soc. A*, 365(1860):2695–2716, November 2007.

[77] S. C. Hora. Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*, 50(5):597–604, May 2004.

[78] C. D. Hoyos, P. A. Agudelo, P. J. Webster, and J. A. Curry. Deconvolution of the factors contributing to the increase in global hurricane intensity. *Science*, 312:94–97, April 2006.

[79] A. S. Jarman. Distinguishing between skill and value in hurricane forecasting. Poster, presented at the European Geosciences Union General Assembly, 22-27 April 2012 EGU, 2012.

[80] B. R. Jarvinen, C. J. Neumann, and M. A. S. Davis. A tropical cyclone data tape for the north atlantic basin, 1888-1983: contents, limitations, and uses. Tech. Memo, NWS NHC 22, 1988.

[81] S. Jewson. Comparing the ensemble mean and the ensemble standard deviation as inputs for probabilistic medium-range temperature forecasts. *arXiv:physics*, arXiv:0310059, 2003.

[82] S. Jewson, E. Bellone, T. Laepple, K. Nzerem, S. hare, M. Lonfat, A. O'Shay, J. Penzer, and K. Coughlin. *Hurricanes and Climate Change. Chapter 5: Five Year Prediction of the Number of Hurricanes that make United States Landfall.* Springer US, 2009.

[83] S. Jewson, R. Binter, S. Khare, K. Nzerem, and A. O'Shay. Predicting basin and landfalling hurricane numbers from sea surface temperature. *arXiv:physics*, arXiv:0701170v2, 2007.

[84] S. R. Johnson and M. T. Holt. *Economic value of weather and climate forecasts. Chapter 3: The value of weather information.* Cambridge University Press, Cambridge, UK ; New York :, (1997).

[85] I. T. Joliffe. Uncertainty and inference for verification measures. *Weather and Forecasting*, 22(3):637–650, June 2007.

318

[86] I. T. Joliffe and D. B. Stephenson. Comments on discussion of verification concepts in forecast verification: A practitioners guide in atmospheric science. *Weather and Forecasting*, 20(5):796–800, October 2005.

[87] I. T. Jolliffe and D. B. Stephenson. *Forecast Verification - A Practitioner's Guide in Atmospheric Science. Chapter 1: Introduction.* John Wiley and Sons, (2003).

[88] K. Judd, C. A. Reynolds, T. E. Rosmond, and L. Smith. The geometry of model error. *Journal of the Atmospheric Sciences*, 65(6):1749–1772, June 2008.

[89] K. Judd and L. Smith. Indistinguishable states i. perfect model scenario. *Physica D*, 151:125–141, 2001.

[90] K. Judd and L. Smith. Indistinguishable states ii. imperfect model scenario. *Physica D*, 196:224–242, 2004.

[91] M. Kalos and P. Whitlock. *Monte Carlo Methods.* Wiley-VCH, 2nd edition, (2008).

[92] H. Kantz and T. Schreiber. *Nonlinear time series analysis.* Cambridge University Press, (1997).

[93] J. Kaplan and M. DeMaria. Large-scale characteristics of rapidly intensifying tropical cyclones in the north atlantic basin. *Weather and Forecasting*, 18(6):1093–1108, December 2003.

[94] R. W. Katz and A. H. Murphy. *Economic value of weather and climate forecasts. Chapter 1: Introduction.* Cambridge University Press, Cambridge, UK ; New York :, (1997).

[95] J. L. Kelly. A new interpretation of information rate. *Bell Systems Tech.*, 35:917–926, 1956.

[96] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. 63(3):425–464, January 2001.

[97] P. J. Klotzbach. Revised prediction of seasonal atlantic basin tropical cyclone activity from 1 august. *Weather and Forecasting*, 22(5):937–949, October 2007.

[98] P. J. Klotzbach and W. M. Gray. Multidecadal variability in north atlantic tropical cyclone activity. *Journal of Climate*, 21(15):3929–3935, August 2008.

[99] T. R. Knutson, J. L. Mcbride, J. Chan, K. Emanuel, G. Holland, C. Landsea, I. Held, J. P. Kossin, A. K. Srivastava, and M. Sugi. Tropical cyclones and climate change. *Nature Geoscience*, 3:157–163, March 2010.

[100] T. R. Knutson, S. T. Sirutis, Joseph J.and Garner, G. A. Vecchi, and I. M. Held. Simulated reduction in atlantic hurricane frequency under twenty-first-century warming conditions. *Nature Geoscience*, 1(6):359–364, June 2008.

[101] J. P. Kossin, S. J. Camargo, and M. Sitkowski. Climate modulation of north atlantic hurricane tracks. *Journal of Climate*, 23(11):3057–3076, June 2010.

[102] J. P. Kossin, K. R. Knapp, D. J. Vimont, R. J. Murnane, and B. A. Harper. A globally consistent reanalysis of hurricane variability and trends. *Geophysical Research Letters*, 34(4):L04815, February 2007.

[103] J. P. Kossin and D. J. Vimont. A more general framework for understanding atlantic hurricane variability and trends. *Bulletin of the American Meteorological Society*, 88(11):1767–1781, November 2007.

[104] H. C. Kunreuther and E. O. Michel-Kerjan. *At War with the Weather*. Massachusetts Institue of Technology, (2009).

[105] C. W. Landsea. Can we detect trends in extreme tropical cyclones? *Science*, 313(5786):452–454, July 2006.

[106] C. W. Landsea. Counting atlantic tropical cyclones back to 1900. *Journal of Climate*, 88(18):197–208, May 2007.

[107] C. W. Landsea and co authors. A reanalysis of the 1911-20 atlantic hurricane database. *Journal of Climate*, 21(10):2138–2168, May 2004.

[108] C. W. Landsea and co authors. A reanalysis of the 1921-1930 atlantic hurricane database. *Journal of Climate*, 25(3):865–885, February 2012.

[109] C. W. Landsea, co-authors: The Atlantic Basin The Atlantic Hurricane Database Re-Analysis Project: Documentation for the 1851-1910 alterations, and additions to the HURDAT database. *Hurricanes and Typhoons: Past, Present and Future*. Columbia University Press, (2004).

[110] J. B. Lang. Score and profile likelihood confidence intervals for contingency table parameters. *Statistics in Medicine*, 27(28):5975–5990, December 2008.

[111] C. E. Leith. The standard error of time-average estimates of climatic means. *Journal of Applied Meteorology and Climatology*, 12(6):1066–1069, September 1973.

[112] D. Letson, D. Sutter, and J. K. Lazo. The economic value of hurricane forecasts: an overview and research needs. *Natural Hazards Review*, 8(3):78–86, August 2007.

[113] L.-Y. Leung and G. R. North. Information theory and climate prediction. *Journal of Climate*, 3(1):5–14, January 1990.

[114] M. Leutbecher and T. N. Palmer. Ensemble forecasting. *Journal of Computational Physics*, 227(7):3515–3539, March 2007.

[115] D. V. Lindley. *Making Decisions*. John Wiley and Sons, 2nd edition, (1985).

[116] Lloyd's. Forecasting risk: the value of long-range forecasting for the insurance industry, (2011).

[117] M. Lonfat, A. Boissonnade, and R. Muir-Wood. Atlantic basin, u.s. and caribbean landfall activity rates over the 2006-2010 period: an insurance industry perspective. 59(4):499–510, August 2007.

[118] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, March 1963.

[119] H. T. M. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, 14(2):155–167, April 1999.

[120] H. T. M. and J. S. Whitaker. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, 134(11):3209–3229, November 2006.

[121] L. Magee. Nonlocal behavior of polynomial regressions. *The American Statistician*, 52(1):20–22, February 1998.

[122] M. Mann and K. Emanuel. Atlantic hurricane trends linked to climate change. *Eos, Transactions, American Geophysical Union*, 87(24):233–244, June 2006.

[123] M. Mann, T. A. Sabbatelli, and U. Neu. Evidence for a modest undercount bias in early historical atlantic tropical cyclone counts. *Geophysical Research Letters*, 34(22):L22707, November 2007.

[124] I. Mason. A model for assessment of weather forecasts. *Australian Meteorological Magazine*, 30(4):291–303, 1982.

[125] S. J. Mason and O. Baddour. *Seasonal Climate: Forecasting and Managing Risk. Chapter 7: Statistical Modelling*, volume 82 of *Nato Science Series*. Springer Academic, (2008).

[126] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261:459–467, June 1976.

[127] R. M. Mazo. *Brownian Motion, Fluctuations, Dynamics and Applications*. International Series of Monographs on Physics. Oxford University Press, (2002).

[128] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman And Hall, 2nd edition, 1989.

[129] O. Mestre and S. Hallegate. Predictors of tropical cyclone numbers and extreme hurricane intensities over the north atlantic using generalized additive and linear models. *Journal of Climate*, 22(3):633–648, February 2009.

[130] A. P. Morse, J. Doblas-Reyes, Francisco, M. B. Hoshen, R. Hagedorn, and T. Palmer. A forecast quality assessment of an end-to-end probabilistic multi-model seasonal forecast system using a malaria model. *Tellus A*, 57(3):464–475, May 2005.

[131] S. L. Mullen and R. Buizza. Quantitative precipitation forecasts over the united states by the ecmwf ensemble prediction system. *Monthly Weather Review*, 129(4):638–663, April 2001.

[132] A. H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, June 1973.

[133] A. H. Murphy. A sample skill score for probability forecasts. *Monthly Weather Review*, 102(1):48–55, January 1974.

[134] A. H. Murphy. Forecast verification: Its complexity and dimensionality. *Monthly Weather Review*, 119(7):1590–1601, 1991.

[135] A. H. Murphy. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8(2):281–293, 1993.

[136] A. H. Murphy. General decompositions of mse-based skill score: Measures of some basic aspects of forecast quality. *Monthly Weather Review*, 124(10):2353–2369, October 1996.

[137] A. H. Murphy. *Economic value of weather and climate forecasts. Chapter 2: Forecast Verification.* Cambridge University Press, Cambridge, UK ; New York :, (1997).

[138] A. H. Murphy. The early history of probability forecasts: Some extensions and clarifications. *Weather and Forecasting*, 13(1):5–15, March 1998.

[139] A. H. Murphy and M. Ehrendorfer. On the relationship between accuracy and value of forecasts. *Weather and Forecasting*, 2(3):243–251, Septmber 1987.

[140] A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society*, 26(1):41–47, 1977.

[141] A. H. Murphy and R. L. Winkler. Probability forecasting in meterology. 79(387):489–500, September 1984.

[142] A. H. Murphy and R. L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338, 1987.

[143] N. O. National Hurricane Center and A. Administration.

[144] N. O. National Hurricane Center and A. Administration. 2013 national hurricane center forecast verification report, (2013).

[145] J. Neyman. On the problem of confidence intervals. *The Annals of Mathematical Statistics*, 6(3):111–116, 1935.

[146] N. Nicholls. commentary and analysis: The insignificance of significance testing. *Bulletin of the American Meteorological Society*, 82(5):981–986, May 2001.

[147] M. Office. North atlantic tropical storm seasonal forecast. Met Office, Exeter, UK [Available online at http://wwwpre.metoffice.gov.uk/weather/tropicalcyclone/northatlantic.html].

[148] I. P. on Climate Change. Managing the risks of extreme events and disasters to advance climate change adaptation (srex). Special Report of the IPCC, 2012.

[149] B. F. Owens and C. W. Landsea. Assessing the skill of operational atlantic seasonal tropical cyclone forecasts. *Weather and Forecasting*, 18(1):45–54, February 2003.

[150] T. N. Palmer. Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, 63:71–116, 2000.

[151] T. N. Palmer, F. J. Doblas-Reyes, R. Hagedorn, and A. Weisheimer. Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Phil. Trans. R. Soc. B*, 360(1463):1991–1998, November 2005.

[152] T. N. Palmer, F. J. Doblas-Reyes, A. Weisheimer, and M. Rodwell. Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bulletin of the American Meteorological Society*, 89:459–470, April 2008.

[153] T. N. Palmer and P. D. Williams. Introduction. stochastic physics and climate modelling. *Phil. Trans. R. Soc. A*, 366(1875):2149–2425, July 2008.

[154] R. Peirolo. Information gain as a score for probabilistic forecasts. *Meteorological Applications*, 18(1):9–17, March 2011.

[155] R. R. Picard and R. Dennis Cook. Cross-validation of regression models. 79(387):575–583, 1984.

[156] R. A. Pielke. United states hurricane landfalls and damages: Can one- to five-year predictions beat climatology? *Environmental Hazards*, 8:187–200(14), September 2009.

[157] R. A. Pielke, J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, and R. Musulin. Normalized hurricane damage in the united states: 1900-2005. *Natural Hazards Review*, 9(1):29–42, 2008.

[158] R. A. Pielke and C. W. Landsea. Normalized hurricane damages in the united states: 1925-95. *Weather and Forecasting*, 13:621–631, September 1998.

[159] P. Pinson, P. McSharry, and H. Madsena. Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society*, 136(646):77–90, January 2010.

[160] W. H. Press, S. A. Tuekolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition, (2007).

[161] C. Primo, C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson. Calibration of probabilistic forecasts of binary events. *Monthly Weather Review*, 137(3):1142–1149, March 2009.

[162] U. C. C. S. Program. Weather and climate extremes in a changing climate. Synthesis and Assessment Product 3.3, June (2008).

[163] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, May 2005.

[164] N. Ranger and F. Nierhoerster. Deep uncertainty in long-term hurricane risk: Scenario generation and implications for future climate experiments. *Global Environmental Change*, 22(3):703–712, August 2012.

[165] R. Ranjan and T. Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society*, 72(1):71–91, January 2010.

[166] E. N. Rappaport and co authors. Advances and challenges at the national hurricane center. *Weather and Forecasting*, 24(2):395–419, April 2009.

[167] M. Re. Topics geo: Natural catastrophes 2011, analyses, assessments, positions, (2012).

[168] M. Re. Topics geo: Natural catastrophes 2012, analyses, assessments, positions, (2013).

[169] D. S. Richardson. Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563):649–667, January 2000.

[170] D. S. Richardson. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127(577):2473–2489, October 2001.

[171] P. J. Roebber and L. F. Bosart. The complex relationship between forecast skill and forecast value: A real-world analysis. *Weather and Forecasting*, 11(4):544–559, 1996.

[172] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660, 2002.

[173] M. S. Roulston and L. A. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55A:16–30, 2003.

[174] P. Royston and A. D. G. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, 43(3):429–467, 1994.

[175] A. B. Schumacher, M. DeMaria, and J. A. Knaff. Objective estimation of the 24-h probability of tropical cyclone formation. *Weather and Forecasting*, 24(2):456–471, April 2009.

[176] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[177] R. Seaman, I. Mason, and F. Woodcock. Confidence intervals for some performance measures of yes/no forecasts. *Australian Meteorological Magazine*, 45:49–53, 1996.

[178] L. J. Shapiro and S. B. Goldenberg. Atlantic sea surface temperatures and tropical cyclone formation. *Journal of Climate*, 11(4):578–590, April 1998.

[179] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, (1986).

[180] J. S. Simonoff. *Analyzing Categorical Data*. Springer-Verlag, (2003).

[181] D. M. Smith, R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife. Skilful multi-year predictions of atlantic hurricane frequency. *Nature Geoscience*, 3(12):846–849, November 2010.

[182] L. Smith. Disentangling uncertainty and error: On the predictability of nonlinear systems. In A. I. Mees, editor, *Nonlinear Dynamics and Statistics*, pages 31–64. Birkhuser Boston, 2001.

[183] L. A. Smith. The maintenance of uncertainty. In Proc. International School of Physics Enrico Fermi, Course CXXXIII, (1997).

[184] L. A. Smith. *Chaos: A very short introduction*. Oxford University Press, (2007).

[185] L. A. Smith. Guidance, information or probability forecast: Where do ensembles aim? Presentation, International Conference on Ensemble Methods in Geophysical Sciences, Toulouse, France, 12-16 November 2012, (2012).

[186] S. R. Smith, J. Brolley, J. J. O'Brien, and C. Tartaglione. Ensos impact on regional u.s. hurricane activity. *Journal of Climate*, 20(7):1404–1414, April 2007.

[187] T. M. Smith, R. W. Reynolds, T. C. Peterson, and J. Lawrimore. Improvements to noaas historical merged land-ocean surface temperature analysis (1880-2006). *Journal of Climate*, 21(10):2283–2296, May 2008.

[188] J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford university Press, 1st edition, (2003).

[189] P. B. Stark. The law of large numbers. Department of Statistics, Berkley College, (2005).

[190] D. M. Stasinoupos and R. A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of the Royal Statistical Society*, C54(3):507–554, June 2005.

[191] D. Stephenson. *Seasonal Climate: Forecasting and Managing Risk. Chapter 9: An introduction to probability forecasting*, volume 82 of *Nato Science Series*. Springer Academic, (2008).

[192] H. J. Thiébaux and F. W. Zwiers. The interpretation and estimation of effective sample size. *Journal of Applied Meteorology and Climatology*, 23(5):800–811, May 1984.

[193] J. C. Thompson and G. W. Brier. The economic utility of weather forecasts. *Monthly Weather Review*, 83(11):249–253, November 1955.

[194] J. Tödter and B. Ahrens. Generalization of the ignorance score: Continuous ranked version and its decomposition. *Monthly Weather Review*, 140(6):2005–2017, June 2012.

[195] Z. Toth. Long-range weather forecasting using an analog approach. *Journal of Climate*, 2(6):594–607, June 1989.

[196] Z. Toth, O. Talagrand, G. Candille, and Y. Zhu. *Forecast Verification - A Practitioner's Guide in Atmospheric Science. Chapter 7: Probability and ensemble forecasts*. John Wiley and Sons, (2003).

[197] K. E. Trenberth. Some effects of finite sample size and persistence on meteorological statistics. part i: Autocorrelations. *Monthly Weather Review*, 112(12):2359–2368, December 1984.

[198] T. S. R. (TSR). http://www.tropicalstormrisk.com/index_.html.

[199] S. V. Weijs, R. van Nooijen, and N. van de Giesen. Kullback-leibler divergence as a forecast skill score with classic

reliability-resolution-uncertainty decomposition. *Monthly Weather Review*, 138(9):3387–3399, September 2010.

[200] H. M. Van Den Dool. A new look at weather forecasting through analogues. *Monthly Weather Review*, 117(10):2230–2247, October 1989.

[201] G. A. Vecchi and T. R. Knutson. On estimates of historical north atlantic tropical cyclone activity. *Journal of Climate*, 21(14):3580–3600, 2008.

[202] G. A. Vecchi and T. R. Knutson. Estimating annual numbers of atlantic hurricanes missing from the hurdat database (1878-1965) using ship track density. *Journal of Climate*, 24(6):1736–1746, March 2011.

[203] G. A. Vecchi and B. J. Soden. Effect of remote sea surface temperature change on tropical cyclone potential intensity. *Nature*, 450:1066–1070, December 2007.

[204] G. A. Vecchi, K. L. Swanson, and B. J. Soden. Whither hurricane activity? *Science*, 322:687–389, October 2008.

[205] G. A. Vecchi, M. Zhao, H. Wang, G. Villarini, A. Rosati, A. Kumar, I. M. Held, and R. Gudgel. Statistical-dynamical predictions of seasonal north atlantic hurricane activity. *Monthly Weather Review*, 139(4):1070–1082, April 2011.

[206] G. Villarini and G. A. Vecchi. North atlantic power dissipation index (pdi) and accumulated cyclone energy (ace): Statistical modeling and sensitivity to sea surface temperature changes. *Journal of Climate*, 25(2):625–637, January 2012.

[207] G. Villarini, G. A. Vecchi, and J. A. Smith. Modeling the dependence of tropical storm counts in the north atlantic basin on climate indices. *Monthly Weather Review*, 138(7):2681–2705, July 2010.

[208] G. Villarini, G. A. Vecchi, and J. A. Smith. U.s. landfalling and north atlantic hurricanes: Statistical modelling of their frequencies and ratios. *Monthly Weather Review*, 140(1):44–65, January 2012.

[209] J. Von Neumann and O. Morgenstern. *Theory of games and economic behaviour.* Princeton University Press, (1947).

[210] P. J. Webster, G. J. Holland, J. A. Curry, and H.-R. Chang. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, 309:1844–1846, September 2005.

[211] P. J. Webster and C. D. Hoyos. Beyond the spring barrier? *Nature Geoscience*, 3(3):152–153, March 2010.

[212] Wikipedia. Butterfly effect. http://en.wikipedia.org/wiki/Butterfly_effect.

[213] D. S. Wilks. Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models. *Monthly Weather Review*, 119(7):1640–1662, July 1991.

[214] D. S. Wilks. Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128(586):2821–2836, October 2002.

[215] D. S. Wilks. Comparison of ensemble-mos methods in the lorenz 96 setting. *Meteorological Applications*, 13(3):243–256, January 2006.

[216] D. S. Wilks. Sampling distributions of the brier score and brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 136(653):2109–2118, October 2010.

[217] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*, volume 100 of *International Geophysics*. Academic Press, 3rd edition, (2011).

[218] D. S. Wilks and T. M. Hamill. Comparison of ensemble-mos methods using gfs reforecasts. *Monthly Weather Review*, 135(6):2379–2390, June 2007.

[219] G. U. Yule. On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Phil. Trans. R. Soc. A*, 226:267–298, 1927.