The London School of Economics and Political Science

Extreme Insurance and the Dynamics of Risk

TREVOR J MAYNARD

London, April 4, 2016

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

Acknowledgements

I would like to express my gratitude to my employer, Lloyd's of London, who funded the first part of my research. I hope the findings in this thesis can repay that debt with some interesting insights into the insurance industry. I would also like to thank the following people:

- Tom Bolt, my boss, for his support and Paul Nunn who originally enthusiastically supported my request to commence writing a thesis.
- David Simmons and Nigel Ralph for considering my questions about common practices in insurance/ reinsurance markets; your experience was invaluable.
- Ed Wheatcroft and Hailiang Du, thank you for reading my thesis prior to submission and for useful comments made.
- My Supervisor, Professor Leonard Smith: thank you for the great questions you have asked me to consider along the way and also for your help and advice.
- Finally, to my wife Rhona and children Hannah and Rebecca, thank you for your support and belief and for putting up with a sometimes distracted husband/father.

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 consider the work submitted to be a complete thesis fit for examination.

I authorise that, if a degree is awarded, an electronic copy of my thesis will be deposited in LSE Theses Online held by the British Library of Political and Economic Science and that, except as provided for in regulation 41 it will be made available for public reference.

I authorise the School to supply a copy of the abstract of my thesis for inclusion in any published list of theses offered for higher degrees in British universities or in any supplement thereto, or for consultation in any central file of abstracts of such theses.

I declare that my thesis consists of less than 100,000 words.

Abstract

The aim of this thesis is to explore the question: can scientific models improve insurance pricing? Model outputs are often converted to forecasts and, in the context of insurance, the supplementary questions: 'are forecasts skillful?' and 'are forecasts useful?' are examined.

Skill score comparison experiments are developed allowing several scores in common use to be ranked. One score is shown to perform well; several others are shown to have systematic failings; with the conclusion that these should not be used by insurers. A new skill score property 'Feasibility' is proposed which highlights a key shortcoming of some scores in common use. Variables from a well known dynamical system are used as a proxy for an insurable index. A new method relating the system and its models is presented using skill scores to find their score optimal piecewise linear relationship. The index is priced using both traditional techniques and new methods that use the score optimal relationship. One new method is very successful in that it produces lower prices on average, is more profitable and leads to a lower probability of insurer failure. In this context the forecasts are both skilful and useful. The efficacy of forecast use is further explored by considering hurricane insurance. Here forecasts are shown to be useful only if very simple adjustments to pricing are made. A novel agent based model of a two company insurance industry containing many key features in the real world is presented enabling the impact of regulation and competition to be assessed. Several common practices are shown to reduce expected company lifetime.

Contents

List of Figures

List of Tables

List of Variable Names

The following table lists the main variables that are used within this thesis. Some variables are not listed here if they are used briefly in a chapter with no lasting meaning. The column 'variable name' is used for chapter 6 to enable easy translation of axes labels on the graphics. Where appropriate equation numbers are also given.

equation number	**mathematical notation**	**description**	**chapter**
2.1	$S(p, v)$	skill score for forecast p and observation v	2
	p, q	probabilistic forecasts	2
2.37	f_u	underlying distribution created by kernel dressing a set of data points	2
2.36	$\hat{\theta}$	score optimal parameter value	2
2.55	f_{w_1,w_2,w_3}	weighted average of Gamma, Lognormal and Pareto distributions, with weights w_i	2
2.54	G_S	Skill Gap when skill score S is used	2
2.62	$r_n(\alpha, \sigma)$	Climatology blended forecast with blending parameter α and kernel width σ	2
3.2	X_k	Lorenz 96 'slow' variable, with index $k \sim \{1, ..36\}$	3
3.2	$Y_{k,j}$	Lorenz 96 'fast' variable associated with slow variable X_k and with index $j \sim \{1, ...10\}$	3
3.3	IEF_k	Instantaneous Effective Forcing within Lorenz full system, associated with variable X_k	3
Page 200	f_k	transformation of observed system variable to decision relevant quantity	4
4.1	R	insurance index based on observed system variables	4
4.3	λ	relative frequency of ensemble members falling within a chosen interval	4
Page 201	Ω	space of values taken by insurance index R in the full system	4

Page 202	Θ	space of values take by the insurance index $\hat{R}$ in the model	4
4.18	D	Ground Up Loss	4
4.19	L	Gross Loss	4
Page 253	N_B	Number of hurricanes generated in the atlantic basin	5
Page 253	N_L	Number of hurricanes that make landfall	5
Page 253	N_C	Number of hurricanes that hit an urban or commercial centre	5
Page 254	sa	Saffir Simpson hurricane strength category (1-5)	5
Page 254	$S(sa)$	Insurance industry loss as a function of hurricane strength	5
5.6	$f(n_B)$	Low fidelity forecast of number of basin storms into category high, medium and low	5
5.5	P_0	Price for hurricane insurance using climatology information only	5
5.8	P_2	Price for hurricane insurance based on approximate season frequency forecast $f(n_B)$	5
5.9	P_3	Price for hurricane insurance using perfect forecast of number of basin storms N_B	5
5.10	P_4	Price for hurricane insurance using perfect forecast of number of landfalling storms N_L	5
5.12	P_5	Price for hurricane insurance based on an adjustment to P_4 which uses approximate season severity forecast	5
	t	index for the simulated year of business	6
6.2 , 6.22	C_t^{main}	claims in year per unit of risk for 'main ' company; definition for competitor (superscript 'comp') is not shown.	6
6.3	P_t^{main}	premium charged per unit of risk	6
	$e^{-\mu}\hat{E}(C_t^{main})$	expected claims discounted to start of year	6
6.5	$e^{-\mu}$payback	portion of premium relating to 'payback'	6
6.8	$e^{-\mu}\hat{K}^{main}$	capital required per unit of risk	6
6.16, 6.30	N_t^{main}	number of policyholders within the market (since all of equal risk)	6
	$e^{-\mu}$	discount factor at the risk free rate where μ is the force of interest	6
	e^{μ}	accumulation factor at the risk free rate	6
6.1	$\mathbb{C}_t^{main}$	main company's share of industry claims	6

6.20	$\mathbb{P}_t^{main}$	total premium written by main company (for all policies)	6
6.23	$\mathbb{I}_t^{main}$	investment return in the year from all sources (for profit and loss account)	6
6.24	π_t^{main}	profit in year for main company	6
6.25	$\mathbb{D}_t^{main}$	dividend paid by main company to its shareholders, in year t	6
6.29	$\mathbb{K}_t^{main}$	capital required by main company at Beginning Of Year (BOY)	6
6.21	$\mathbb{J}_t^{main}$	capital injection in year t for main company	6
	$\mathbb{K}_{t+1}^{main}$	capital held at the End Of Year	6
	$\mathbb{K}_t^{main} + 1 > 0$	boolean test to check whether the main company is alive or dead at the end of the year	6
6.6	VaR	Value at Risk (see Glossary)	6
6.7	$TVaR$	Tail Value at Risk (see Glossary)	6

Glossary of terms

Term	Definition
Additive Observational Noise	A time series x_t can be thought of as a series of draws from a random variable X_t with mean μ_t. The differences $e_t = X_t - \mu_t$ are then a random variable with zero mean. The terms e_t are often called innovations, error terms, white noise or dynamic noise - here they will be called Additive Observational Noise. They are often assumed to have a Gaussian distribution but this is not a necessary condition.
Attractor	Let f be a map. An attractor is 'the set of points to which most points evolve under iterates of f' (Milnor 1985, quoting Collet and Eckmann 1980).
Capital	Money held against the possibility that the premium charged is insufficient to pay for expenses and any insurance claims during the period the insurance is in force.
Categorical forecast	A probabilistic forecast where the possible outcomes are discrete items.
Claim	The monetary amount which the insurer will pay to a policyholder if an insurable event has occurred
Climatology	A climatology is an empirical distribution based on past observations over a defined period of time
Climatology blending	A weighted average between a probabilistic forecast and a climatology forecast. The blending parameters can include the weight variable and also any parameters of the forecast (such as the kernel width if kernel dressing is used). These can chosen to optimise a particular skill score for a give set of observations.
Convective event	Atmospheric event such as thunderstorm or tornado.
Exceedance probability	The exceedance probability (EP) curve is commonly used in general insurance. It denotes the probability that a loss will be greater than or equal to a given amount. If the Cumulative Distribution Function (CDF) of insurance losses is F then the EP curve is $1 - F$.
Fahrenheit (F)	A temperature scale where water freezes at 32 degrees and boils at 212 degrees. One hundred and five degrees fahrenheit is expressed as 105F.

Forecast A prediction of the value of an observed variable at some point in the future.

Free assets Insurance jargon: investments held by an insurer in excess of reserves and any statutory capital. In theory this money could be paid back to shareholders and the insurer would remain solvent and able to trade; in practice rating agencies typically expect free assets to be greater than zero in order to grant a BBB or above rating.

Gamma A continuous probability distribution with pdf $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ with $\alpha, \beta > 0$ and support $= (0, \infty)$

GBP Great British Pounds - currency in the UK; also known as 'Sterling'.

Gross Loss (GL) An measure of insurance loss after allowance for terms and conditions such as deductibles and limits- but before the deduction of any reinsurance.

Ground Up Loss (GUL) Financial damages following an insured event before the application of any deductibles or other relevant terms and conditions. The deductible is paid by the insured (also sometimes called the excess)

Hazard A class of event that has the potential to cause economic or insured loss (or injury, loss of life). Examples include: Earthquakes, Hurricanes and Flooding.

IID Independent and Identically Distributed (IID). Two events are independent if their joint probability is the product of their marginal probabilities. Two random variables are IID if they are independent and have the same marginal distribution.

Insolvency An insurer/reinsurer is insolvent if its regulator decides that it is no longer capable of trading. There are various triggers that would lead to this decision, some qualitative and some quantitative. If a firm's reserves and minimum capital requirements exceed the assets available this would be a key trigger.

Kernel Dressing A method to create a probabilistic forecast from an ensemble of forecast outcomes. Formed by the summation of kernel's which are normalised probability distributions each with mean equal to the observation and a chosen kernel width.

Lognormal A continuous probability distribution, with support $= (0, \infty)$. If $X \sim N(\mu, \sigma^2)$ then $Y = e^X$ is lognormally distributed

Long tailed business Policies where it can take a number of years to establish whether a claim has occurred. For example liability policies where a court case can take many years to make a final ruling to determine who is liable.

Losses Insurance jargon: The claims paid out following an insurable event are sometimes referred to as losses by insurers. This is an unfortunate phrase when compared to general usage of the word because insurance losses may still be profitable if premium rates were adequate.

Model A collection of mathematical expressions that attempt to describe the key features of a system.

Normal A continuous probability distribution with pdf $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\mu-x)^2}{2\sigma^2}}$, for $\sigma > 0$, with support $= (-\infty, \infty)$. Also described as a Gaussian Distribution.

Observation	A data point that can be used to evaluate a forecast.
Optimum score estimation	The process of finding the parameters of a forecast that lead to best possible skill score given a series of observations.
Pareto	A continuous probability distribution with pdf $f(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}$, for $\alpha, \beta > 0$, support $= (\beta, \infty)$
Premium	The amount paid to the insurer for the provision of insurance - this is the price of insurance.
Probabilistic forecast	A forecast where the value of the variable is not predicted deterministically - but is described by a distribution function that can either be discrete or continuous
Probability Forecast	A probabilistic forecast whose support is a continuum.
Rating Agency	Examples include Standard and Poor's, AM Best, Fitch and Moody's. These firms consider financial information along with qualitative management information in order to give a rating (i.e. comparative score) for different firms. They rate many financial firms/ assets including credit ratings and financial strength ratings. Each firm has its own rating scales - but these are broadly compatible. Ratings such as AAA indicate the highest quality assets.
Regulator	Insurance regulators oversee the insurance industry in their country. They impose restrictions on types of business that can be covered and have various reporting and capital requirements. Examples include the Prudential Regulation Authority in the UK and OSFI, the Office for the Superintendent of Financial Institutions, in Canada.
Reinsurance	Insurance purchased by insurers to offset specified insurance claims.
Reinsurance to close	Specific reinsurance bought in the Lloyd's market to transfer all the business from one syndicate into another.
Relative ignorance	The difference between the average skill score for a given forecast and the skill score obtained though using climatology as the forecast
Reserves	Reserves are held to pay insurance claims in the future. They are either (1) unexpired risk reserves, money held against risks that may arise in a future period of time for policies on-risk; (2) case reserves, money held against known insurable events that have been advised to the insurer by the policyholder and (3) Incurred but not reported reserves IBNR, statistical reserves held against events that are likely to have occurred already but which have not yet been advised to the insurer.
Run off	Insurance jargon: A firm is in 'run off' if it ceases to write new business and continues to trade until any residual claims are paid. Modelling the run off means letting the computer model run until all claims are settled.

Short tailed business	Business that is not 'long tailed' typically includes policies which cover damages that are quickly identified such as property damage.
Skill	A measure of the quality of a forecast, defined fully in chapter 2 (2.1)
Skill Gap	The sum over many observations of the difference between the skill score for a given observation/forecast pair and the expected skill score given the forecast
Statutory minimum capital	Regulators around the world will specify the method that insurers have to use to calculate the minimum amount of capital that must be held.
Syndicate	Type of insurance vehicle at Lloyd's of London.
Sum Insured	Insurance jargon: this is the monetary value of the insured item.
System	A set of processes, either real or artificial, that produce a number of observable variables.
Time Series	A series of Observations $X = \{x_1, ...x_N\}$ made at times $T = \{t_1, ...t_N\}$. The differences $d_i = t_i - t_{i+1}$ may not all be equal. A time series will be denoted $\{X, T\}$. N is described as the 'length' of the time series.
TVaR	Tail Value at Risk - the expected value of a variable given that it is in excess of the Value at Risk for a specified probability
USD bn	Billions of United States Dollars
VaR	Value at Risk at probability p. A term used in banking and insurance to indicate the value (K) of a financial random variable (X) such that $P(X > K) < p$.
$X \sim N(0, 1)$	A random variable X that has the standard unit normal distribution.

Chapter 1

The insurance industry and the problem of extremes

'***Critics suggest that financial models were ineffective ....[highlighting] the need for a close look at the role and effectiveness of financial models and the expertise of modellers ...A stubborn disregard for the dynamic reality of market relationships will lead to poor outcomes. A reliance on the persistence of historical relationships is dangerous....Those who seek an interdisciplinary, multidimensional approach to modelling the dynamics of markets and the macroeconomy are more likely to prove successful.***'

Financial models useful but limited (Financial Times) Sullivan 2011 [246]

The discipline of statistics is vital to and deeply embedded in the insurance industry. Its methods are used in analysis of past claims data to calculate premium rates and to determine the funds required to survive adverse events be they man made or natural [3, 7, 18, 141]. Statistical methods are used in combination with those of engineering and physical sciences to create catastrophe models [76] that are now widely used by insurers [132, 159].

The insurance industry aims to provide protection for society against all manner of hazards including: earthquakes, the costs of fighting legal action and paying damages, compensation and physical damage in marine and aviation disasters and terrorism [138, 175]. A number of these hazards relate to atmospheric perils such as flooding, tropical cyclones or hailstorms [248]. The scientific community have

developed models to explore the range of potential outcomes and to try to make predictions. An important question insurers are now asking: **can scientific models improve insurance pricing?** [60, 139]

This thesis explores that question. First the concept of 'skill' is considered in detail in Chapter 2. The chapter lists key skill scores that are used in practice and then carries out a series of experiments to determine which perform well in various circumstances. One score, I J Good's, 'Ignorance' [20, 29, 97], is shown to perform well in most circumstances; several others are shown to have systematic failings; arguably these should not be used by insurers. It is important to know when a model produces forecasts that are statistically unlikely. A new method is introduced, which compares the running average observed skill score to the distribution of the expected score allowing model rejection[1] once observed values are outside of expected confidence intervals.

An ideal system-model pair consists of a dynamical system (which may arise in the real world or be a stated set of mathematical relationships) and mathematical approximations, called 'models' which attempt to capture one or more critical features of the system. Models can take many forms, they may be statistical, physical, or based on coupled partial or ordinary differential equations. The initial conditions of such models can be set to approximate the current state; the ensuing behaviour of the model is then a 'forecast' of the system. A range of model parameters can be trialled each, likely, leading to different model behaviour. Such a collection of model outputs is called an 'ensemble' from which probability forecasts can be created.

A collection of system-model pairs, based on models introduced by Lorenz [155], are introduced and explored in Chapter 3. The skill of forecasts arising from the models is assessed using the Ignorance score and this is shown to be improved by mixing with the empirical distribution of past observations, a process known as Climatology Blending [30]. Chapter 4 then uses these system-model pairs to explore whether such models can be useful in an insurance context. An artificial hazard is created from system variables and an insurance index is created where the payout depends on the value of the index (as with some catastrophe bonds [13]). The

[1] 'Rejected' is used here to imply that a forecast system has been shown to be inconsistent with the distribution from which the observations are drawn. This does not imply that the forecasts contain no useful information and, if they do, they may still be used in practice.

model is not the system, however, and there is no a-priori reason why model values should correspond precisely to those of the system [243, 245]. A mapping between the model variables and the system is therefore required; the approach presented (called ϕ-transformation) is thought to be a new approach. Traditional insurance methods [128] are used to price the index and also two approaches using the ϕ-transformed forecasts. The results are encouraging for one model: prices are on average 10% lower, yet the insurer becomes 16% more profitable and goes insolvent with one fifth of the frequency.

Chapter 5 introduces a different method to that described in Chapter 4 to explore hurricane risk [148, 208]. A simple model of the hurricane process is developed with forecasts that become increasingly accurate - then pricing methods are considered that use the forecasts. One business volume scaling method results in a 7.9% increase in profits and a 9.1% reduction in negative profits. The other, more sophisticated, pricing methods, however, cause a reduction in profitability leading to the conclusion that forecasts are useful only if they are used carefully.

Whilst the Lorenz 96 system is an interesting test bed [103] the insurance setting in Chapter 4 remains highly simplified. Chapter 5 introduces alternative pricing methods but still avoids consideration of the complexity of the insurance industry. Chapter 6 moves closer to reality by introducing a novel agent-based model of an insurance industry with two competitors; complications such as dividend payment, payback rules and customer loyalty are included. A variety of real world inspired regulatory regimes, pricing methods and claims processes are tested leading to some unexpected results relative to a control experiment. For example, a traditional pricing method is shown to reduce company lifetime by 10% and company value by 2%; conversely a company that uses the wrong underlying claims distribution has a 50% longer lifetime and 6% higher company value despite having a 3.5% lower market share of business. This new model therefore highlights that competitive effects impose strong constraints on what can be achieved in practice through regulation and pricing.

In short, the question of how to determine whether scientific models can provide decision relevant information to insurers is considered throughout this thesis. To provide effective decision-relevant information models must be skillful [29] and ideally

that skill must be demonstrated quantitatively and coherently (Chapter 2). Second, the insurance relevant aspects of the model must be isolated, interpreted and deployed; methods to do this are illustrated by analogy (using Lorenz 96) in Chapters 3 and 4. Hurricane forecasts are ubiquitous in the insurance sector [126, 221, 260], the limited utility of the targets of today's forecasts are noted in Chapter 5, making them of limited use even if they had skill beyond climatology. The complexities of a simplified insurance market using a more traditional actuarial approach are illustrated with a novel agent based model in Chapter 6. The findings and relevance to insurers are summarised and discussed in Chapter 7.

1.1 Description of the insurance industry and key players

The following section is written based on the author's working knowledge of the Insurance Industry over 20 years and gives an introduction of the key stakeholders in the insurance industry that are relevant to this thesis; readers with a background knowledge of insurance may wish to skip this section. The insurance industry is active in all developed world countries [188]; with a nascent 'micro insurance' industry in the developing world [176, 269]. The main sub-sectors can be grouped as: life, health and general insurance. Insurers themselves purchase insurance, the companies offering this cover are called Re-insurers [143]. The Lloyd's insurance market is a unique mutual market place and offers both insurance and reinsurance [133].

New business comes to insurers through a variety of channels including: direct (via the telephone or internet) or through brokers [147]. Insurance operations are regulated in a variety of ways around the world, some regulators are elected whilst others are fixed bodies appointed by governments.

General Insurers As noted earlier general insurers provide insurance against a wide variety of hazards. Business is typically on a one year renewable basis [164] and is provided to commercial businesses ('commerical lines') and the public ('personal lines') (though longer term policies are also common in some classes of business). Property damage policies pay out when property (buildings or contents) are dam-

aged or lost when certain specified insurance events occur. Events can include: wind damage, flood damage, earthquake, subsidence, theft, fire, escape of water (burst pipes etc), vandalism [15]. Insurability is hard to define [22] but usually requires that losses be fortuitous; i.e. there needs to be some variance of outcome or there is no need for the insurance - but the variance should not be too large. The loss should ideally have a maximum size. The International Actuarial Association define [113] 'normal' claims as those resulting from a number of independent events. Catastrophes such as hurricanes, earthquakes and floods do not fit this definition. In some regions some hazards are, therefore, excluded from standard cover, but can either be covered through national pooling arrangements [71, 78, 279] or via specialist 'Excess and Surplus' markets [202]. Liability policies (called 'Casualty' in the US) pay out when the insured is deemed to have harmed a third party [143]. Other forms of insurance include: motor, aviation, marine, agriculture and political risks [145].

Reinsurers Reinsurers provide insurance to insurers. The majority of coverage is to direct insurers (those providing insurance to either the public or commercial businesses) - some reinsurers provide reinsurance to reinsurers - this is called 'retrocession' [143]. Two major reinsurance markets are in London [150] and Bermuda [9] with multiple companies in each locations. Two of the largest reinsurers are Swiss Re and Munich Re and these are based in Switzerland and Germany [112]. Reinsurance contracts are written on various bases such as: quota share, excess of loss and stop loss [100, 143]. With quota share contracts the direct insurer retains a fixed proportion of all claims and 'cedes' the remainder [143]. The direct writer is sometimes also called the 'cedant'. The cedant passes on premiums in accordance with the proportion ceded, but also often with an additional amount. There is typically no limit to the claims payable by either party although some quota share contracts do have limits. Quota share reinsurance is often used by insurers with limited capital who wish to use the expertise of the reinsurer in a given class of business before accepting more of the risk in due course [79]. With excess of loss contracts the cedant retains the first portion of losses up to a specified 'retention' level [143]. The reinsurer then pays for claims in excess of this level, up to a specified amount. The difference between the upper threshold and the retention is called the 'limit'. Direct insurers often have a reinsurance 'programme' where they will pur-

chase reinsurance in a number of layers, often with different reinsurers [100]. Stop loss contracts typically apply to the total underwriting profits of a defined 'book' of business [177].

Lloyd's of London The Society of Lloyd's is an insurance market [79, 133]. The Lloyd's building, in London, houses many competing entities offering insurance. The capital to support the business comes from 'Names' who allocate their funds to one or many 'Syndicates', the syndicates are open for a single 'Year of Account' after which they cease to write new business [135]. They remain 'open' typically for a further 2 years and then declare profits and close (often by purchasing reinsurance to cover any remaining liabilities, called 'Reinsurance To Close', this is typically from the next open year of the syndicate). The Syndicates are run by 'Managing Agents' who provide the underwriters and other staff to carry out the insurance operations. New business comes to Lloyd's via many channels including brokers who typically take a risk round several underwriters who agree to take a certain share of the risk (and record this on a paper record called a 'slip') [147]. Hence insurance in the Lloyd's market often follows a 'Co-insurance' model with many separate insurers each taking a share of a risk. The Lloyd's market writes both direct business and reinsurance. Lloyd's maintains a large fund called the 'Central Fund' [134]. This fund is paid into by the Syndicates. If any of the Syndicates becomes insolvent the Central Fund has the discretion to pay the claims. This is why no valid claim has ever gone unpaid in Lloyd's history even though certain Syndicates have failed. In this sense the Society of Lloyd's is a mutual organisation. The Corporation of Lloyd's is a body which represents the market (for example negotiating licences to trade around the world) but also sets and enforces standards, admits new syndicates and calculates capital requirements.

Regulators Insurance is a highly regulated market because insurers provide no tangible product or service at the time of purchase [203]. Instead they provide a 'promise' to pay for damages should an insured event occur [106]. An unscrupulous insurer could take the premium, pass it to shareholders and fail to have sufficient funds to meet claims when they are due. For this reason insurers must hold additional money ('capital') in case the premiums are not sufficient to pay the

claims [203]. There is a variety of insurance regulation around the world [77]. In the EU insurers are required to calculate the capital required so that their balance sheet at the end of the year is solvent with 99.5% probability; also expressed as a 1-in-200 test [83, 252].

1.2 Use of forecasts within Insurance

This thesis seeks to explore whether forecasts can be useful within insurance. This section discusses briefly how forecasts are (or can be) used at present. The various uses are considered under four headings: (1) Immediate use, (2) Short term pricing, (3) Medium term forecasts within Catastrophe models; and (4) Longer term strategic uses.

Immediate use Some insurance events occur without warning (e.g. Marine or Aviation accidents); compared to others, such as hurricanes, for which there is usually several days if not weeks of warning. Once a Hurricane is observed in the Atlantic (or Typhoon in the Pacific and Indian Oceans) insurers will start to consider the implications for their business [223]. From my own experience at Lloyd's, forecasts of storm track [182, 257, 261] and likely damages [257] are useful to enable management to brief their Boards of Directors and to respond to questions from the Media, Regulators and Rating Agencies. Short term planning (for example deciding to mobilise a team of loss adjusters in a region) can also be assisted [223]. Should 'live hurricane trading' [223] ever be available the forecasts would be useful in assessing whether to purchase them.

Short term pricing In 2010 Lloyd's [139] convened a group of underwriters, scientists, exposure managers and catastrophe modellers to consider the use of seasonal forecasts, I facilitated this meeting. Several of the attendees noted they would be nervous to use forecasts if they were the first institution to do so, due to the potential for harm to their company's reputation should the forecast prove incorrect. The attendees felt that forecasts should be of value to the industry but noted that competitive pressure could prevent the benefits being 'fully realised'. The Lloyd's paper suggested that insurers 'should educate themselves on the ways in which the

skill of forecasts can be measured '; chapter 2 lists multiple skill scores, properties of those scores and then seeks to objectively compare them, an insurance perspective is given in sub-section 2.1.4.

A more recent paper by major reinsurer SCOR [223] considers whether seasonal hurricane forecasts can be useful and considers their level of skill. They make the following comments:

- Forecast skill is low until after the 1 July yet this is key reinsurance contract renewal date; so forecasts will have little practical value until they have skill prior to this date;
- 'It is inconceivable that [Regulatory capital] would be reset on the basis of a forecast' (They are referring here to annual forecasts. See the medium-term rate discussion below where, I believe, forecasts are already affecting capital);
- Marginal reinsurance purchase decisions could be assisted by a skillful forecast (e.g. where other metrics do not lead to a clear decision);
- The pricing of third of fourth event covers[2] would be especially sensitive to the forecast.

In light of these (scarce) publications it appears that annual forecasts are not used actively for pricing within the insurance workflow and this would agree with my own experience. However it is possible that other companies are using such forecasts and not stating this publicly.

Crop insurance is one area where there has been considerable academic debate about the role of forecasts and there have clearly been some attempts to use them (or to assess whether they would have been useful after the fact). Osgood [191] notes that seasonal forecasts were not used in the pricing of weather insurance in Malawi due to 'operational constraints and limitations of launching a pilot' and goes on to say that average payouts in practice were very different in El-Nino and La-Nina years (an order of magnitude different according to Osgood [190]). As such, they conclude, a skillful ENSO forecast could undermine traditional insurance due to Inter-temporal

[2]For example a third event cover pays out only after two other major hurricanes have made landfall in the year.

adverse selection [157,214]. Carriquiry [37] suggests that forecasts and insurance can be combined to good effect but note that 'implementation of forecast-contingent insurance policies will require non- trivial innovation'. Skees [232] suggests that the price of insurance should be adjusted to take account of season-specific information or sold before such information becomes available [233]. Daron [57, 58] successfully developed a Bayesian Network approach to crop insurance in Kolhapur (India). Daron built forecasts from the HadRM3[3] into the model but found 'Given the size and nature of the uncertainties associated with climate predictions on the scales relevant for [Weather Index Insurance], it is only conceivably appropriate to use [Bayesian Networks] to promote improved understanding rather than attempting to provide optimal decisions'. It appears that the use of forecasts is not yet widespread within insurance pricing of crop or other parametric products. Chapters 3 and 4 develop an insurance index within the Lorenz 96 systems [155] and show, in section 4.6, that forecasts can improve profitability and reduce the risk of insolvency, so there may be circumstances where such forecasts are useful. Chapter 5 considers hurricane forecasts and whether they can be used to amend pricing or business volume decisions, demonstrating, in section 5.3, that suitably skillful forecasts could improve profitability if carefully used.

Medium term forecasts within Catastrophe models Following the 2005 hurricane season a number of studies were published suggesting that hurricane risk was elevated above long term averages [73, 266]. The insurance industry called for better modelling (for example Benfield [230]). In light of this the major Catastrophe Model providers updated their approach to modelling by providing a 'medium term'[4] view of risk [226], to augment the climatology view already provided. Lloyd's has required its market to consider the capital implications of such raised levels of hurricane risks for some time [137]. According to evidence given to the NAIC in 2011 [212] from 2005 to 2008 RMS carried out an 'Expert Elicitation' process [8, 119, 151] to calculate their medium-term view[5]; after this they switched [264] to a blend of 9 (and

[3]UK Hadley Centre Regional Model

[4]These are also referred to by some in the insurance industry as a 'near term' view.

[5]Under the expert elicitation, respected hurricane scientists were each asked to state how they expected hurricane frequencies to change (relative to climatology) over the next 5 year period - these views were then average with equal weight and the model was adjusted accordingly.

later 13) separate forecast-based models weighted according to their hind-cast skill. AIR (the other major Catastrophe model provider in common use) offers a warm sea surface temperature conditioned hurricane catalogue [4]. The AIR approach does not explicitly use forecasts but allows their clients to decide whether to do so. Catastrophe models are used in practice for Pricing and Capital setting; therefore, since forecasts are embedded within such models, it is clear that Insurers *are* using forecasts in their workflow although, in my experience, many would not say or recognise that they are. To my knowledge, North Atlantic Hurricane risk is the only natural peril for which catastrophe models build forecasts in, in this way.

Longer term strategic uses The insurance industry has been concerned about climate change and the long-term impact on its operations for a number of years [67, 136, 173]. For example the impact on coastal communities at four locations around the world of potential extreme sea level rise by 2030 has been considered by Lloyd's [149]. A study [2] commissioned by the Association of British Insurers in collaboration with AIR and the UK Met Office used climate forecasts to adjust a catastrophe model to examine the impact on Hurricane, EU Windstorm and Typhoon risk. The impact on UK flood risk was explored in 2009 by the ABI [54] again by taking climate model output and adjusting a catastrophe model. UK Flood risk was also explored by Lloyd's in 2015 [142] working with flood modeller JBA to assess the changing nature of Thames food risk by 2080. As a final example the UK Insurance Regulator, the Prudential Regulation Authority published their assessment of the impact of climate change on the UK Insurance industry in 2015 [204] and this was grounded in long-term climate forecasts. Therefore, it is clear that long-term forecasts do have an influence on strategic thinking within insurers.

1.3 History of computer modelling in the insurance sector

Over the past 35 years the complexity of computer modelling within the insurance industry, driven by improvements to computer power and speed, has increased enormously. The insurance industry was slow to adopt the methodology in mainstream

business practice[6] however: ultimately a major encouragement to use arose in 2004 when the Financial Services Authority in the UK moved to a probabilistic formation of regulatory capital setting.

Standard insurance company internal models generally assess risk over a one year time horizon. Some assume modelled insurance claims reach their 'Ultimate' levels (i.e. allow for the final settled values after many years of legal debate); but they still typically just take one year of business. Some companies (typically Life Companies) do test the run off[7] of their longer term policies but these still often exclude new business or assume it arises in a largely deterministic way. This time horizon and model definition limits the questions that can be asked of such models. Indeed the models are generally designed to answer the single question 'How much capital should I hold to survive a 1 in 200 level of risk over the next year?'.

The insurance industry is a collection of a large number of competing companies. Competition is a major force and (particularly for larger companies) the actions of one company can affect the whole industry. Yet, typical insurance company models do not include competition effects (as shown by the following literature review). These omissions from models mean that (for example) questions about the long term impact of regulatory decisions or pricing rules are not assessed. Given the vast uncertainty in such models they cannot be considered predictive; but the insights they offer can still be valuable. They can even challenge long held views of appropriate management actions in given situations by demonstrating that they are unlikely to work even in a simplified model world. In some cases data is plentiful and past claims from prior years can be projected forward into the future allowing for inflation to assess the risk and calculate a price. This 'burning cost' method is used frequently. For some classes of business (e.g. motor) the methodology is sophisticated and uses Generalised Linear Models (GLMs) [31] to derive rating factors so that individuals can be differentiated.

Insurance pricing has to combine past claims information of a particular policyholder with average behaviour of representative cohorts [170]. There are many methods to do this, for example credibility weighting [16]. A similar concept has

[6]Based on personal experience of over 20 years in the financial services industry.

[7]'Run off' occurs when all policies have expired. This can take decades depending on the type of insurance written.

been introduced for forecasting termed Climatology Blending [30]. The recent past is a very short sample of the full distribution of potential outcomes and insurers may be misled by a series of below average claims and this can lead to losses or even insolvency once claims return to more typical values (see Chapter 6). This is particularly the case for companies that are exposed to natural catastrophes where the law of large numbers may take many years to even out (if, indeed, it applies at all). This leads to difficulties in estimating underlying distribution families (model structure error) and also the appropriate parameters of those families. Parameter and model structure errors become even more significant when estimating extreme quantiles of claims distributions. In some areas of insurance (e.g. life assurance or catastrophe insurance) the industry has addressed this challenge by breaking down the risk in a reductionist way. For example as described by Risk Management Solutions [211], hurricane catastrophe models first create synthetic hurricanes based on past hurricane statistics (genesis location, likely track, intensity statistics, radius etc). They then take idealised exposure databases including: location of property, construction characteristics, age, estimated value etc. Using past hurricanes they estimate how different wind speeds have led to different levels of destruction and from this they calculate so called 'vulnerability' functions which relate one to the other. Using vulnerabilities they combine the simulated hurricanes with the property information to create simulated economic damages. They finally take (often idealised) insurance terms and conditions and from this simulate insurance losses. The resulting distributions are a starting point for pricing calculations. Another challenge with the burning cost[8] method is that if risks are non-stationary the past may not be a good guide to the future (a regular disclaimer on policy projections). In fact, risks are changing all the time: climate change is affecting the weather, urbanisation is affecting the built environment, population growth is affecting the density of humanity exposed to risk, inflation affects the costs of replacement, wealth is increasing leading to greater values exposed, technology can reduce risks and costs.

To explore these issues a novel insurance market model is described in Chapter 6 containing two companies which compete on price for a share of the available busi-

[8]The 'Burning cost' is a jargon term used by the insurance industry to mean a moving average of historical claims.

ness based on a simulated (and realistically short) past claims history. In order to motivate the design of the model a literature review of the history of insurance modelling has been undertaken and is presented as background material below; readers familiar with insurance modelling may wish to skip this.

Discussion of the reasons for modelling Daykin et al [61] note, that the General Insurance Study Group of the Institute of Actuaries was founded in 1974 and turned its collective thoughts to the problem of assessing and modelling insurance company solvency. Initial papers listed the reasons for solvency issues but quickly concluded that it is difficult to assess the impact of non-stationary and random risks using deterministic methods which by their nature relate to single scenarios or best estimates. For example, Daykin et al [63] state 'It is important for the management to view the company as a dynamic entity' when they compare this to the traditional static accounting view. Having made the case for dynamic functional analysis it was quickly realised that Montecarlo techniques would be necessary, for example Ryan [220] stresses 'Since even the convolutions of claim frequency and severity in the most simple cases cannot be determined in closed forms, in practice the only approach is to use simulation.' This point is emphasised throughout the literature.

Ryan's model [219] was published in 1980 in a paper for the International Congress of Actuaries and contains many of the key features that later models adopted. Regulators in Finland (see Pentekainen et al [194]) appear to be the first to formally consider such models from a regulatory perspective in the insurance industry following initial work in Germany in banking in the 1960s according to Hooker et al [106]. The 'Finnish Solvency Study' as it became known in the UK appeared to spur UK actuaries on to develop stochastic models [64]. This was clearly part of an international collaboration, work presented in the US by Coutts and Devitt in 1986 [49] was presented at the first international conference on Insurance Solvency held at the Wharton School an event which many of the authors referred to in this literature review attended. In 1987 a major paper was presented to UK actuaries [62], this model included a complex investment model developed by Wilkie [270] for the maturity guarantees working party. This model was the core asset model of much actuarial work until market consistent models gained precedence in the late 1990s and early 2000s [234], [104]. Coutts and Clarke [51] investigated

asset allocation in 1991 using the 1987 model. Meanwhile Daykin et al [63] continued to develop their 1987 model and published a new model in 1990 which is the first reviewed paper to contain a simple insurance market rather than a single company. Daykin and Hay use the 1990 model in various investigations stressing 'Traditional accounting models are of limited value in assessing the financial strength of a general insurance company, partly because of difficulties of measurement and valuation and partly because they do not help the user to understand the inherent uncertainty. The accounting model is even less promising as a vehicle for providing management with the information necessary to make proper decisions about future business strategy...'.

Despite the fact that shortcomings to deterministic methods were identified in the 1980s, when the first 'Model Offices' [9] were developed, the approach took many years to gain acceptance. Coutts and Thomas [50] were still urging actuaries in 1997 to consider '...the power of Stochastic modelling as a technique for examining the overall risk...' noting the trend in the US and also at Lloyd's to consider 'risk based capital' introduced by the NAIC[10] in 1993 and Lloyd's in 1995, as described in Hooker et al [106]. Lowe and Stannard [156], also in 1997, illustrated the use of Dynamic Functional Analysis (DFA) modelling in one particular company - as an encouragement for others to follow. In 2001 Ryan et al [218] considers how actuaries can play a role in Financial Condition Assessments focussing on general insurers and Spiers et al [244] illustrated in 2004 how stochastic modelling could be used to investigate the implications of different management actions on life assurance companies. In the autumn of 2004 the Financial Services Authority introduced the Individual Capital Adequacy Standards (ICAS) [83] and this effectively required larger companies to use stochastic modelling techniques in the UK. Only after this regulatory intervention did this methodology become fully embedded in the board room of UK insurers - a process that still continues in the EU at the current time with the forthcoming introduction of Solvency II. However, the models used by

[9]The phrase Model Office is used in the insurance industry to refer to a computer model of the underwriting, claims, investment and other relevant processes within a firm.

[10]The National Association of Insurance Commissioners [180] is 'the U.S. standard-setting and regulatory support organisation created and governed by the chief insurance regulators from the 50 states'.

firms for regulatory reporting tend to focus on one insurance company (or group) at a time (their own). Indeed Daykin, Pentikainen and Rantala [198] described the addition of multiple insurers as 'largely intractable ' in their seminal 1994 book. Despite this, recent model developments have attempted to include competition effects by considering multiple insurers with rules of interaction for example the work of Taylor [250] and Wright et al [277].

Approaches taken to modelling insurance companies Insurance is an unusual industry because when policyholders pay their premiums they receive nothing tangible in return. The insurer has taken on some of the risks they face and offers to pay out some or all of the costs should they materialise. As such the policyholder has purchased peace of mind. By pooling, the insurer tames a variable risk in return for a fixed payment (effectively a certain loss for the policyholder). The reason why anyone would wish to make such a trade was largely (but not wholly [65]) explained by Friedman in 1948 [87] through consideration of the policyholder's utility function (assuming, as is typical, that it is monotonic increasing though possibly with decreasing slope as wealth increases). The small premium reduces utility with certainty - but by far less than the large decrease in wealth, and hence utility, that would occur should an insurance risk (e.g car crash or house fire) arise. Depending on the price, the expected utility can be higher after the purchase of insurance than before, despite the fact that the premium is calculated to exceed expected losses. This latter point, and the fact that the pooled risk is less volatile, explains why the expected utility of the insurer is also higher. It is critical, however, that the insurer be solvent to pay the claims should the policyholder make them and for this reason regulators require that insurers hold capital over and above reserves (which are held to meet expected payouts only) to ensure a sufficiently high probability that claims can be paid. This reason for holding capital is emphasised in Hooker et al [106] '...A major reason why this is required of insurance businesses is that insurers are regarded as trustees for what is in effect policyholders' money, whereas in many other businesses the goods or services are delivered either in advance of, or very soon after, the consideration is paid. In other words 'trust me' is a major element of what the insurer is selling.'

There are many stakeholders as listed in Daykin [62] : legislators, regulators,

policyholders, third party claimants, intermediaries (brokers), company creditors, shareholders, management and employees. Hooker et al [106] differentiate between existing and future policyholders: the first wants security only (for now), the second wants a mixture of security and cheap insurance. Hitchcox et al [105] also consider rating agencies as an additional stakeholder as they now exert a major impact on management decisions. Coutts and Devitt [49] include future investors and company analysts in the list. None of the papers reviewed explicitly mentions the Inland Revenue as a stakeholder although some do consider taxation directly. The extent to which these various stakeholders are modelled varies across the literature. Typically, legislators and regulators are modelled exogenously. Managers are modelled explicitly (e.g. through premium rating calculations or dynamic choices for asset allocation). Policyholders actions are modelled explicitly in two of the models (Spiers et al [244] and Wright et al [277]) - but are often modelled implicitly via an impact on modelled volumes of business as a function of premium rate. Third party claimants, intermediaries and company creditors are not modelled in the papers reviewed.

Despite being central to the running and prudential regulation of the industry there was no standard definition of solvency in the literature. A practical definition is given in Daykin [62] 'When it comes to definition most people would readily agree that being solvent implies having assets sufficient to meet the liabilities...in practice, therefore, it may be true to say that a company is solvent if the supervisor says that it satisfies his requirements' although Hitchox et al [105] extend this by noting that unless rating agencies are happy with the level of capitalisation the company will effectively cease to be able to write business.

As noted above, insurers hold capital in addition to reserves in order to maintain solvency. Hooker et al [106] list the following additional reasons: claims paying ability, desire to maintain dividend payments in times of unprofitability, the desire to be able to invest in other projects and to support other risks[11] the business runs. Hitchox et al [105] adds that a higher financial strength rating may be achieved with a larger capital buffer[12] and also suggests that an additional buffer above regulatory

[11]'Other risks' might include those associated with investing in more volatile asset classes like equities or lower investment grade corporate bonds.

[12]A higher financial strength rating may be desirable if it enables the (re)insurer to appear

minimum could be held to reduce the probability of costly, time consuming and intrusive regulatory intervention. It is clear from the reviewed literature, and my own experience, that insurers do hold such a buffer in practice.

Daykin et al [61] note the following factors which affect claims paying ability: adequacy of premium rates, claims frequency, weather and natural disasters, court case judgements, inflation, reinsurance recoveries and negotiations on individual claims. All models reviewed have included premium adequacy[13] in some form. Ryan [220], Pentekainen and Rantala [195], Coutts and Thomas [50], Taylor [250] and Wright et al [277] incorporate claim frequency specifically (separating it from severity). Few models explicitly allow for catastrophes (these are modelled explicitly in [50] and [250]). Inflation (both retail price and additional claims related) is modelled explicitly in several models, particularly those defined in the 1980s a time of quite high inflation [184]. Inflation is of particular relevance for long tailed classes[14] (e.g. liability) when high inflation can be a major cause of insolvency - the models which focus on shorter tailed classes (e.g. Taylor [250] and Wright et al [277]) will not be materially affected by their choice to exclude specific consideration of inflation. Reinsurance is modelled explicitly only in Coutts and Thomas [50] despite being a common feature in the insurance market in practice. Court case judgements and individual claims negotiations are not modelled explicitly in any of the reviewed papers; however the impact of these can be (broadly speaking) thought to have been included in the choice of stochastic claims parameters.

Insurers often split modelled claims into three classes [144] : Attritional, Large and Catastrophe. Catastrophe claims relate to major events, such as hurricanes, which lead to multiple policies claiming simultaneously, often modelled with a Catastrophe model [211]. Large claims are used to describe events that affect policies with

stronger to customers and to, therefore, sell more insurance business.

[13]'Premium adequacy' arises when the premium is greater than the sum of the average claim amount and expenses.

[14]Long tailed classes include contracts whose policy term exceeds one year (such as many Life Assurance contracts) but also includes General Insurance policies whose policy term may be only year but where claims can still be payable if damages can be shown to have arisen during the period of insurance. In the second case it can take many years for an insurer to be sure they will get no more claims from such policies. Policies that are not long tailed are often described as 'short tailed'.

very large sums insured (such as oil refineries or sky scrapers), often modelled with a 'frequency/severity' approach [140] (such as a compound Poisson process). Attritional claims cover all causes and policies not covered by Large or Catastrophe models, they are often modelled by sampling from a fitted claims distribution [140]. In summary there are a variety of methods of modelling insurance claims as stochastic processes and, from the papers reviewed, these can be grouped as follows:

- Model total annual claims ratio [219], [62], [63] and [50] (for attritional claims);
- Model frequency and severity separately, [220], [50] (for large and catastrophe claims), [250] and [277].

The methodology in Chapter 6 is similar to the stochastic claims ratio (or loss ratio) approach but uses Lognormal distributions which several papers state they would have liked to use but for run time issues. For example, Ryan [219] and Daykin [62] both mention they used a Normal distribution due to run times associated with other approaches. This can lead to a reduction in the number of simulations considered, however, Daykin et al [63] argue this need not be a problem and useful comparisons can be drawn provided the random seeds are held fixed between different experiments.

The main cash flows within an insurer are listed in Daykin et al [62] (see also [198]): premium income, interest and dividends on assets, reinsurance and other recoveries, claims settled, reinsurance premiums, expenses, tax and dividends. Daykin et al [63] further divide expenses into fixed and variable components which are not treated stochastically in their model.

Insurers have to retain premiums to pay for future claims; they also hold capital in addition as described above. These are both invested in a variety of assets in order to provide: a return to shareholders, inflation protection and liquidity. Some asset classes are more risky than others of course; if insurers choose to invest in volatile asset classes they risk suffering a fall in asset values at a time when they need the money to pay claims. Within general insurers, premium reserves are typically invested in highly liquid low risk assets whereas a portion of shareholder funds is often invested in higher risk equities or corporate bonds. For with-profits[15] life assurers

[15]'With-profits' business is a form of life assurance/ saving policy where the claim payment is

the asset allocation and benefits are intimately linked and so must be modelled together using an asset model of sufficient sophistication [244]. For general insurers the main linkage between investment returns and claims payments is through the common driver of inflation. This tends to affect longer tailed policies (e.g. liability contracts) [93]. Hence papers with a focus on longer tailed business may see the need to include a more complex asset model (as with Daykin et al [63]) whereas those focussing on shorter tailed business (e.g. property catastrophe covers or motor) do not need a complex model; or indeed (as with Wright et al [277]) arguably no model at all.

Dividends are payments back to shareholders. They are at the discretion of the management of the company and reflect the philosophy of the company [198]. If dividends are overpaid the capital base of the insurer is weakened and this can contribute to insolvency. Benjamin [21] suggests the discounted value of future dividends provides a measure of the value to the shareholders. The models reviewed take a variety of approaches to dividend modelling and these can be split as follows:

- Dividends not modelled at all: [219], [220], [195];
- Reduction in investment return: [62];
- Percentage of profits: [50], [277];
- Set dividends to reduce excess capital to a floor: [250];
- Percentage of premium written: [63].

In practice finance directors take multiple factors into account when setting dividends including stability, inflation, real growth, links with recent profitability, size of capital held [198].

Taxation is clearly an important feature for insurers. Changes to taxation (particularly if retrospective) can put a strain on company finances in any industry. All the models reviewed, however, assume taxation levels are held constant throughout the simulation. The methods taken for modelling taxation in the models reviewed are as follows:

variable and dependent on the investment performance of the assets in which the premiums are invested.

- Not modelled explicitly: [219], [220], [195], [250] and [277];
- Reduction to investment return: [62];
- Explicit model, complex treatment of gains and income: [63].

In the case where taxation is not modelled explicitly this need not be considered a major flaw, if changes to the taxation regime during the simulation are ruled out, since premiums would be calculated to cover the taxation arising. Coutts and Thomas [50] are silent on taxation - this is likely to imply they also do not model it explicitly. Expenses are not modelled explicitly in any of the papers reviewed apart from Daykin et al [63].

As described above the amount of effort different research groups or authors put into modelling different features depends on the focus of their study. For example Coutts and Clarke [51] base their study on asset allocation on the model used by Daykin et al [62], presumably because it contained Wilkie's quite detailed asset model. Coutts and Thomas [50] focus on modelling reinsurance so need to develop a detailed model of gross claims and hence model attritional, large and catastrophe claims explicitly. Pentikainen and Rantala [195] focus on the running off of a long tail book of business so develop a reasonably sophisticated reserving algorithm.

Shareholders own the capital held by an insurer and stand to lose all their investment if claims exceed reserves. To compensate for this risk the premium must provide for expected losses and a return on capital (as described in Kreps in 1990 [128]). Only Ryan's 1984 paper [220] and Daykin et al [63] model this feature explicitly.

Hooker et al [106] introduces the concepts of: process risk, parameter uncertainty and specification error. The definitions they give are:

- process risk: variability due to the random nature of the outcomes;
- parameter uncertainty: even if the 'true model' were known you would have to estimate the parameters of the model from a sample of past data - the extent to which these differ from the 'true parameters' leads to a potential error which they call parameter error;
- specification error: The model itself may not be correct. If you chose a different model you would get different view of risk - the difference between these views

they call specification error.

Hooker et al also note that 'Trends and cycles contribute to the overall risk. The perils which give rise to insurance claims and the forces behind them are not static, but change over time. The causes of change may be legal, technological, social, economic, fiscal, political or environmental. The effect of these changes may be retrospective as well as prospective. Changes can be exhibited as trends or cycles and it is often not easy to distinguish between the two.'. Retrospective changes would include, for example, a new law that allows past settled claims to be re-opened and re-negotiated. Hence past losses become higher than initially estimated. The impact of cycles and trends (except for the, so called, 'underwriting cycle') is not investigated further in any of the papers or in this thesis. A critical step towards making allowance for trends is their detection. Some methods in common use require strong assumptions of Gaussian residuals and fulfilment of the Gauss-Markov assumptions [207] in order to use the t-test [121,131,207]. Appendix C gives some thoughts on trend detection and proposes a permutation method to calculate significance which does not rely on strong assumptions. The appendix also suggests some graphical methods for data exploration using the permutation method.

All insurers will fail eventually (unless additional capital is made available in times of crisis which often happens in practice). All models in the reviewed literature truncate the simulation after a small number of years (the number increases as computing power increases). This will fail to show up any features that emerge after the end of the simulation.

The framework of stochastic modelling developed by the actuarial profession in the context of insurance, described above, is closely linked to methods to determine real option values [168] and optimal control techniques [41]. Such techniques extend traditional Net Present Value methods in corporate finance [110] by recognising the value within projects of: delay, abandonment and expansion [56]. Indeed, real option techniques have been incorporated into Life Assurance modelling since the early 2000s. The use of state price deflators [12] rather than risk neutral pricing methods [19, 110], enabled traditional actuarial models and option prices to be merged [118, 235]. Christophides and Smith [44] use DFA modelling to compare two business mix strategies in the presence of stochastic uncertainty. My own experience, whilst

working in the life assurance industry from 1997 to 2004, included using real option methods to assess the efficacy of different bonus[16] decision rules and the efficacy of different investment strategies. Sinha [231] has used the binomial method and real option pricing to explore the purchase of Ahisa, a Mexican life assurance company, by Met Life; concluding that traditional Net Present Value methods would suggest they overpaid but when embedded options are explicitly allowed for, the value was fair. This thesis does not explicitly consider real option theory but it may be useful to explore this in future, for example to quantify the financial value of forecasts within insurance in comparison to their price.

Modelling insurance markets Insurance companies do not operate on their own. They are embedded within a global marketplace of insurers and reinsurers. As such, as with any industry, they are subject to the forces of competition. Insurers serve a variety of customer groups whose sophistication is wide ranging: corporate clients may have risk functions and understand their individual risks better than their insurer does.

The majority of models used in practice by insurers for regulatory purposes focus on the single company in question and do not consider either competitors or customers. Some will include expected levels of new business but these generally enter as exogenous parameters rather then emergent features. This approach reflects the assessment of regulatory solvency over one year (which is the current test). The interaction of the marketplace is likely to be of increasing importance, however, when assessing the value of longer term strategies, new regulation etc. For example Taylor [250] notes that a model office framework allows for testing of proposed regulation before imposing it on the marketplace, stressing that 'regulatory controls need to be applied with great caution lest they induce perverse effects, possibly even the reverse of those intended.'

None of Ryan [219] [220], Pentekainen and Rantala [195], Daykin [62] and Coutts and Thomas [50] model an insurance market. Daykin [63] considers a single company and its behaviour relative to a market that it is deemed too small to influence (though the market does influence it). Taylor [250] considers 20 competing insurers. Wright

[16]Bonuses are a contractual addition to the Basic Sum Assured in a With Profits Policy which is granted following positive investment, expense or mortality experience.

et al [277] have considered markets with 4 and 10 insurers.

Various methods have been developed to introduce competition into the models. In Daykin [63] a process is defined which reduces the amount of business assumed to be written when the company premium rates are above market rates and vice versa. Each company within Taylor's [250] formulation is given a well defined peer group - its own premium rates are a blended average of the peer group's rates and its own target rate which is an expected loss plus a loading which takes account of the company's solvency position. More business is written when rates are cheap and less when they are expensive relative to competitors. Wright et al [277] have the most complex market dynamics, they model both competing insurers and also customer groups (the only model to consider customers) - the customers choose insurers based on both price and also non-price factors as a proxy for customer preferences (such as loyalty) they combine these into the 'total cost' which customers seek to minimise.

It is common practice to carry out profit sharing between the insurer and the customer. This is often done in an informal (non contractual) way and is described as 'payback' as described in Hitchcox et al [105]. Of the models noted here, only Daykin [63] models this explicitly.

The insurance industry is well known [150, 278] for going through periods where premium rates are high (or 'hard') across the industry and other periods when they are low and less profitable ('soft'). Lowe and Stannard [156] describe the process of pricing within catastrophe markets:

- If results are good, prices will decline from their current level;
- Prices will continue to decline until results are poor, at which point they will rise;
- The rate of decline will be greater in a period of benign claims;
- Rises in prices include nominal increases in premium rates and also implicit increases through higher retentions and other coverage reductions.

A major catastrophe may cause some capital providers to reduce their appetite to invest in insurance thus reducing competition and allowing prices to rise. Higher profits are made, attracting new capital providers, increasing competition, lowering

prices and hence profitability. At this point some companies may lose business and the cycle begins again. It is not clear whether this is a repeating behavioural cycle or a series of unprofitable periods punctuated by catastrophes leading to short periods of high profits. Daykin [63] have hard coded underwriting cycles into their model (these can vary by class of business). In contrast Taylor [250] and Wright et al [277] test to see whether cycles emerge from their model set up - and find that they do so in some circumstances.

Consideration of Bayesian methods This thesis does not explicitly make reference to Bayesian methods, though the Climatology Blending approach in sections 2.7 and Chapters 3 and 4 is similar to 'credibility weighting' methods discussed by Bailey [16] who made an explicit link to Bayesian methods in 1950 [17], which was extended by Mayerson [160] in 1964. In 1996 Scollnik [222] used Monte Carlo Markov Chain methods to give a Bayesian prediction of frequency counts in workers compensation insurance. It appears, however, that Bayesian methods had not been taken much further by practicing actuaries since, in 1999 Pereira [196] sought to convince actuaries that Bayesian statistics 'could be useful for solving practical problems', adding 'Bayes is a powerful branch of statistics not yet fully explored by practitioner actuaries'. Pereira's paper considers several standard actuarial procedures from a Bayesian perspective, specifically: 'graduation' (smoothing empirical probabilities), claims reserving and credibility weighting. In the same year Klugman et al [127] carried out a survey of the use of credibility techniques over 190 US insurers and concluded '...that credibility theory is not widely adopted among surveyed actuaries at United States life and annuity carriers to date in managing mortality, lapse and expense related risks'. In 1999, Reiss et al [209] used Bayesian methods in the context of Excess of Loss reinsurance and Verral [262] reinterpreted the well known Bornhuetter-Ferguson [26] method of reserving within a Bayesian framework in 2001. Such methods do not seem to have found their way into mainstream practice in the UK [36], however. Studies have continued to explore Bayesian methods in an insurance context for example Fellingham et al [80] have applied Bayesian methods to Health Insurance Claims Costs in 2007, Puustelli et al [205] to Financial Guarantee insurance in 2008 and Luoma et al [158] to Equity-linked life insurance contracts in 2011. Daron's work on Bayesian Networks in relation to crop

insurance [57, 58] has already been noted. Lloyd's have recently started to consider Bayesian techniques [92] to assess the credibility of forecasted loss ratios given past experience. Given increasing computing power such methods seem to have growing promise to enable continuous adjustment of insurance risk models; it would therefore be interesting to extend the work in this thesis in future to consider Bayesian methods.

1.4 Major new results in this thesis

Details of new work are given in every chapter; this section briefly describes the major new results in this thesis. The insurance question: 'can scientific models improve insurance pricing?' can be focussed into two sub-questions: **(1) Are forecasts skillful? and (2) Are forecasts useful?**

The concept of forecast skill is shown, in Chapter 2, to be relevant and important to insurers. There are many skill scores with different properties, these properties are listed and discussed in an insurance context. Scores that are 'proper' have an expected score that is minimised when a forecast has the same probability density function as the process that generates the observations [94]. This has previously been shown to be a key property [29] and the results of Chapter 2 are in agreement. A new property called 'feasibility' is introduced, requiring a poor score to be given to forecasts that systematically ascribe non-trivial probability to unlikely events. Two common scores are shown to fail this property, one of these (the Continuous Ranked Probability Score - CRPS) is a proper score. Three new robust conclusions are proposed in an insurance context: (1) Multiple, proper, feasible, local, skill scores should be used to assess forecast skill; (2) the Ignorance skill score should certainly be included in the scores used and (3) The CRPS score, since not feasible, should not be used.

Three experiments are designed to provide a ranking amongst scores. The first takes an arbitrary set of real numbers and defines the underlying distribution as a sum of normalised Gaussian kernels centred on them, to produce a probability density function from which a sample of observations is drawn. Several forecasts are produced using the method of Kernel Dressing [30] one of which has the same kernel

width as used to generate the observations. Different skill scores are used to assess the best scoring forecast. Score A is deemed to have 'beaten' score B if A's best scoring forecast has a kernel width that is closer to the true value than B's. The experiment is repeated multiple times and this allows the number of wins, draws or losses to be counted. Score A is ranked higher than B over all if it wins more often. This method finds that the Ignorance score has the highest rank.

A second experiment is carried out where the number of observations is small. This sparse data setting is typical of insurance where we often have only a few years of relevant data [53]. Observations are drawn from a standard Gaussian distribution and three forecasts are tested, also Gaussian but with varying standard deviation, one of which is correct. Similar to the first experiment the skill score that more often picks the correct distribution based on very little information is ranked highest. Apart from data sets with only one or two observations, the Ignorance score again performs best.

The Skill Gap is defined to be the sum over many observations of the difference between the skill score for a given observation/forecast pair and the expected skill score given the forecast, this is defined for any score and is the same as the Information Deficit [216] if the Ignorance score is used. An experiment is set up to test how quickly the Skill Gap can reveal forecasts that are inconsistent with the process generating the observations (termed 'rejection') with a prescribed confidence and probability of rejection. The Ignorance score is shown to reject certain types of forecast more quickly than others, but not all, leading to the conclusion that it is appropriate to use multiple skill scores when using the Skill Gap in this way.

Atmospheric hazards arise from complex, chaotic dynamical systems rich in non-linearities and thresholds [153, 154]. The Lorenz 96 systems (I and II) [155] are used as a proxy for such processes in Chapters 3 and 4. A key component of System II, the 'Instantaneous Effective Forcing' is explored and, using System I, three models are developed: (1) constant forcing, (2) a functional relationship based on the slow variables and (3) an Auto-Regressive process of order 4 (AR(4)) [32] based on prior forcing values. From these, five parameterisations are explored (three for the constant forcing term, and one each for the functional and AR(4) models). An ensemble of forecasts at each lead time is created and these are Climatology

Blended and then scored using the Ignorance score. Six different parameterisations of Lorenz System II are explored in this manner. Initially four parameterisations were considered in two pairs (1) Lower forcing and (2) Higher forcing. A second parameter, 'coupling' determines the influence of the fast variables on the slow variables and this was varied in each pair between high and low coupling. Models in the low coupling systems were expected to perform better but unexpectedly did worse for reasons explained in Chapter 3. Two further systems (each with five models) were introduced in response. The chapter concludes that the models are skillful at least for part of the observation period.

Armed with skillful forecasts from the Lorenz system, Chapter 4 explores whether they can be useful in an insurance context. To facilitate this a new concept, ϕ-transformation, is introduced. This is the best scoring relationship between system values and those of a given model. It explicitly recognises that the model is not the system and that when models take certain values the system may tend to take different values. An insurance index is developed which has hypothetical claims payments depending on values of the observed Lorenz variables. This is analogous to existing insurance indices such as those used for catastrophe bonds [116], but the use of ϕ-transformation is novel. Prices for the index are calculated first using traditional, climatology, methods and then using two variants which make use of ϕ-transformed, forecasts. The first variant takes the transformed ensemble mean to update expectations and the second uses climatology blending [30]. The first method is shown to be the most successful in this case and, using the most skillful forecasts, outperforms climatology pricing in several ways: (1) the company is 16% more profitable on average, (2) it goes insolvent with one fifth of the frequency and (3) the prices charged are on average, 10% cheaper. This method therefore benefits shareholders, insurance regulators and policyholders respectively. The chapter concludes that skillful forecasts can be useful for insurers.

The efficacy of forecasts for insurers is further explored in Chapter 5 where a simple yet informative model of hurricane damages is presented along with a series of forecasts and pricing methods which make use of them. The model incorporates each of the processes: generation, landfall, city hit and strength of storm. Pseudo-forecasts are produced at several of these stages and attempts are made to

improve on climatology insurance pricing. The results highlight that skillful forecasts, may not be useful to insurers unless very carefully used. Indeed a key new finding demonstrates that using such skillful forecasts can reduce profitability and financial stability rather than increase it, as the impact of lower variance is passed to policyholders through lower premiums which can prove inadequate on multiple occasions. The new findings are discussed in the context of: chance of insolvency, career performance of underwriters; expected profitability and return on capital.

A novel insurance industry agent based model is presented and explored in Chapter 6 that has many of the key structures observed in the insurance industry and discussed in section 1.3. Realistic looking behaviours emerge which are informative and enable various insurance questions to be explored. The model considers two competing companies that suffer the same underlying claims process but have different approaches to pricing. One company (the main company) is the focus and a series of single competitors are included to ensure that realistic checks and balances on profitability are included. Each simulation is run until the main company dies and this is repeated multiple times in a Monte Carlo approach. Average company lifetimes are significantly lower than the regulatory target due to the impact of parameter uncertainty on estimating capital requirements. In the particular experiments explored it is shown that: (1) it is possible to create such a model which displays realistic looking, yet subtle, market behaviours; (2) the insurance pricing method of 'payback' [253] reduces company value and leads to shorter company lifetimes; and (3) the TVaR (see Glossary) regulatory metric, commonly thought to be robust due to its desirable mathematical properties, leads to shorter company lifetimes due to difficulties in its estimation, with the conclusion that it should be used cautiously. This work suggests that further development of agent based models of the insurance industry would lead to considerable insights regarding: resilient forms of regulation and robust pricing methodologies.

Appendix C presents some comments on a stand-alone problem of interest in the insurance sector, 'Trends' which play a key role in insurance. For example, their 'detection' is important to ensure premium rates remain adequate and capital sufficient. Modern computing power can reduce the need for often restrictive statistical assumptions about the process underlying a potential trend. A permutation method

for estimating the slope probability is presented in the appendix along with three novel graphical techniques which are explored for various pseudo and real data sets. The example of tide gauge data is explored in detail and the new methods used to give evidence for accelerating sea level rise at multiple locations.

In summary, based on the work presented here I believe the question posed at the start of this chapter can be answered affirmatively. Provided they are used appropriately, scientific models can improve insurance pricing.

Chapter 2

Measuring forecast skill

'***A reasonable fee to pay to an expert who has estimated a probability as*** p_1 ***is*** $k\log(2p_1)$ ***if the event occurs and*** $k\log(2-2p_1)$ ***if the event does not occur ...When making a probability estimate it may help to imagine that you are to be paid in accordance with the above scheme.***'

Good 1952 [97]

Forecasts are often carried out to aid decision making or test a theory [274]. In this context 'skill scores' have been developed for a variety of tasks including [274] comparing forecasts, training forecasters, encouraging their honesty and finding 'optimal' parameters [69, 94].

According to Murphy [178], J.P.Finley (an American Meteorologist) was the first to attempt to appraise a forecast by defining a score equal to the 'percentage correct' for his own tornado forecasts in 1886. In his case, however, skeptical reviewers noted that he would have achieved a better score by forecasting 'no tornado' every time, illustrating that his proposed skill score was flawed. The situation is even more challenging for continuous probability forecasts. Many probability skill scores have been proposed and discussed over the last 100 years. Since the 1950s a wide variety of properties of skill scores have also been defined. All the properties found in a literature review are listed (see table 2.2) and then interpreted in the context of insurance. A new property 'Feasibility' is proposed which certain common scores do not possess, to their detriment.

There are infinitely many skill scores [224] and their merits and uses have been

discussed in the literature. This chapter adds to that debate in three ways (1) by comparing different scores in various ways to provide a ranking amongst scores, (2) by interpreting scores and their properties in the context of insurance and (3) by illustrating methods where skill scores allow additional insight or robustness compared with traditional methods. The original elements of this chapter are believed to be as follows:

- Section 2.2: Extending the method, first introduced by Benedetti [20], of using the Taylor Expansion to contrast the Ignorance Score and the Brier Scores. Thereby showing that **the Brier score should be passed over in favour of the Ignorance score**;
- Section 2.4: A kernel dressing experiment whereby multiple skill scores are compared objectively, creating a ranking amongst scores, **demonstrating that I J Good's Ignorance score [69, 97, 216] performs best in that setting**;
- Introduction of a new skill score property 'Feasibility' (defined in section 2.1) which highlights scores that penalise impossible or highly improbable events. Later sections (2.3.1 and 2.3.2) show that some skill scores fail this property. This work has been partially published by the author in [240];
- Section 2.5: Illustration of how scores perform under conditions of sparse data and demonstration **the Ignorance score again performs best**;
- Section 2.6 describes an experiment, using a measure known as the 'Skill Gap' (defined later), to calculate how quickly a given score will show that a forecast system is different from the distributions underlying observations, showing that the Ignorance score performs well in some cases but the Proper Linear, Spherical and Naive Linear scores do better in some circumstances, leading to the conclusion that **the use of multiple proper scores to assess forecast skill can be useful**;
- The poor performance of the MSE and Naive scores in some or all of the experiments described suggest that non-proper scores should not be used.
- Section 2.7: Introduces an algorithm to produce forecasts of varying 'quality' and then uses the climatology blending method [30] to improve the skill score for each of these forecasts.

2.1 Definitions and score properties

This section defines key skill scores and some of their properties [23, 34, 81, 84, 90, 94, 165, 224, 274]. There is nothing original in this section except for the presentation and the definition of the 'Feasibility' property in section 2.1.2.

2.1.1 Definition of a skill score

Let Ψ denote the space of probability density functions. If p is a probabilistic forecast (i.e. $p \in \Psi$) and v an observation of the system with sample space Ω (so that $p : \Omega \to [0, 1]$) then a skill score is a functional $S : \Psi \times \Omega \to \mathbb{R}$. All the skill scores below can be thought of as having three components (one or more of which are often zero) (1) A local component which is the contribution of the forecasted probability of the event that actually occurred, only, (2) A forecast only component which is a contribution of the structure of the forecast to the score value - without reference to what occurred and (3) A mixed term which relates the forecast distribution and the event that occurred in a non-separable way. The following equation uses this split where f represents component (1), g component (2) and h component (3).

$$S(p, v) = f(p(v)) + g(p) + h(p, v) \tag{2.1}$$

2.1.2 Properties of scores

The following is a list of all the skill score properties in the reviewed literature, which may not be exhaustive.

Orientation Scores can be either positively or negatively oriented. In the negative case scores operate like cost functions - the lower the better. In this chapter all scores are given a negative orientation. This is always an arbitrary choice and the literature contains examples of each. With negative orientation a poor score is one with a high value and a good score has a low value. Forecast A has a better score than forecast B if it is lower, and a worse score if it is higher. The terms: Good, Poor, Better and Worse will be used consistently in the following chapter. In some cases there is a forecast which gives the lowest possible score and in this case the term Best will be used given the choice of negative orientation.

Propriety A score S is **'proper'** (see for example [274]) if, for any pair of forecast pdfs p and q, the following inequality holds.

$$\int q(z)S(p(z),z)dz \geq \int q(z)S(q(z),z)dz \tag{2.2}$$

A score is **'strictly proper'** if the inequality is strict, this concept is made clear in Bernardo [23] but he uses the term Proper. Brown [34] uses the term 'admissable' instead of proper. The implication of this property is that if forecasters are incentivised using Proper scores they will do best to give a forecast p that is equal to their true beliefs q. McCarthy [165] calls this 'keeping the forecaster honest'. Selten [224] uses the phrase 'Incentive Compatible' in place of strictly proper. By this he notes that Improper scores encourage forecasters to deliberately issue a forecast that does not agree with their true beliefs (q). In recent times many forecasts have been fully automated largely removing such behavioural features. As will be illustrated in this chapter, however, it is also advisable to use Proper scores because, in the limit of many observations, they favour the true distribution above other forecasts. Here scores that do not have the Propriety property will be referred to as either Improper or non-Proper.

Locality A score is **'Local'** if it depends only on the forecast density of the observation[1] ($p(v)$). This concept was mentioned by McCarthy in 1956 [165] when he commented that I J Good's skill score (defined in equation 2.9) depends only 'on the probability assigned'. Winkler and Murphy [274] also distinguish scores which take account of all the forecasted probabilities (i.e. including $p(x)$ for $x \neq v$ and those which are concerned only with the outcome that occurred $p(v)$). The latter are Local, the former are not Local. Winkler and Murphy call these 'partial' and 'total measures' (respectively). The first paper in the reviewed literature that gives a formal definition of Local is that of Bernardo [23] where he defines a Local score as one where,

$$S(p,v) = S(p(v),v) \tag{2.3}$$

Using the characterisation in equation 2.1 above, a Local score only has component 1.

[1] 'Observation' is used here in a broad sense which includes estimates of outcomes, for example during data assimilation.

Equity Introduced in 1992 by Gandin and Murphy [90] the equity property relates to categorical forecasts. In this case there are n complete and mutually exclusive categories, probabilities are assigned to each category and one occurs. Following Gandin and Murphy [90], define a 'performance matrix' P where p_{ij} is the relative frequency of the event in which the ith category is forecast and the jth arises. A scores matrix S has elements s_{ij} which denotes the score given in the situation stated. They note that 'climatology' is defined as $p_j = \sum_i p_{ij}$ and the 'predictive probability vector' is $q_i = \sum_j p_{ij}$. A random forecast is, they define, one where $p_{ij} = p_j q_i$. Then a score is **'equitable'** when the level of skill ascribed to a constant forecast is the same as the average level of skill ascribed to a random forecast. They argue that both these situations represents a zero skill situation and hence should not differ. Specifically [90],

$$\sum_i \sum_j q_i p_j s_{ij} = \sum_j p_j s_{ij} \tag{2.4}$$

The left hand side being the average score of a random forecast and the right hand side the score when the forecast is constantly set at category i. Equation 2.4 imposes some constraints on the values s_{ij}, but still allows for many score types particularly for larger values of n. Jolliffe and Stephenson [122] show that for a large class of Proper scores it is impossible for the score to be both equitable and Proper at the same time.

Regularity Gneiting and Raftery [94] define a **'regular score'** (for a categorical forecast) as one where $S(., i)$ is real valued for $i = 1, ...m$, except possibly that $S(p, i) = -\infty$ if $p_i = 0$. They are using the positive orientation rule; the infinite score would be $+\infty$ in the negatively oriented case. In other words a score can only take an infinite value if the event that occurred was designated as impossible in the forecast.

Insensitivity Selten [224] defines scores in a categorical setting where there are n possible states. He defines a forecast as a vector $p = (p_1, ...p_n)$ where each p_i is the probability ascribed to event i occurring, summing to unity over n since the events are a complete description of all possible events and are mutually exclusive.

He defines Δ_n to be the set of all such probability distributions and then defines the score given to forecast p if event i occurs as $S_i(p)$. The expected score of p under another distribution r is defined as $V(p|r) = \sum_i^n r_i S_i(p)$, the **'expected score loss'** is then defined as $L(p|r) = V(r|r) - V(p|r)$. A score is then strictly Proper if $L(p|r) > 0 \ \forall p, r \in \Delta_n$ such that $p \neq r$.

Given p is such that $p_j = 0$ and q another forecast such that $q_j > 0$ Selten then suggests a skill score is **'insensitive'** if $L(p|(1-\alpha)q + \alpha p) = +\infty$ for any $0 \leq \alpha < 1$; here he uses the positive orientation and defines $L(p|r) = +\infty$ if $V(r|r) > -\infty$ and $V(p|r) = -\infty$.

This is motivated as follows. Let $r = (1-\alpha)q + \alpha p$ for an arbitrary p and q. Then he suggests that as α tends to 1, r is 'closer' to p. So that $L(p|r(\alpha))$ should, he argues, decrease as α increases. A score is insensitive if this does not happen for some p. Selten argues that insensitivity is undesirable.

Hypersensitivity Selten [224] then defines the property **'hypersensitivity'**; using the notation above. If $r, p \in \Delta_n$ are two distributions with $r_j > 0$ and $p_j = 0$ for at least one j then the score is hypersensitive if both the following hold: (a) $V(p|r) = -\infty$ and (b) for every $\epsilon > 0$ and $M > 0$ it is possible to find $r, p \in \Delta_n$ with $r_i > 0$ and $p_i > 0$ such that $|r - p| < \epsilon$ and $L(p|r) > M$.

A score is hypersensitive if forecasts can be 'close' (defined by the Euclidean distance between them as n-vectors), yet their scores arbitrarily far away.

Symmetry Selten [224] defines a symmetric score as one in which , given a permutation of the numbers $1, ..n$ (π say) then: $S_{\pi(i)}(\pi(p)) = S_i(p)$. The term is also used differently by Ferro [81] who states that the score should not depend on the ordering of ensemble members.

Elongation invariance Another property defined by Selten [224] considers extensions to the sample space. Specifically a given score may be considered to operate on both Δ_n and Δ_{n+1}. Then a distribution $p \in \Delta_n$ can be mapped by an 'elongation function' to a distribution $\theta(p) \in \Delta_{n+1}$ by adding zero as the nth component. So $\theta(p) = (p_1, ...p_n, 0)$. Then a score is **'elongation invariant'** if $S_i(\theta(p)) = S_i(p)$.

Neutrality Selten [224] defines a **'neutral'** score to be one in which for any pair of forecasts p, q it is true that $L(p|q) = L(q|p)$. He motivates this by considering two theories p and q with one correct. Whichever is wrong should be considered as far from the truth as if the other were wrong, he argues.

Feasibility Motivated by new work in this chapter the following definition (illustrated in figure 2.1) is introduced for a negatively oriented score.

A score S is **'Feasible'** if it assigns bad scores to forecasts that give material probability to highly improbable events. Specifically, let $\lambda = \inf\{p(z)|z \in \text{supp}(p)\}$, this is the probability density of the least likely outcome, the infimum (where $\text{supp}(p)$ denotes the support of the random variable with pdf p). For any $\epsilon > 0$ define a set $M_\epsilon := \{z|p(z) < \lambda + \epsilon\}$; when ϵ is small these are the set of observations that the forecast ascribes small probability density to. Let $\mu = \inf\{S(z,p)|z \in M_\epsilon\}$, the best score amongst the minimal probability events. Then a score is Feasible if $S(z,p) \leq \mu$ $\forall z \notin M_\epsilon$, that is, for any observation that is not in M_ϵ the skill score ascribes a better or equal score than μ to the forecast.

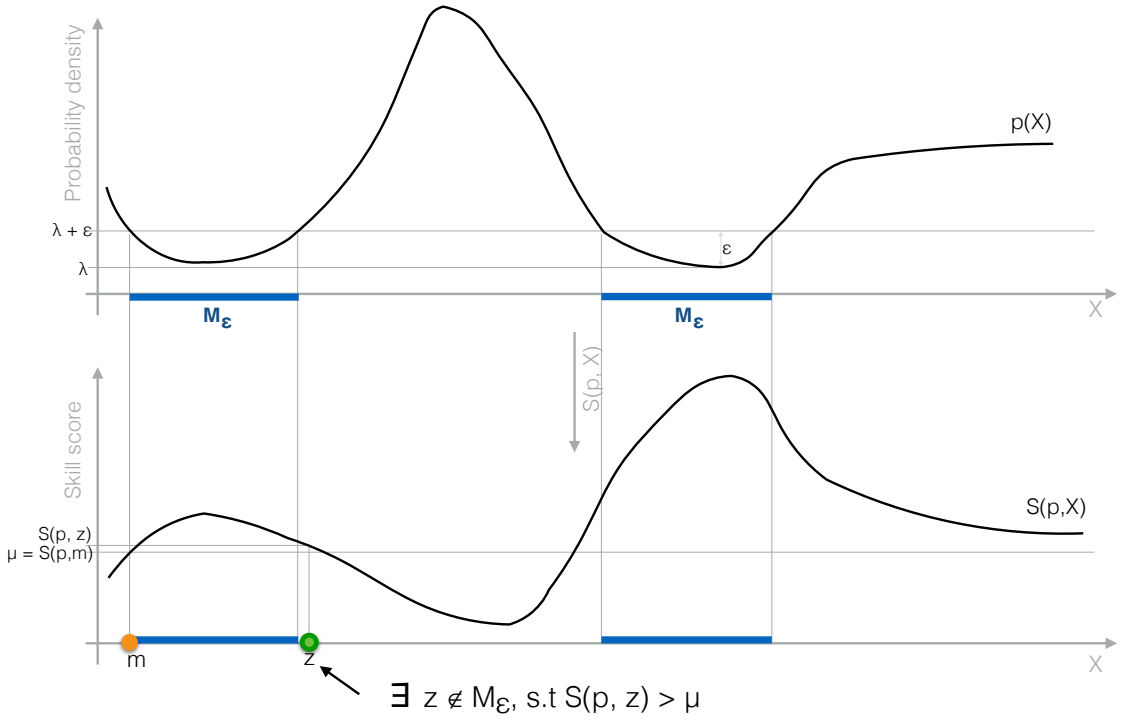


Figure 2.1: Illustration of Feasibility property for a skill score that is **not** Feasible. The top graphic shows the forecast probability density (p) of the observed variable X, where $\lambda = \inf\{p(z)|z \in \text{supp}(p)\}$ the probability density of the least likely observation. ϵ is a given small real number and M_ϵ is the set of values with probability density within ϵ of λ, informally, the set of observed values that are expected to arise with low probability, or 'minimal probability events'. The lower graphic shows the skill score value arising for different observed values X and $\mu = \inf\{S(z,p)|z \in M_\epsilon\}$, is the best score amongst the minimal probability events, the observed value m which corresponds to this best score is illustrated by an orange solid dot. This skill score is not 'Feasible' because the value z (illustrated by a green dot with dark border) is outside of the minimal probability events M_ϵ yet has a worse (i.e higher) score than m, formally $S(p,z) > \mu$.

2.1.3 Ensemble evaluation

The following properties are ensemble evaluation techniques and do not relate to probability forecasts.

Fairness Ficker et al [84] define a property they call Fairness. The property Fairness relates to the evaluation of ensembles. They state there are at least three interpretations of ensembles (1) that they define the only outcomes the forecaster believes are possible and are equally probable, (2) the ensembles represent a collection of functionals of the distribution (such as mean, median or quantiles and (3) that the ensemble outcomes are a random sample from the predictor's belief distribution. Using interpretation 3 they say that a skill score will be Fair if given a forecaster is required to issue a random sample from some distribution then the expected score (using the distribution they truly believe is correct) will be maximised when they choose a distribution equal to their beliefs. They define strictly fair when the expectation is uniquely optimised by the belief distribution.

Accountability Let a **'perfect model'** be defined as one which captures completely the dynamics of a system. **'Model error'** can then be defined as the difference in predictions that arise from models that are not perfect. Smith [237] notes that a prediction from a model that has initial condition uncertainty will not be exact, even if made by a perfect model. In the case of a perfect model this form of error is, however, different to model error since, provided [238] the initial conditions are consistent with observations and chosen to lie on the attractor, the forecasted values will reflect the true Probability Density Function (PDF) of the system. An ensemble that samples the true PDF of a system is defined by Smith [237] as **'Accountable'**, he (and others) note that imperfect models cannot be accountable [123, 237].

Uselessness If an ensemble cannot be distinguished from a sample from the climatology then it is **'Useless'** [237].

Attribute	Definition	Who defined	Also known as
Proper	Expected score (under r) should be as high as possible when the forecast (p) is equal r. L(p|r) < 0	Winkler and Murphy 1968	Desirable property (Good 1952) Honesty (McCarthy 1956), Admissible (Brown 1974)
Strictly proper	Proper and best expected score only achieved when p=q	Bernardo 1979	Proper (Bernardo's definition of proper is strict), Incentive Compatible (Selten 1998)
Local	Score only depends on the probability assigned to the observation that occurs	Bernardo 1979	Payoff depends only on the probability actually assigned (McCarthy 1956) Partial measure of validity (Winkler and Murphy 1968)
Feasible	Score is worst possible for impossible events	Maynard and Smith 2015	
Negative orientation	Lower scores imply better forecasts - like cost functions.		
Equity	Ensures that constant forecasts have the same score as those where the category is chosen at random	Gandin and Murphy 1992	
Regular	Score is real valued except can be infinite (depending on orientation) if the observation arises at a value with forecast probability of zero.	Gneiting and Rafferty 2007	
Insensitive	The score for a convex combination (1-a)p + ar of one forecast (p) with the true forecast (r) should tend to the best score as a tends to 1.	Selten 1998	
Hypersensitive	If a sequence of forecasts can be found where the forecast is close to the true score but such that L(p|r) is as large as required (categorical forecasts)	Selten 1998	
Elongation invariant	The score for n categories should be unchanged by adding a new category and giving it probability zero	Selten 1998	
Symmetric	If a categorical forecast p1,..pn is scored then the score should be the same for any permutation of 1,..n	Selten 1998	
Neutral	L(p|r) = L(r|p) for any pair of forecasts p and q	Selten 1998	
Ensemble Evaluation properties			
Fair	Ensemble evaluation property. Expected score is maximised when the forecaster chooses the distribution underlying the ensemble to be consistent with their beliefs	Ficker et al 2013	
Accountable	An ensemble is accountable if the limiting pdf as the number of ensemble members increases tends to the true PDF	Smith 1995	
Uselessness	A prediction from an ensemble is "useless" if it cannot be distinguished from climatology	Smith 1995	

Figure 2.2: Summary of skill score properties and ensemble evaluation techniques

2.1.4 Score properties from an insurance perspective

Propriety is an important property for insurers. Non-proper scores can reject forecasts that match the underlying distribution of outcomes and, in an insurance context this could lead to mis-pricing of risks, reducing profitability and even potentially leading to insolvency. It is surely nonsensical to use a skill score that is known to favour forecasts that are systematically wrong; insurers should not do so. Also, although, as noted, many forecasts are automated now there is still a strong role for judgement in many fields and any feature that discourages misrepresentation of an expert's beliefs (deliberate or otherwise) is desirable.

Locality is also an important concept for insurers. Profitability over a period will often be assessed when judging whether an underwriter is competent, this is only influenced by what actually happened, not what didn't happen. As noted, non-Local scores take account of aspects of the forecast that did not arise. It should be noted, however, that in one of the uses described in this chapter the, non-Local, Proper Linear score performs better than the, Local, Ignorance score so it may be premature to rule them out completely, see section 2.6.

As noted, some scores cannot be both Equitable and Proper; non-proper equitable scores should therefore be passed over by insurers. Also, it is not intuitively obvious to me that constant scores and random ones (defined earlier) should have the same score. Whilst they both appear to have zero skill this ascribes too much weight on the heuristic interpretation of 'zero'.

Regularity, elongation invariance and symmetry are fairly technical requirements but are sensible. In particular the regularity condition which requires that an infinitely bad score should only arise in cases where an observation forecasted to be impossible actually occurs. In an insurance context such a forecast could be very significant. For example a prediction that there is no possibility of losing a court case would support a zero reserve for that possibility. If the case is eventually lost there will be no earmarked funds to pay the claim which must then be paid out of general cashflow thereby reducing profitability. Effectively this is what happened in the asbestos cases of the 1980s which cost the insurance industry hundreds of billions of dollars.

Insensitivity and Hypersensitivity seem designed to penalise the Ignorance score

which only fails them because it gives an infinitely bad score to events that were stated to be impossible (argued above to be a positive feature of the Ignorance score in an insurance context). These properties are therefore not important in an insurance context.

Selten's Neutrality property is the critical property that leads to recovery of the Proper Linear score as being the unique score which is proper, elongation invariant and symmetric. Later in this chapter it will be shown that the Ignorance score performs better in a variety of settings. Given this, the fact that Ignorance does not have Neutrality suggests this property does not have any features that could be considered important from an insurance perspective, since it rules out arguably the most successful score.

It is not clear why the Fairness property is useful in an insurance setting. Any ensemble can be converted to a probability forecast (for example using kernel dressing or blending described in Brocker and Smith [30]) and then the usual probabilistic scores can be used. In insurance it is necessary to have probability distributions to assess capital requirements.

The Feasibility property, novel to this chapter, is important from an insurance point of view. Rewarding a forecast that consistently ascribes very low probabilities to events that actually arise may lead to that forecast being selected for use in pricing or capital setting. If this is done then the events that obtain will consistently be those that were not expected. In my experience Boards of Directors take a negative view of plans (i.e. forecasts) that are consistently different from the outcome that arises. Scores that are not Feasible might be useful to assess forecasting methods that could be adjusted to be better (i.e. to avoid ruling out forecasting methods that are in development) but such forecasts should not be used in practice unless they had been corrected and then scored well when tested with proper, Local, Feasible scores.

2.1.5 Various skill scores

The following describes various skill scores, these are defined and it is highlighted whether they are proper, Local and/or Feasible. These properties have been highlighted due to the importance for insurance as discussed above. Feasibility or oth-

erwise for each score is demonstrated in section 2.3.

Naive Linear Score [Not proper, Local, Feasible]

$$S(p, v) = -p(v) \tag{2.5}$$

The Naive Linear score has been used for many years, for example it is discussed in Friedman 1983 [86]. This score uses the probability density that the forecast would ascribe to the occurrence of a given observation value. This is intuitive in that: if an observation occurs at a point that the forecast thought highly unlikely then the forecast would score badly - in other words this is a Feasible score. This is a Local score in that it depends only on the value of PDF at the outcome, i.e. only has component 1 of the three described in equation 2.1. This score is not Proper however [29] (i.e. a forecaster can get a better score by giving a forecast that is different to the probability of the event that occurs).

Proper Linear Score [Proper, Not Local, Feasible]

$$S(p, v) = \int_{-\infty}^{\infty} p^2(z)dz - 2p(v) \tag{2.6}$$

The improperness of the Naive Linear score can be 'fixed' by adding the integral term to define the Proper Linear Score in equation 2.6; as described in in Friedman 1983 [86]. By fixing the propriety of the score it is no-longer Local due to the integral term, the integral term is component 2 of equation 2.1 above as it does not refer to the event that occurs at all. For a given p the integral term is constant, hence if the observation, v, occurs at a point of zero probability the score is maximised, hence the score is Feasible. Selten [224] shows that the Proper Linear score is the unique scoring rule up to affine transformation that is symmetric, elongation invariant, Proper and neutral.

Power Rule Scores [Proper, Not Local, Feasible]

$$S(p, v) = (\alpha - 1) \int_{-\infty}^{\infty} p^{\alpha}(z)dz - \alpha p(v)^{\alpha - 1} \tag{2.7}$$

The power rule score family is a generalisation of the Proper Linear score defined for $\alpha > 1$; all are proper. The rule is discussed in Selten [224] although his presentation is for categorical forecasts; the definitions have been converted to a continuous

formulation above. When $\alpha = 2$ the Power Rule score is equal to the Proper Linear score. In this chapter: a power rule score with parameter α is denoted **'powerrule α'**. Since each value of α defines a Proper scoring rule, Selten notes '... the power family shows the there are infinitely many incentive compatible scoring rules'. Again the integral terms implies that the score is not Local and is component 2 of equation 2.1. As with the Proper Linear score the integral term is constant for any particular forecast hence this score is Feasible.

Mean squared error Score [Not Proper, Not Local, Not Feasible]

$$S(p, v) = \int_{-\infty}^{\infty} (v - z)^2 p(z) dz \tag{2.8}$$

The Mean Squared Error (MSE) (see for example Ferro et al [82]) is another example of an Improper score. The further the observation (v) is from the part of the forecast distribution that has the highest density, the greater weight will be given to the squared 'error' term. Hence a high (bad) score is given when high density is ascribed to values far from where the outcome actually occurs. The MSE score can be seen as a generalisation of the Root Mean Squared Error average score (see section 2.4.1). The latter is not a score at all but an average over many forecasts - it suffers from non-properness and as it includes information about the whole forecast distribution is not Local. It is easy to show that $S_{MSE}(p, v) = \sigma^2 + \mu^2 + v(v - 2\mu)$, where σ^2 is the variance of the forecast and μ the mean. Therefore, the Mean Squared Error has both components 2 and 3 as described above. This score is not Feasible as shown in section 2.3.2.

Ignorance Score [Proper, Local, Feasible]

$$S(p, v) = -\log_2(p(v)) \tag{2.9}$$

The Ignorance score, so named in Brocker and Smith [29] is the same as the logarithmic score proposed by I J Good [97] in 1952. This is a Feasible, Proper and Local score. Being Local it only has component 1 of equation 2.1. In his paper Scoring Rules for Forecast Verification, R Benedetti [20] notes that only the Ignorance score is Local and Proper (up to affine transformation).

Continuous Ranked Probability Score [Proper, Not Local, Not Feasible] The Continuous Ranked Probability Score (CRPS) is defined by:

$$S(p, v) = \int_{-\infty}^{\infty} (\int_{-\infty}^{z} p(t)dt - H(z - v))^2 dz \tag{2.10}$$

Where the Heaviside (step) function H is defined as follows:

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \tag{2.11}$$

Hence, the CRPS is defined as the square of the L^2 distance between the distribution function of the forecast and a step distribution function centred on the outcome. See Ferro et al [82] for a similar definition to the one above. The CRPS score, whilst proper, is not Local and as with MSE the squared term in the integral can be expanded to show that this score has components 2 and 3 of equation 2.1. Also, CRPS is not Feasible as shown in section 2.3.1. Ficker et al [84] show that CRPS is not Fair for ensembles, but they also derive an extension (not shown) to CRPS that does have this property.

Spherical Score [Proper, Not Local, Feasible]

$$S(p, v) = \frac{-p(v)}{(\int_{-\infty}^{\infty} p^2(z)dz)^{\frac{1}{2}}} \tag{2.12}$$

As noted in Friedman 1983 [86] the spherical score provides another 'correction' to the naive score to convert it to a Proper score. Note the integral term is the L_2 norm of the forecast PDF. Due to the integral this is not a Local score and can naturally be thought as only having component 3 of equation 2.1. For any forecast the integral term is constant, so as for the Proper Linear score, it is Feasible.

Brier Score [Proper, Feasible] The above skill scores may all be used to score forecasts with continuous PDFs. The following score is used for binary events in a categorical setting. Let p denote the forecast probability of an event occurring. Then there is a forecast probability of $1 - p$ that it does not occur. Following the notation in Benedetti [20] the event occurring is represented by the vector $(1, 0)$ and its non-occurrence by $(0, 1)$. Then given that event

$$\hat{e}_k \in \{(1, 0), (0, 1)\}$$

occurs (one of these two vectors) the Brier score is calculated as:

$$S(p, \hat{e}_k) = |(p, 1-p) - \hat{e}_k|^2 \tag{2.13}$$

Where $|.|$ denotes Euclidean distance between two vectors. Ferro [81] shows that the Brier score is not Fair. The concept of Locality does not apply to Binary scores since once p is specified, the remaining probability is immediately specified.

Table 2.1: Key properties for selected skill scores from insurance perspective

Skill score	**Proper**	**Local**	**Feasible**	**Continuous**
Naive Linear	×	✓	✓	✓
Proper Linear	✓	×	✓	✓
Power rule(s)	✓	×	✓	✓
Mean squared error	×	×	×	✓
Ignorance	✓	✓	✓	✓
CPRS	✓	×	×	✓
Spherical	✓	×	✓	✓
Brier	✓	NA	✓	×

2.2 Ignorance and the Brier score

Benedetti [20] uses Taylor's theorem to expand the mean value of the Ignorance score[2], for a binary event, around a forecast that ascribes equal probability to each of the two possibilities. He shows that the second order Taylor approximation of the average Ignorance score equals that of the expected Brier score indicating that for forecasts with close to equally probable outcomes the expected scores from both scores will behave similarly. He argues this may explain the popularity of the Brier score in circumstances when the two outcomes are close to equally likely. He also notes that the Brier score fails to adequately praise good forecasts of extreme events, however he does not use the Taylor polynomial approach to gain further insights

[2]In the current section an alternative definition of Ignorance is used to ease the notation, i.e. $S(p, v) = \ln(v)$, Since $\log_2(v) = \frac{\ln(v)}{\ln(2)}$ this score is a scalar multiple of the Ignorance score defined above and so the minima etc below will be the same for each.

here. This section builds on his approach to show that the Taylor expansion of the mean Ignorance is very different for an extreme forecast. Later sections in this chapter show that the Ignorance score performs well under a number of conditions which suggests the use of the Brier score should be foregone in place of the Ignorance score. In summary, for forecasts close to equally likely the two scores will give similar results, while for extreme events the Brier score will fail to reward better forecasts and penalise worse ones.

2.2.1 Expected Brier score

Following Benedetti's notation and given a background climatology frequency for the event (denoted f) the expected value of the Brier score in this setting is:

$$E[BS(f,p)] = f|(p,1-p)-(1,0)|^2 + (1-f)*|(p,1-p)-(0,1)|^2 \tag{2.14}$$

$$= f((p-1)^2+(1-p)^2) + (1-f)*(p^2+p^2)$$

$$= 2(f-2fp+p^2)$$

Statement of Taylor's multivariate theorem If $g : D \subseteq \mathbb{R}^m \to \mathbb{R}$ is n times continuously differentiable then.

$$g(x+h) = g(x) + \sum_{r=1}^{n-1} \frac{1}{r!}[(h_1\frac{\partial}{\partial x_1} + ... + h_m\frac{\partial}{\partial x_m})^r g](x) + R_n \tag{2.15}$$

Where R_n is the remainder term.

In two dimensions this reduces to the following. Let $x = (\phi, \pi)$, $h = (f-\phi, p-\pi)$:

$$g(f,p) = g(\phi,\pi) + \frac{\partial g}{\partial f}|_{(\phi,\pi)}(f-\phi) + \frac{\partial g}{\partial p}|_{(\phi,\pi)}(p-\pi) \tag{2.16}$$

$$+ \frac{\partial^2 g}{\partial f^2}|_{(\phi,\pi)}(f-\phi)^2 + 2\frac{\partial^2 g}{\partial f \partial p}|_{(\phi,\pi)}(f-\phi)(p-\pi) + \frac{\partial^2 g}{\partial p^2}|_{(\phi,\pi)}(p-\pi)^2 \tag{2.17}$$

Derivation of Taylor expansion of Ignorance Following Benedetti's notation the expected value of the Ignorance score is: $\Psi(f,p) = f\ln(p) + (1-f)\ln(1-p)$ where f is the frequency the event occurs in the climatology (assumed to be stationary) and p is the probability the forecaster assigns to the event. Let

$$A(f,p) := \Psi(f,p) \tag{2.18}$$

$$\frac{d\Psi}{df} = \ln(p) - \ln(1-p) =: B(p) \tag{2.19}$$

$$\frac{d\Psi}{dp} = \frac{f}{p} - \frac{1-f}{1-p} =: C(f,p) \tag{2.20}$$

$$\frac{d^2\Psi}{dpdf} = \frac{1}{p} - \frac{1}{1-p} =: D(p) \tag{2.21}$$

$$\frac{d^2\Psi}{dp^2} = -\frac{f}{p^2} - \frac{1-f}{(1-p)^2} =: -E(f,p) \tag{2.22}$$

$$\frac{d^2\Psi}{df^2} = 0 \tag{2.23}$$

Suppose the forecaster believes that the correct value for the event probability is $p = \pi$ and the observed value of the climatology for the event is $f = \phi$, then the second order Taylor polynomial for the score value $\Psi(f,p)$ around the point (ϕ,π) is given by:

$$\Psi(f,p) = A(\phi,\pi)+B(\pi)(f-\phi)+C(\phi,\pi)(p-\pi)+D(\pi)(f-\phi)(p-\pi)-\tfrac{1}{2}E(\phi,\pi)(p-\pi)^2 \tag{2.24}$$

Multiplying out and gathering terms this can be re-written as:

$$\Psi(f,p) = (A-B\phi-C\pi+D\phi\pi-\tfrac{1}{2}E\pi^2)+(B-D\pi)f+(C-D\phi+E\pi)p+Dfp-\tfrac{1}{2}Ep^2 \tag{2.25}$$

Comparison of Taylor polynomial to Brier score Following Benedetti if $(\phi,\pi) = (\frac{1}{2},\frac{1}{2})$ then:

$$A = \tfrac{1}{2}\ln(\tfrac{1}{2} + (1-\tfrac{1}{2})\ln(1-\tfrac{1}{2}) = -\ln(2) \tag{2.26}$$

$$B = \ln(\tfrac{1}{2}) - \ln(1-\tfrac{1}{2}) = 0 \tag{2.27}$$

$$C = \frac{\frac{1}{2}}{\frac{1}{2}} - \frac{1-\frac{1}{2}}{1-\frac{1}{2}} = 0 \tag{2.28}$$

$$D = \frac{1}{\frac{1}{2}} + \frac{1}{1-\frac{1}{2}} = 4 \tag{2.29}$$

$$E = \frac{\frac{1}{2}}{\frac{1}{2}^2} + \frac{1 - \frac{1}{2}}{(1 - \frac{1}{2})^2} = 4 \tag{2.30}$$

So around $(\frac{1}{2}, \frac{1}{2})$ as in Benedetti's paper:

$$\Psi(f, p) \simeq -(\ln(2) + \tfrac{1}{2} - 2(f - 2fp + p^2) \tag{2.31}$$

Benedetti then notes that the polynomial in f and p is exactly the same as the formula for the expected Brier score and hence for situations where the event is neither extremely likely or unlikely (i.e. f is close to 50%) then the Ignorance score and Brier scores will give similar behaviour.

2.2.2 Expansion for rare events

Benedetti notes that the Brier score 'becomes inadequate as a skill score' when events are rare and considers a rare event with probability $\phi = \frac{1}{1000}$. He notes that the difference in Brier score between a forecast that considers such an event impossible and one that ascribes the correct probability is just 0.1%, a tiny gain in score. The Ignorance score will severely penalise a forecast that ascribes zero probability density to an event that occurs. The Taylor polynomial of the mean Ignorance score around the point $(\frac{1}{1000}, \frac{1}{1000})$ is derived as follows: $A = -0.007907255$, $B = -6.906755$, $C = 0$, $D = 1001.001$, $E = 1001.001$. Substituting these value into the formula the following expression arises:

$$\Psi(f, p) \simeq 3.143236 - 2(3.953878f - 500.5005 * fp + 250.2502 * p^2) \tag{2.32}$$

In this case the polynomial in equation 2.32 is very different to that of the expected Brier score (equation 2.14) which helps to illustrate why the two scores behave differently in the rare (or near certain) event setting.

2.3 Feasibility of various score types

The following subsections show whether the various scores described above are Feasible. In summary, the following scores are Feasible: Ignorance, Naive and Proper Linear, Power Rule, Spherical and Brier. The CRPS and Mean Squared Error scores are not Feasible.

2.3.1 CRPS - not Feasible

This subsection demonstrates that the CRPS is not Feasible by providing two counterexamples.

Counterexample 1: Gaussian mixture The following sequence of graphics illustrate a major shortcoming of the CRPS score. Consider a bimodal forecast with PDF ($f(v)$) and Cumulative Density Function (CDF) ($F(v)$), for a given observation v. The PDF (figure (a)) and CDF (figure (b)) of one such distribution is shown in figure 2.3. Figure (b) also shows a Heaviside function (H(v) shown blue) centred at an observation value of $v = -1$. The difference, $\delta(z) = F(z) - H(z - v)$, between the CDF of the bimodal forecast and the Heaviside function is shaded green. It is this difference term which is squared and integrated over the whole support of the forecast which defines the CRPS. Note that the CDF has an inflection point at v=0 which corresponds to the trough between the two peaks.

Figure 2.4 illustrates various observations $v_1, ...v_9$ increasing from negative, through zero, to positive, these are shown by a blue vertical line plotted at the observation value. The graphic also shows the forecast f with a dotted red line and the integrand of the CRPS (i.e. $\delta(v)^2$) as a shaded green region. The CRPS score is the area of the green region. This is clearly least when $v = 0$ which is at the at the median of the forecast distribution. The fact that the score is minimised at the median is easily shown. Differentiate S to get $\frac{dS}{dv} = 2.\int_{-\infty}^{v} p(t)dt - 1$ (see appendix A for a proof), this is zero when the integral is equal to $\frac{1}{2}$, i.e. at the median of the forecast. In the above example the median occurs in the middle of the two peaks when the density is close to zero (the pdf of the forecast, in red, has been superimposed on the graphic to illustrate this). Suppose that an outcome of x=0 arose (which would be likely

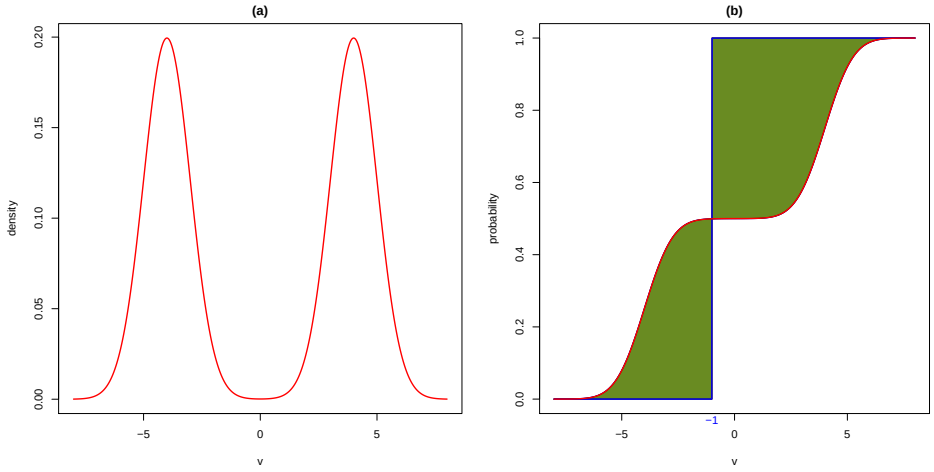


Figure 2.3: Figure (a) bimodal Gaussian distribution. Figure (b), the CDF of the bimodal distribution is shown by the red line, the Heaviside function is drawn at an observation of -1. This illustrates the region that is taken into the CRPS integral when the observation is -1.

if the process generating the observations (the 'truth') was unimodal for example). Then the CRPS score for the forecast, observation pair would be the best possible value. This is despite the fact that the forecast ascribed close to zero probability to the event that occurred. It can therefore be seen that CRPS ascribes a good score (in fact the best) to a highly improbable event, which is opposite of the behaviour required for a skill score to be 'Feasible', clearly CRPS does not have this desirable property.

Counterexample 2: Sawtooth The following extends the discussion presented in Smith et al [240]. Suppose the support of a forecast PDF (f) is comprised of 17 intervals from 0 to 17 each of length 1. Also suppose that the forecaster believes that values are highly likely to arise from only 8 of these intervals ('high density blocks') and that the likelihood is close to uniform within each block. In between these are 'low density blocks' with density approximately $f(x) \approx \frac{1}{1000}$. Each block is not quite uniform with a small positive or negative adjustment, each an order of magnitude lower than the average density within the block. This example is illustrated in figure 2.5(a),(b) and (c). The adjustments are made to make figure (c) clearer. The PDF f is illustrated in figure (a) with a red line - its median is shown with a black vertical line. The score value for each of the potential observations from 0 to 17 is shown in the figure (b): the Ignorance score is shown in green and the CRPS in black. Note that the Ignorance score is inversely related to the density

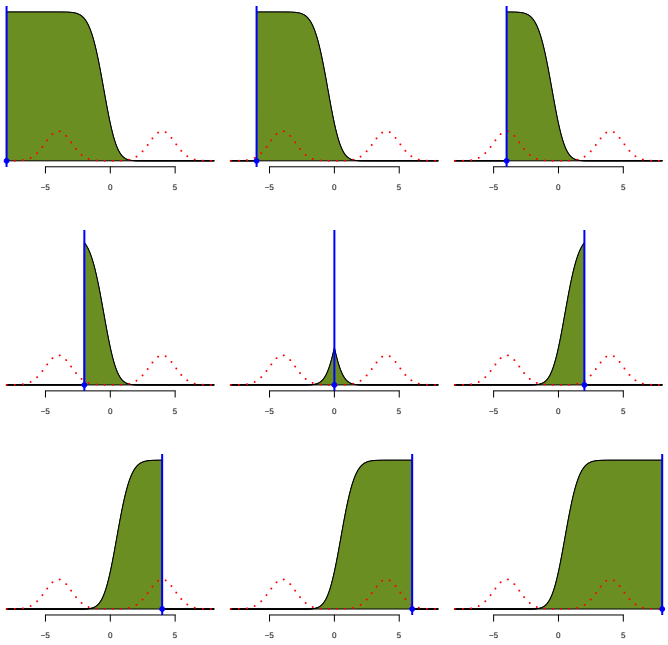


Figure 2.4: Integrand of the CRPS integral (shown green) for various observation values with respect to a Bimodal Gaussian forecast (shown red). This highlights visually that the CRPS score is best (i.e. a minimum) at the median of the distribution; in this example this is at a point of very low forecast density.

of the forecast. The CRPS, however, behaves very strangely. First, as shown above, the best score is at the median of the forecast. This is in a region the forecaster believes is highly unlikely to occur. Also the score smoothly decreases from values to the left of the median and then smoothly increases after this. It passes through several other intervals of near-zero forecast probability, but there is no indication from CRPS the graph that this has occurred. CRPS is indifferent between intervals that are forecast to be highly probable and near impossible.

Assume the the true distribution is shifted one unit to right of the forecast. So that for every outcome that can occur the forecast assigned a near-zero probability. The *average* CRPS score (i.e. integrated over all possible outcomes) will be close to that of a perfect forecast; the average Ignorance score will be close to the worst possible score recognising that the forecast was completely wrong. Figure (c) illustrates this point a different way; the CRPS and Ignorance score values are plotted as $x(v)$ and $y(v)$ coordinates as functions of increasing observation values v. As the observation increases from 0 to 17 the Ignorance value fluctuates whilst the CRPS gradually descends to its minimum at the median, then the process reverses and the

path is approximately traversed in the opposite direction. The small positive and negative adjustments from the multi-modal uniform distribution described above were made so that these two paths are discernible.

The sawtooth example can be used to prove that the CRPS is not Feasible as follows. Using the notation in the definition of Feasibility note that $\lambda \approx \frac{1}{1000}$ and let $\epsilon = \frac{2}{1000}$. In this case $M_\epsilon = \bigcup_{i=0}^{8}(2i, 2i+1)$ and from Figure 2.5(b) it is clear that $\mu < 2$, since all values on the black line, which represents the CRPS score for the low probability region $(8, 9) \subset M_\epsilon$, are less than 2. Choose $t = 1.5$ then $t \notin M_\epsilon$, then $S(p, t) > 4$. In conclusion, there exists a probability density p such that

$$\exists t \notin M_\epsilon \text{ s.t. } S(p, t) > 2 > \mu$$

and so CRPS is not Feasible.

Gneiting and Raftery [94] see this behaviour of CRPS as desirable. They note that with Local scores 'no credit is given for assigning high probabilities to values near but not identical to the one materialising'. In the above example the sawtooth forecast looks very similar to the truth, it is just located in the wrong place. So in a sense the forecast *is* close, and this is what CRPS rewards. In the bimodal case, above, (with a unimodal truth) CRPS is not close, however. In decisions that require accuracy ('is outcome X likely at location Y or not?') this definition of 'close' is not helpful. It may therefore be arguable that the CRPS score could be useful in the context of model development where a forecast that resembles reality should be highlighted for further improvement. Using CRPS in a production setting (i.e. where real decisions are to be made) is inadvisable due to its lack of Feasibility.

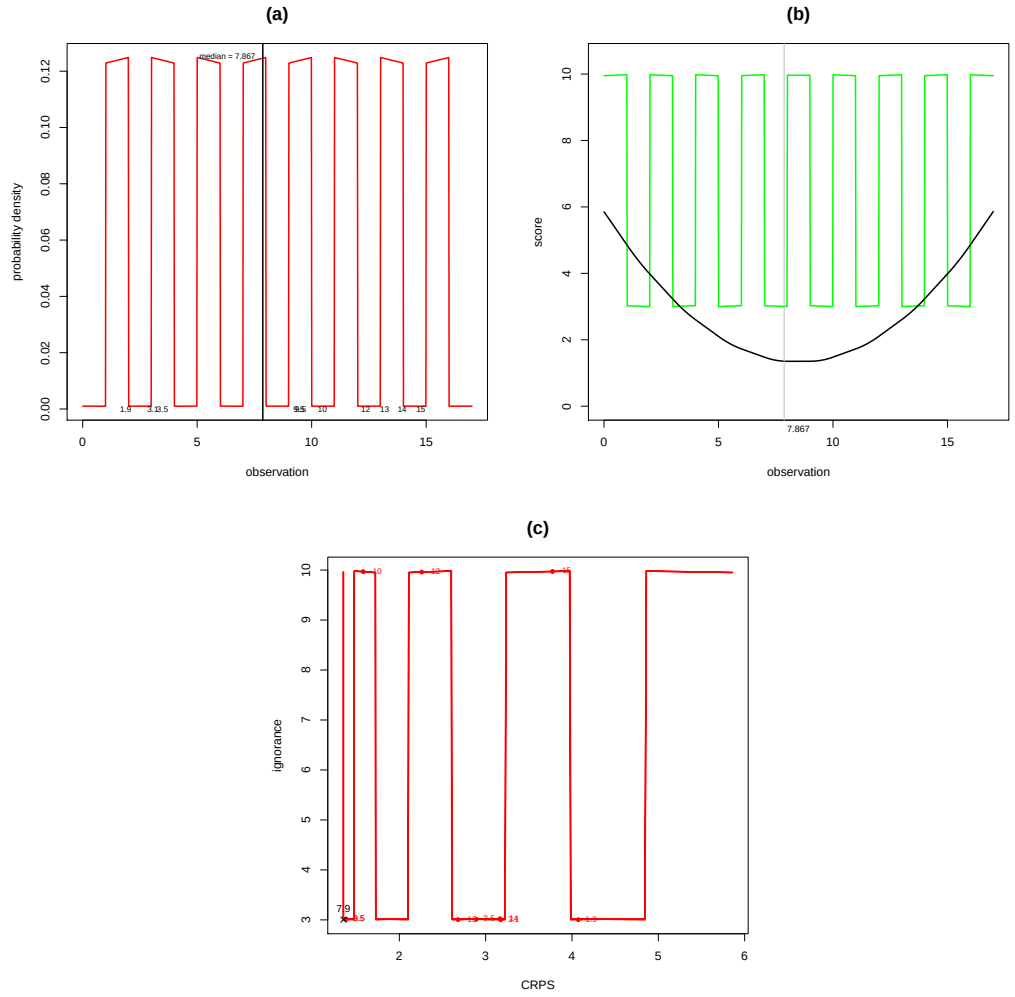


Figure 2.5: Implications of lack of 'Feasibility' - CRPS (not Feasible) vs Ignorance (Feasible). Top left: Probability density function of 8-uniform forecast - with median shown as vertical black line. Top right: Score value for various observed values. Ignorance shown in green and CRPS shown in black. Note that the Ignorance score reacts sensitively to the probability density of observations whereas the CRPS gives similar scores to observations that vary from highly likely (the peaks in probability density) and highly unlikely (the troughs) Bottom: CRPS score vs Ignorance score as the observation moves from lowest to highest value.

2.3.2 Mean Squared Error - not Feasible

The Mean Squared Error skill score S_{MSE} is not Feasible and this is shown by way of the following counterexample. This section uses the notation used in the definition of Feasibility. Recall, to show that a skill score is not Feasible we have to find a value outside of the minimal probability events (i.e. $z \notin M_\epsilon$) for which the score is worse than the best score μ for all points within M_ϵ. In short we must find a probable event that scores worse than an improbable one. Where 'Improbable events' are defined for a given value of ϵ.

Let $\epsilon = \frac{1}{4}$ and consider a bimodal uniform distribution p defined for $\delta < \frac{\epsilon}{8}$, as:

$$p(z) = \begin{cases} \frac{1}{2} - \frac{\delta}{2} & z \in [-2, -1] \cup [1, 2] \\ \frac{\delta}{2} & z \in (-1, 1) \end{cases} \tag{2.33}$$

Consider two possible outcomes 0 and 2. Note that, the probability density of the least likely outcomes $\lambda = \frac{\delta}{2}$ which is the density for all outcomes in the open interval $(-1, 1)$ so that $M_\epsilon = (-1, 1)$. In particular note that $0 \in M_\epsilon$. Note, however, that $p(2) = \frac{1}{2} - \frac{\delta}{2}$ and so $2 \notin M_\epsilon$. By integration, $S_{MSE}(p, 0) = \frac{7}{3} - 2\delta$, and $S_{MSE}(p, 2) = \frac{19}{3} - 2\delta$. The outcome with the best score in M_ϵ must have score μ that is lower than or equal to the score for the observation 0, by definition of the infemum. Therefore $\mu \leq \frac{7}{3} - 2\delta = S(p, 0) < \frac{19}{3} - 2\delta = S(p, 2)$.

The above has shown that $\exists z \notin M_\epsilon$ (i.e. $z = 2$) such that $S(p, z) > \mu$. Therefore MSE is not Feasible.

2.3.3 Ignorance - Feasible

By defintion, $\forall t \notin M_\epsilon$ we have $p(t) \geq \lambda + \epsilon$, and also $\forall z \in M_\epsilon$ we have $p(z) < \lambda + \epsilon$. Since Ignorance is defined as $S(p, v) = -log_2(p(v))$ and by the continuity and monotonicity of log it is always the case that $S(p, t) \leq -log(\lambda + \epsilon) < S(p, z)$ where t and z are defined as above. Hence $S(p, t) \leq \inf(S(p, z)) = \mu$. Therefore the Ignorance score is Feasible, since the above inequality is true for any $t \notin M_\epsilon$.

2.3.4 Linear scores - Feasible

'Linear' scores are taken to include: Naive Linear, Proper Linear, Spherical and Power Rule scores. In each case the score $S(p,v) = A(p) + \frac{-p(v)^\alpha}{B(p)}$ where α is a positive real number and $A(p)$ and $B(p)$ are non-negative terms that depend only on p (and for a given p are therefore constant terms). In particular, with the spherical score A(p)=0 and for the other scores B(p)=1. Note that $p(v)^\alpha$ is monotonic increasing and continuous. Since both A and B are non-negative and constant for all observations (z or t), dividing by B and adding A does not change the inequality so, with the same definition of t and z in section 2.3.3:

$$S(p,t) \leq A + \frac{-p(\lambda+\epsilon)^\alpha}{B} < S(p,z) \tag{2.34}$$

Hence, $S(p,t) \leq \inf(S(p,z)) = \mu$ and since the above inequality is true for any $t \notin M_\epsilon$ the Linear scores are all Feasible.

2.3.5 Brier - Feasible

In the case of the Brier score there are only two possible observations - the event occurs or it does not. These are denoted $e_1 = (1,0)$ and $e_2 = (0,1)$ respectively. The assigned probability that the event occurs is denoted p. There are then three cases to consider: $p < 0.5$, $p = 0.5$ and $p > 0.5$. **Case 1:** If $p = 0.5$ then $\lambda = 0.5$ and $M_\epsilon = \{e_1, e_2\}$ for all $\epsilon > 0$, there are therefore no points $z \notin M_\epsilon$ and the condition for Feasibility is met trivially. **Case 2:** consider the case when $p < 0.5$, then $\lambda = p$. If $\epsilon > 1 - 2p$ then $M_\epsilon = \{e_1, e_2\}$ and again the condition is met trivially. Consider the other situation when $\epsilon \leq 1 - 2p$ then $M_\epsilon = \{e_1\}$ and $\mu = |(p, 1-p) - (1,0)|^2 = 2(p-1)^2$. In this situation $e_2 \notin M_\epsilon$ and $S(e_2) = 2p^2 < 2(p-1)^2 = \mu$ since $p < 0.5 < 1 - p$. **Case 3:** the case for $p > 0.5$ is similar, here $\lambda = 1 - p$. If $\epsilon > 2p - 1$ then $M_\epsilon = \{e_1, e_2\}$ and, as before, the Feasibility criteria is trivially met. If $\epsilon \leq 2p - 1$ then $M_\epsilon = \{e_2\}$ and, by definition, $\mu = |(p, 1-p) - (0,1)|^2 = 2p^2$. Under these conditions $e_1 \notin M_\epsilon$ and
$S(e_1) = 2(1-p)^2 < 2p^2 = \mu$. Therefore in all possible cases the Feasibility criteria is met and the Briar score is Feasible.

2.4 Comparison via 'optimum score estimation'

Gneiting and Raftery [94] note that one of the uses of skill scores is 'optimum score estimation' where parameters are found by finding those which produce the best average score given observations (defined in full in equation 2.36 below); this is also explored in Du and Smith [69]. This section describes a situation where the true parameters are known by construction and then the optimum score estimation technique is carried out using a variety of scores. The scores which lead to parameters that are closer to the true parameters are deemed to have done better than those which find parameters further away. Such experiments lead to a clear ordering amongst scores (in each example). The Ignorance score is shown to perform well in this setting.

Optimal Score Estimators Given a forecast $f(x, \theta)$ of a variable x with parameters θ and observations $X_1, ...X_n$, and a Strictly Proper scoring rule $S(p, X)$ Gneiting and Raftery [94] define the 'Optimal Score Estimator' ($\hat{\theta}$) as follows. Let

$$\mathcal{S}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} S(f(x, \theta), X_i) \tag{2.35}$$

then define

$$\hat{\theta}_n = \text{argmin}_\theta(\mathcal{S}_n(\theta)) \tag{2.36}$$

argmin($\mathcal{S}(\theta)$) is a function that returns the minimum value of $\mathcal{S}_n(\theta)$ over all possible values of the parameter θ. Let $\tilde{\theta}$ be the true parameter underlying a data generating process, then Gneiting and Raftery [94] note that $\hat{\theta}_n \to \tilde{\theta}$ for Strictly Proper scores. (Note their expression uses argmax because they use positive orientation for their scores.)

The following procedure creates an underlying distribution, $f_u(x)$ for an observed variable x whose functional form is defined by a single parameter σ_u. Multiple probability forecasts are also produced using the same procedure. Observations are sampled from the underlying distribution and Optimal Score Estimation is used to choose the forecast with the best score. For a given set of observations this is repeated for multiple score types to test which score chooses the forecast that is 'closest' (defined on page 96 below) to the underlying distribution.

Defining the underlying distribution - Kernel Dressing The following method is known as **'Kernel Dressing'** [30]. Define a PDF f_u as follows. Given any set of N real numbers $S_u = \{r_1, ..., r_N\}$, let σ_u be a kernel width and let the **'underlying distribution'** be defined as:

$$f_u(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sigma_u} \phi(\frac{x - r_i}{\sigma_u}) \tag{2.37}$$

Where,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{2.38}$$

Create observations Sample N_{obs} values $x_1, ...x_{N_{obs}}$ from the underlying distribution f_u - call these **'observations'**. This process can be repeated N_{seed} times to produce multiple sets of observations to quantify the impact of sampling error.

Create a family of forecasts Let $\sigma_m \in \{\sigma_1, ...\sigma_{M_{fcsts}}\}$ be one of M_{fcsts} positive real numbers and use the same set S_u to define forecast distributions as follows:

$$p_m(\sigma_m, x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sigma_m} \phi(\frac{x - r_i}{\sigma_m}) \tag{2.39}$$

This creates a family of probability forecasts indexed by σ_m and when $\sigma_m = \sigma_u$ the forecast p_m is exactly equal to the underlying distribution f_u.

Definition of 'closeness' Given forecasts p_i ($i \in \{1, 2\}$) (with kernel widths σ_i). p_1 is **'closer'** to the underlying distribution f_u than p_2 if $|\sigma_1 - \sigma_u| < |\sigma_2 - \sigma_u|$.

Create datasets S_u to avoid serendipity Given that the underlying distributions are defined using Gaussian kernels there would be a danger that creating the data sets S_u by sampling (say) from Gaussian distributions will produce results that are not general because of the common distribution family. Even other well known distributions such as Lognormal or Gamma may be too 'well behaved'. To avoid this unwanted serendipity the data sets S_u are generated from a dynamical process (the 'Duffing map') which produces highly non-Gaussian outputs. The Duffing map, a discrete version of the Duffing equation [70], is defined as follows:

$$X_{k+1} = Y_k \tag{2.40}$$

$$Y_{k+1} = -bX_k + aY_k - Y_k^3 \tag{2.41}$$

Where a and b are parameters and X_0 and Y_0 are given initial values from which iterative values are generated. The following algorithm is used to produce the real numbers in the set S_u:

- **Choose initial values** Let x_0 and y_0 be real numbers chosen to be on the Duffing Map attractor [3]
- For $j \in \{1, ...N_{ens}\}$
  - **Create jth perturbed initial condition** Let $x_{0,j} = x_0 + \epsilon_j$, where $\epsilon_j \sim N(0, \sigma_{duf}^2)$ and let $y_{0,j} = y_0 + \nu_j$, where $\nu_j \sim N(0, \sigma_{duf}^2)$
  - **Evolve forward K steps** Define $r_j = X_K$ where X_K is the Kth iterate of the Duffing Map (2.40) with initial conditions $X_0 = x_{0,j}$ and $Y_0 = y_{0,j}$.
- Repeat until $j = N_{ens}$
- Define $S_u = \{r_j\}_{j=1}^{N_{ens}}$.

It is now possible to define the following experiments:

Experiment C2.1 Test convergence speed of optimal score parameter.

Parameters for observations: Let $\sigma_u = 0.1$, $N_{obs} = 2^{10}$ and $N_{seed} = 10$. The set S_u is generated from the Duffing map with parameters below.

Parameters for forecast family: Let $\sigma_m \in \Sigma$ where $\Sigma = \bigcup_{i=1}^{3} \Sigma_i$ and Σ_i is defined for integer t in the table below:

Σ_1	$0.05 + 0.01t$	$0 \leq t \leq 2$
Σ_2	$0.08 + 0.0025t$	$0 \leq t \leq 23$
Σ_3	$0.14 + 0.01t$	$0 \leq t \leq 3$

Parameters of Duffing Map: Let $a = 2.75$ and $b = 0.2$, $\sigma_{duf} = 0.01$, $N_{ens} = 2^{12}$ and $x_0 = 0.283995145703728$, $y_0 = 1.092899393566238$, $K = 32$.

[3] For a definition of 'attractor' see the glossary, or Milnor [174]. In practice, potential values are found by running the Duffing map until the values have visually settled down and then choosing any values after this point.

Experiment C2.2.p Find optimal score parameters using different scores.

These experiments are designed to produce an ordering amongst scores.

Parameters for observations: Let $N_{obs} = 2^7$ and $N_{seed} = 10$. Seven different sets $S_1, ...S_7$ are produced from the Duffing map as defined in the table below. Experiment C2.2.p refers to the data set S_p.

Parameters for forecast family: Let $\sigma_m \in \Sigma$ where Σ is defined as for experiment 2.1

Parameters of Duffing Map: Let $a = 2.75$ and $b = 0.2$, $N_{ens} = 2^{12}$, $\sigma_{duf} = 0.01$, $K = 32$ and let the initial conditions be defined in the table below. These underlying data sets are illustrated in figure 2.6:

Experiment	Underlying dataset S	x_0	y_0
C2.2.1	S_1	-1.409707255606690	-0.952496328839017
C2.2.2	S_2	-1.237472722490239	-1.375416550272213
C2.2.3	S_3	-0.398660021372058	-0.979897892460767
C2.2.4	S_4	0.075153134286194	-0.113837933918633
C2.2.5	S_5	0.135405448765377	0.700349003561764
C2.2.6	S_6	0.283995145703728	1.092899393566238
C2.2.7	S_7	0.374505007140980	0.666289868430975

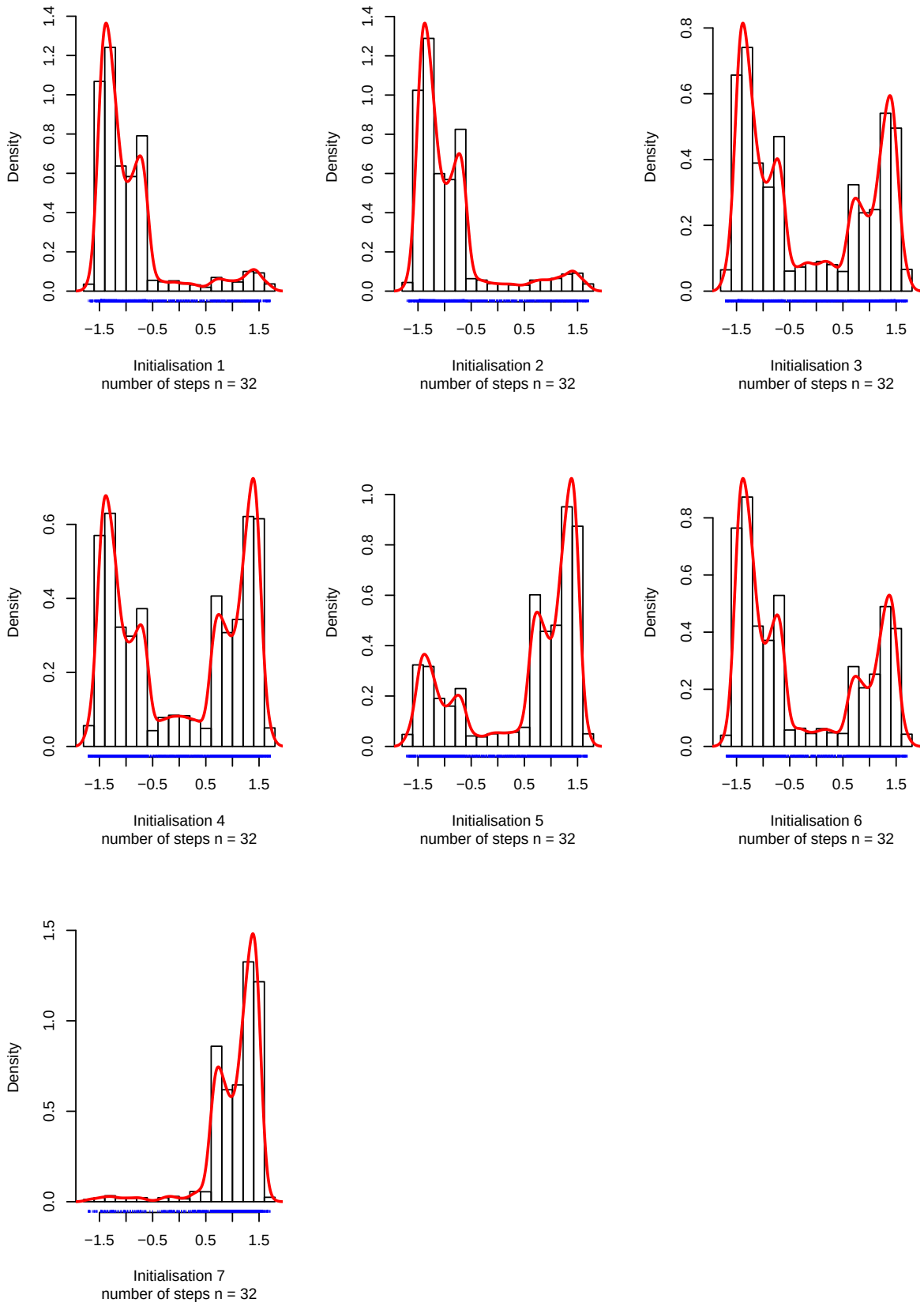


Figure 2.6: Illustration of the 7 underlying sets S_u drawn from a Duffing Map and shown as a blue tick marks and a histogram. The 'underlying distribution' f_u is illustrated by the red line.

Results for experiment C2.1 Figure 2.7 plots the optimal score estimate $\hat{\sigma_u}$ arising from each of the 10 sets of observations, the plot character $0, ...9$ shows result for each observation set. The plot compares $\hat{\sigma_u}$ derived from the Ignorance score (x -axis) and Proper Linear score (y-axis). The cross hairs in the graphic intersect at the true underlying kernel width width (i.e. $\sigma_u = 0.1$). The shaded double-triangular area shows the region where the score type on the x-axis is closer to the true value than the score on the the y-axis. It is clear that even with $N_{obs} = 2^{10}$ observations (far higher than would be available in many practical situations) there is still some scatter around the true value.

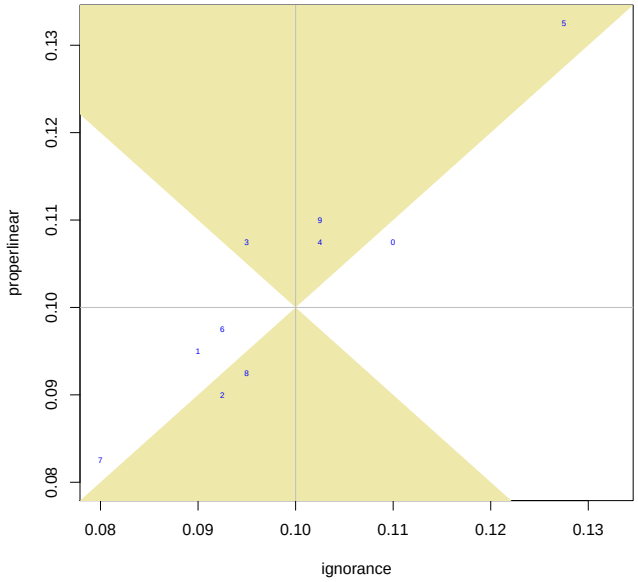


Figure 2.7: Experiment 2.1: Optimum Score Estimates of underlying kernel width - comparison of Ignorance and Proper Linear Scores. Grey cross hairs indicate the true underlying parameter σ_u. The shaded triangular region illustrates the zone where the parameter value derived by the Ignorance score is closer to the true value than the value derived using the Proper Linear score. The Ignorance derives a parameter that is closer to the truth 6 times out of 10.

By aggregating the observations from all 10 seeds, 10,240 equally likely observations are created. Figure 2.8 shows that for both score types (Ignorance and Proper Linear) the average skill score across all outcomes is lowest when σ_m is close to the true underlying parameter of 0.1. Even with this many observations the minimum is not at 0.1 which shows that convergence can be slow.

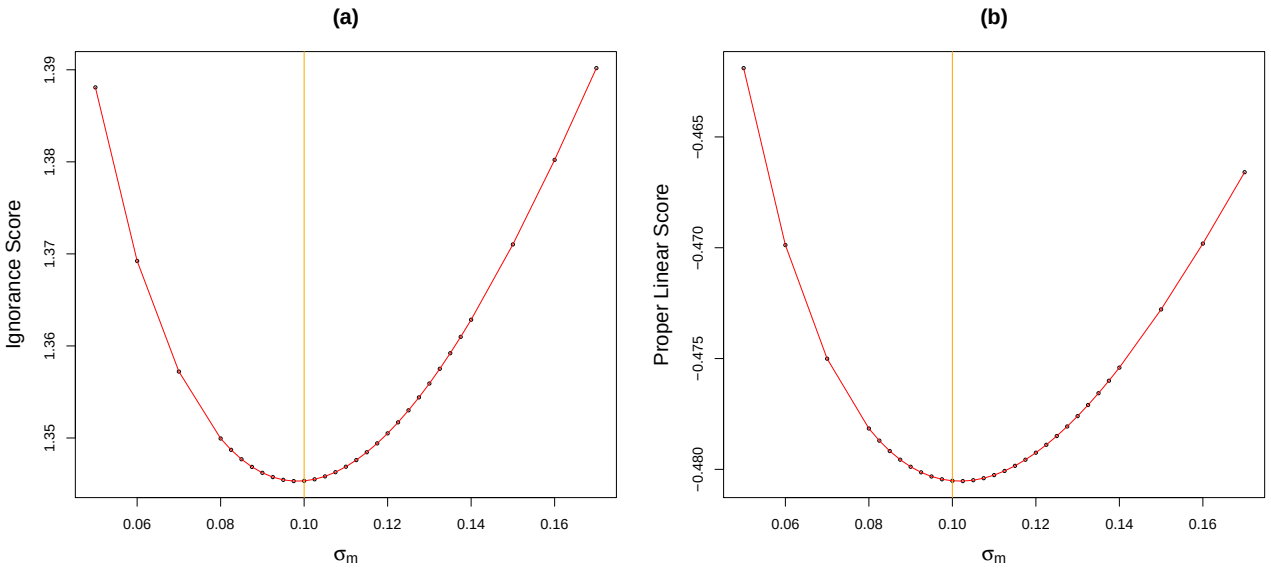


Figure 2.8: Average skill score values for different trial values of the forecast kernel width. Figure (a) shows the Ignorance score and figure (b) the Proper Linear. The best score (minimum) occurs close to the true underlying parameter value ($\sigma_u = 0.1$, shown with vertical orange line) in each case. Average scores for each value of σ_m are calculated over 10,240 simulated outcomes.

Results for experiment C2.2.k These experiments were carried out for all the scores described in this chapter (except the Brier score because the data is not binary). Figure 2.9 shows the results of all the experiments when comparing Ignorance and CRPS. The plot characters are of the form $[k,s]$ where k refers to one of the seven data sets S_k and $s \in 0, ...9$ refers to the N_{seed} observations produced by the seed indexed with s. Various conclusions follow from inspection of the graphic:

- Ignorance beats the CRPS score in this example since more of the plot characters are in the white areas reflecting the fact that Ignorance gets closer to σ_u more often than CRPS; specifically, Ignorance does better 54 times and CRPS 16 times; there are no occasions where they tie. (although one looks close, it is marginally off the diagonal);
- On a number of occasions the optimal score estimate for the CRPS score is at the extremes of the range of tested σ_m values. It is certain, however, that the optimal score estimate on those occasions is equal to or further away from the underlying parameter value σ_u. Truncating the values of $\{\sigma_m\}$ tested is therefore generous to the CRPS score by assuming it picked a parameter closer to the underlying kernel width than it would have with a wider mesh. In these

cases the Ignorance score estimation method finds parameters that are closer to the truth anyway, so there is no miscounting;

- There is a wide degree of scatter arising from the different seeds;
- There is a wide degree of scatter arising from the various underlying data sets.

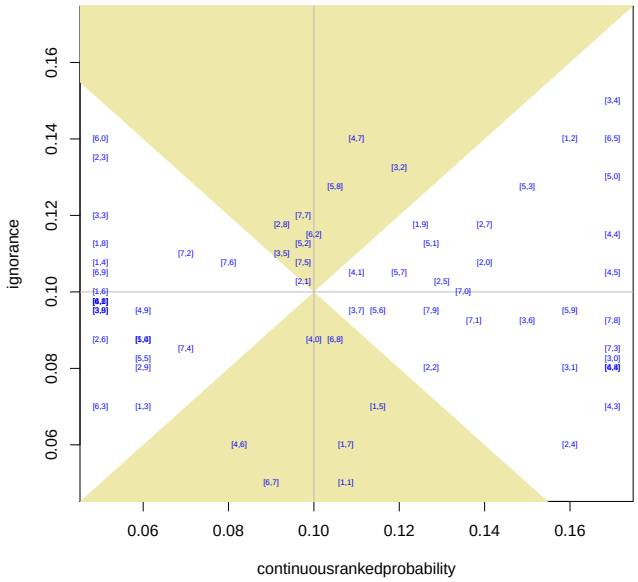


Figure 2.9: Comparison of Optimum Score Estimates $\hat{\sigma}_u$ for CRPS and Ignorance. The plot character is of the form [k,s] where k refers to one of the seven data sets S_k and $s \in 0, ...9$ refers to the N_{seed} observations produced by the seed indexed with s. The shaded area shows the cases where the optimal score estimator for CRPS is closer to the true underlying parameter than the value derived using the Ignorance score. Note that in 54 of 70 cases the result falls in the white area indicating that Ignorance outperforms the CRPS score.

Score comparison metric Figure 2.9 suggests a simple format for comparing the scores. For each pairing of scores, where $score_1$ is on the x-axis (say) and $score_2$ on the y-axis: count the number of times the result falls in the white region, call this N_2 ($score_2$ wins) and the shaded region, call this N_1 ($score_1$ wins). Also count any cases where the coordinate is on the diagonal lines, where the scores draw, call this D. Define the $score_1$ ratio, $R_1 = \frac{N_1 + \frac{D}{2}}{N_1 + D + N_2}$. R_1 is a real number between 0.000 and 1.000. A value of 1.000 denotes a case where $score_1$ produces a better optimal score estimate in every case, a value of 0.000 arises when $score_2$ does best every time. When $R_1 \approx 0.5$ the scores either regularly draw, or they each win a similar

number of times. Using this method, the following table shows the results of many such tests using the same comparison approach. If $R_1 > 0.5$ then score 1 is said to be 'better' than score 2, or 'score 1 wins'.

Based on this method of comparison the following conclusions can be drawn.

- The Ignorance score does best at choosing parameters that are close to the true kernel width;
- Amongst the Proper scores the CRPS does worst;
- The Improper Naive Linear score never outperforms the Ignorance score but does occasionally beat the other Proper scores;
- The Improper Mean Squared Error score does worst (and the reason for this is described in section 2.4.1 below);
- The Proper Linear, Power Rule and Spherical scores all have similar performance in this test;
- As α gets smaller the power rule score performs better on this test.

Table 2.2: Skill score comparison results, Experiment C2.2.k

Scores tested	N_1	**D**	N_2	R_1
CRPS vs Ignorance	16	0	54	0.229
CRPS vs MSE	70	0	0	1.000
CRPS vs naive linear	64	1	5	0.921
CRPS vs powerrule1.5	20	4	46	0.314
CRPS vs powerrule2.0	19	4	47	0.300
CRPS vs powerrule2.5	20	4	46	0.314
CRPS vs properlinear	19	4	47	0.300
CRPS vs spherical	21	6	43	0.343
Ignorance vs MSE	70	0	0	1.000
Ignorance vs naivelinear	69	1	0	0.993
Ignorance vs powerrule1.5	37	5	28	0.564
Ignorance vs powerrule2.0	44	2	24	0.643
Ignorance vs powerrule2.5	44	2	24	0.643
Ignorance vs properlinear	44	2	24	0.643
Ignorance vs spherical	41	5	24	0.621
MSE vs naivelinear	0	4	66	0.029
MSE vs powerrule1.5	0	0	70	0.000
MSE vs powerrule2.0	0	0	70	0.000
MSE vs powerrule2.5	0	0	70	0.000
MSE vs properlinear	0	0	70	0.000
MSE vs spherical	0	0	70	0.000
naivelinear vs powerrule1.5	3	0	67	0.043
naivelinear vs powerrule2.0	3	0	67	0.043
naivelinear vs powerrule2.5	3	0	67	0.043
naivelinear vs properlinear	3	0	67	0.043
naivelinear vs spherical	4	0	66	0.057
powerrule1.5 vs powerrule2.0	28	16	26	0.514
powerrule1.5 vs powerrule2.5	32	14	24	0.557
powerrule1.5 vs properlinear	28	16	26	0.514
powerrule1.5 vs spherical	31	19	20	0.579
powerrule2.0 vs powerrule2.5	27	26	17	0.571
powerrule2.0 vs properlinear	0	70	0	0.500
powerrule2.0 vs spherical	29	20	21	0.557
powerrule2.5 vs properlinear	17	26	27	0.429
powerrule2.5 vs spherical	28	16	26	0.514
properlinear vs spherical	29	20	21	0.557

2.4.1 Behaviour of MSE for a kernel dressed ensemble

This section contains a proof that the MSE always chooses the smallest kernel width available. Equation 2.8 is equivalent to:

$$S_{MSE}(p_m, x_i) = (x_i - \tilde{\mu})^2 + \tilde{\sigma}^2 \tag{2.42}$$

Where $\tilde{\mu}$ is the mean and $\tilde{\sigma}$ is the sd of the forecast p_m (using the same terminology as defined in section 2.39). Now $\tilde{\mu}$ is calculated as:

$$\tilde{\mu} = \int_{-\infty}^{\infty} x p_m(x) dx = \frac{1}{N\sigma_m} \sum_{i=1}^{N} \int_{-\infty}^{\infty} x\phi(\frac{x - r_i}{\sigma_m}) dx \tag{2.43}$$

Change variables in the integral using $s = \dfrac{x - r_i}{\sigma_m}$; then:

$$\tilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} (\sigma_m s + r_i)\phi(s) ds = \frac{1}{N} \sum_{i=1}^{N} r_i =: \bar{r} \tag{2.44}$$

The above derivation makes use of the fact that $\int s\phi(s) = 0$ since ϕ is the PDF of a unit normal distribution and also that $\int \phi(s) = 1$. So $E(p_m) =: \tilde{\mu} = \bar{r}$, i.e. the mean of the forecast is equal to the mean of the ensemble values that gave rise to it. The slight abuse of notation $E(p_m)$ is introduced to help with the next step. To calculate $\tilde{\sigma}$ the formula $\tilde{\sigma}^2 =: VAR(p_m) = E(p_m^2) - (E(p_m))^2$ is used. Now using the same change of variables the following equation arises:

$$E(p_m^2) = \frac{1}{N} \sum_{i=1}^{N} r_i^2 + \sigma_m^2 \tag{2.45}$$

So that,

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} r_i^2 + \sigma_m^2 - \bar{r}^2 = VAR(r) + \sigma_m^2 \tag{2.46}$$

Finally substituting the values of $\tilde{\mu}$ and $\tilde{\sigma}$ into equation 2.47:

$$S_{MSE}(p_m, x_i) = (x_i - \bar{r})^2 + VAR(r) + \sigma_m^2 \tag{2.47}$$

For a given ensemble the terms $(x_i - \bar{r})$ and $VAR(r)$ are constants, hence S_{MSE} can be minimised by letting $\sigma_m \to 0$ which is exactly the behaviour observed.

RMSE The Root Mean Squared Error (RMSE) of the ensemble mean quantifies the distance between the ensemble mean and its corresponding observed value. It is defined as

$$S_{RMSE}(\bar{x}, X) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\bar{x}(i) - X(i))^2}, \tag{2.48}$$

where $\bar{x}(i)$ is the ensemble mean of the i^{th} forecast and $X(i)$ is the observed outcome corresponding to forecast i. While easily interpreted as a distance from the observed value, the ensemble mean is somewhat meaningless in that it does not quantify the distance of any particular forecast trajectory or distribution from what actually happened. The RMSE can be generalised as follows:

$$S^*_{RMSE}((p_1, ...p_m), (X_1, ...X_m)) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\int_{-\infty}^{\infty}(X - z)^2 p(i, z)dz.} \tag{2.49}$$

The original RMSE re-emerges by setting the forecast p as a delta function at the ensemble mean. It should be noted that the RMSE is not a score in the same sense as the others. These are all defined on single forecasts whereas RMSE is defined on multiple forecasts. Note the integral in the summation the Mean Squared Error defined in equation 2.8. The above discussion of MSE showed that the lowest score can always be attained by reducing the standard deviation to zero in the forecast (i.e. reducing the forecast to a delta function as described above)- surely an unfortunate incentive when probabilistic forecasts are intended to illustrate the uncertainty rather than hide it. The RMSE does this for the forecaster automatically which is undesirable.

2.5 Skill score efficacy given sparse data

The previous experiment produced a ranking amongst forecasts. In that example the number of observations (2^7) was relatively large. In many real world problems the number of observations can be much less than this. The following experiment illustrates how well different scores perform when data is sparse, to test whether the skill score rankings are different in this case. The following defines a 'sparse data experiment'.

Create observations Let $x_{k,j}$ be sampled from a unit normal distribution ($N(0,1)$). Define 'Observation Set k' as $O_k = \{x_{k,1}, ...x_{k,2^N}\}$, where N is an integer and $k \in \{1, 2, ...M\}$

Define forecasts Three forecasts distributions $P = \{p_{Narrow}, p_{Perfect}, p_{Wide}\}$ are tested:

1. $p_{Narrow} \sim N(0, \frac{1}{\sqrt{2}})$;
2. $p_{Perfect} \sim N(0, 1)$; and
3. $p_{Wide} \sim N(0, \sqrt{2})$.

Experiment algorithm The following experiment algorithm is used:

- For a given skill score $S(p, x)$
- For $k \in \{1, ...M\}$
  - Calculate the average score $\bar{S}_{p,k}$ for each of the three forecasts over the observations O_k. $\bar{S}_{p,k} = \frac{1}{2^N} \sum_{j=1}^{2^N} S(p, x_{k,j})$, where $p \in P$;
  - For $p \in P$ define $C_{p,k} = \begin{cases} 0 & \bar{S}_{p,k} \leq \bar{S}_{q,k} \text{ for some } q \neq p \\ 1 & \bar{S}_{p,k} > \bar{S}_{q,k} \ \forall \ q \neq p \end{cases}$
- Define $F_{p,S} = \frac{\sum_{k=1}^{M} C_{p,k}}{M}$, where S denotes the skill score.

In words, for each of the three forecasts the average score is calculated for a given observation set; the forecast which has the best average score is deemed to be 'chosen' by the score. This is repeated for M observation sets and the frequency

F_p with which the score chooses each forecast p is calculated. Forecast $p_{Perfect}$ is correct and so scores with a high value of $F_{Perfect,S}$ have done well. F defines a ranking between scores: i.e. if $F_{Perfect,S_1} > F_{Perfect,S_2}$ then we say that score S_1 is better than S_2. The above description now allows experiment 2.3 to be defined:

Experiment C2.3 Sparse data

Parameters for observations: $M = 2^{10}$, $N \in \{0, 1, ..7\}$

Skill scores: The following skill scores are tested: Ignorance, Powerrule ($\alpha \in 1.5, 2, 2.5$), Spherical, CRPS, MSE, Naive Linear

Results for experiment C2.3 Figure 2.10 shows the results for the Ignorance score. When the number of observations is small (up to 2^2) the Narrow forecast is chosen more often. This is not surprising because there is a significant chance with so few observations that all of them are close to the mean, in which case the narrow distribution will give high probability to the observed values. If an event occurs far from the mean the narrow distribution would be penalised more than the wider ones, but such events are rare. Hence the narrow distribution will often avoid being penalised and will therefore be preferred by the Ignorance score. This behaviour is only observed when the number of observations is low. Indeed once the number of observations are equal to or greater than 2^5 the correct distribution is chosen by the Ignorance score over 80% of the time.

Figure 2.11 shows, for multiple skill scores, the proportion ($F_{Perfect}$) that the correct distribution is chosen (for comparison this is the width of the blue segment in figure 2.10). As is typical, the non-Proper scores perform poorly: MSE never picks the correct forecast (in fact it always picks the narrow distribution - not shown) and the Naive Linear score does little better. Apart from situations with very few observations (i.e. two or less) Ignorance performs best out of all the score types. CRPS has similar performance to the power rule score with $\alpha = 1.5$. As the α parameter increases the success rate for the power rule decreases - a similar result to experiment C2.1. Amongst Proper scores the Spherical score performs worse for larger sample sizes. The proportion appears to tend to 1 for all the Proper scores. Selten's criticism [224] of the Ignorance score is, again, unfounded.

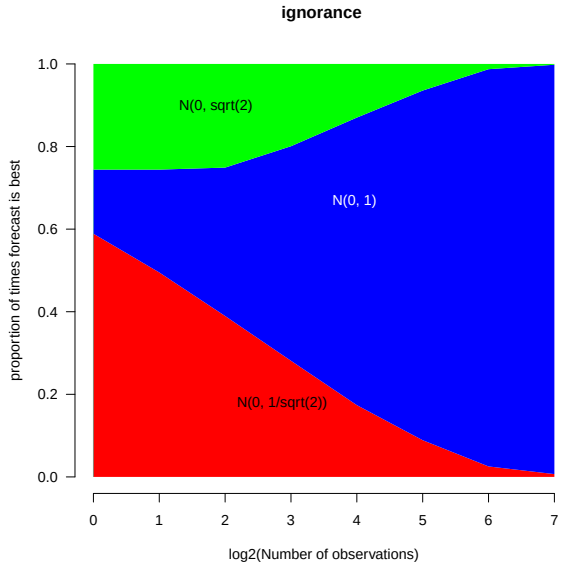


Figure 2.10: Sparse Data Example: Winning proportions for the various forecasts using the Ignorance score. Blue represents $F_{Perfect}$ the relative frequency of choosing $N(0,1)$, Green represents F_{Wide}, $N(0,\sqrt{2})$ and Red represents F_{Narrow}, $N(0,\frac{1}{\sqrt{2}})$. The proportions are shown on the y-axis for a given sample size of observations (N), the x-axis shows $\log_2(N)$. Observations are drawn from $N(0,1)$ distribution.

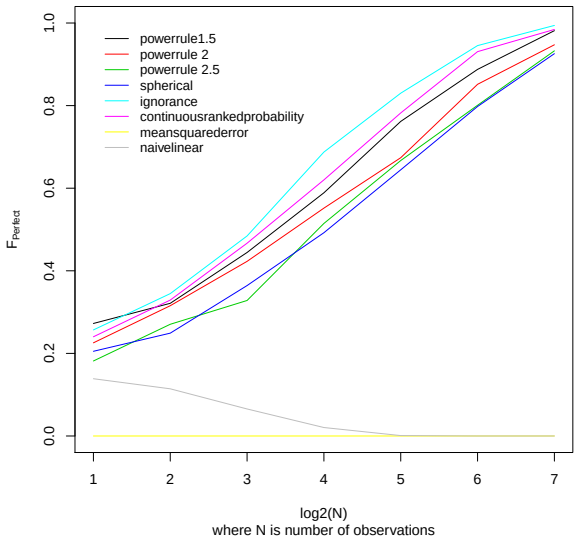


Figure 2.11: Sparse Data Example: Proportion of realisations in which the perfect forecast is correctly identified by different skill scores. Each line corresponds to a different skill score. When more than 2^1 observations are available the Ignorance score has the highest success rate.

2.6 Testing skill scores using the Skill Gap

This section explores the speed with which forecasts from a structurally incorrect model are **'Rejected'**[4] by various scores. A sequence of forecasts from a given model will be referred to as a **'forecast system'** in this chapter. The running average Skill Gap (defined in equation 2.50 below) is calculated. The forecast system can be rejected if the Skill Gap falls outside of a chosen confidence interval. A family of distributions are defined from which the underlying truth and forecasts are chosen. The concept of 'Rejection Time' is introduced as the number of observations after which the forecast system can be rejected with a chosen probability. Initially the Rejection Time is explored using the Ignorance score and then one case is illustrated for the Naive Linear, Proper Linear and Spherical Scores. Other scores are not discussed due to their poor performance in earlier sections of this chapter. The Ignorance score is shown to perform well in some circumstances, but other scores sometimes do better.

Definition of Skill Gap For a forecast system $\underline{p} = \{p_t\}_{t=1}^{\tau}$, skill score S, and observations $\underline{X} = X_1, ...X_\tau$ the Skill Gap(G) is defined as:

$$G_S(\tau, \underline{X}, \underline{p}) = \frac{1}{\tau} \sum_{t=1}^{\tau} \left(S(p_t, X_t) - \int_{-\infty}^{\infty} p_t(x) S(p_t, x) dx \right) \tag{2.50}$$

The integral term is the expected score assuming the observation is drawn from the forecast distribution. Where this converges, define

$$G_\infty(\underline{X}, \underline{p}) = \lim_{t \to \infty} G_S(t, \underline{X}, \underline{p}) \tag{2.51}$$

The score $S(p_t, X_t)$ is a random variable; therefore G_S is also a random variable. If the Ignorance score is used then the Skill Gap is the same as the Information Deficit [69, 216]; the name 'Skill Gap' is used rather than Information Deficit to emphasise that, while the Ignorance score is naturally interpreted as information (in bits) the other skill scores considered below are not.

[4]Here 'Rejected' means that the observed outcomes are inconsistent with the probability distributions from the forecast system. The forecasts may still be useful and they may be informative, but it is not appropriate to use them as probability forecasts.

Let $P(A)$ denote the probability of A and let $\inf(S)$ denote the infimum of the set S. Define the quantile (Q_{G_S}), at time t, of the Skill Gap for a forecast system $\underline{p}$ as:

$$Q_{G_S}(t, \underline{X}, \underline{p}, \lambda) = \inf\{x | P(G_S(t, \underline{X}, \underline{p}) < x) = \lambda\} \tag{2.52}$$

Definition of Rejection If the forecasts are correct in the sense that observations (Y_i say) are drawn from the forecast distributions, so that $Y_i \sim p_i$, then it is possible to calculate $Q_{G_S}(\tau, \underline{Y}, \underline{p}, \lambda)$ (either analytically, or estimated through simulation). Given actual observations $\underline{X}$ let $g = G_S(\tau, \underline{X}, \underline{p})$, $Q_1 = Q_{G_S}(\tau, \underline{Y}, \underline{p}, \lambda)$ and $Q_2 = Q_{G_S}(\tau, \underline{Y}, \underline{p}, 1-\lambda)$. Then the forecast system can be **'Rejected'** if either $g > Q_1$ or $g < Q_2$.

Note that the definitions of the Skill Gap and Rejection make no assumptions about the process generating the observations which may be unknown and even potentially unknowable. The following definition, however, considers a situation where observations are generated from known underlying distributions $\underline{q}$.

Definition of Rejection Time For a forecast system $\underline{p}$, underlying distributions $\underline{q}$ and observations $X_t \sim q_t$, for chosen confidence level λ and probability γ and a given skill score S: the **'Rejection Time'** $RT_S(\underline{q}, \underline{p}, \lambda, \gamma)$ is defined as follows:

$$RT_S(\underline{q}, \underline{p}, \lambda, \gamma) = \begin{cases} \inf\{t | Q_{G_S}(t, \underline{X}, q, 1-\gamma) = Q_{G_S}(t, \underline{X}, p, \lambda)\} & \text{if } G_\infty(q) > 0 \\ \infty & \text{if } G_\infty(q) = 0 \\ \inf\{t | Q_{G_S}(t, \underline{X}, q, \gamma) = Q_{G_S}(t, \underline{X}, p, 1-\lambda)\} & \text{if } G_\infty(q) < 0 \\ \text{undefined} & G_\infty \text{ doesn't converge} \end{cases} \tag{2.53}$$

Note that, for given components $\underline{p}$, $\underline{q}$, λ and γ the Rejection Time is a property of the given skill score S; allowing different scores to be compared. Specifically if $RT_{S_1} < RT_{S_2}$ for scores S_1 and S_2 then score S_1 can be said to be 'better' than S_2 for those particular components. The Rejection Time has been defined in general; in the following examples, however, a forecast PDF (p_0) is chosen and then used for every time t (i.e. $p_t = p_0 \ \forall t$). In this case the integral term is constant for each t so that

$$G_S(\tau, \underline{X}, \underline{p}) = \frac{1}{\tau}\left(\sum_{t=1}^{\tau} S(p_0, X_t)\right) - \int_{-\infty}^{\infty} p_0(x) S(p_0, x) dx \tag{2.54}$$

If observations are drawn from the forecast distribution, for a proper score, the summation term will converge to the integral and hence G_∞ is defined in each case. The following family of distributions will be used to test the Rejection Time concept in a specific controlled environment.

Definition of controlled distribution families Define (f) as the weighted sum of Lognormal, Gamma and Pareto distributions each with the same mean (μ) and variance (σ^2). Specifically:

$$f_{(w_1,w_2,w_3)}(x) := w_1 f_{Lognormal}(x) + w_2 f_{Gamma}(x) + w_3 f_{Pareto}(x) \tag{2.55}$$

Where, $\sum_{i=1}^{3} w_i = 1$ and $w_i > 0 \; \forall i$. Once two weights are chosen the other is determined since they are constrained to sum to unity. The set of weights that meet the criteria above fall in a triangular region of the plane and distributions can therefore be defined uniquely by points (w_1, w_2) in the triangle. Each point represents the distribution $f_{w_1,w_2,1-w_1-w_2}$. The **'distance'** between two distributions f_1 and f_2 is defined to be the Euclidean distance between the points in the triangle defining them.

Figure 2.12 (called a 'Rejection Time diagram') illustrates the calculation of Rejection Time for two points in the triangular distribution space. The series of truth distributions $\underline{q}$ is defined by the red dot in the triangle (note these are all the same) and the forecasts $\underline{p}$ by the blue dot. The quantile lines for various values of t are illustrated for quantiles $\lambda, \gamma \in \{5, 10, 25, 50, 75, 90, 95\}$, these are created by sampling and are therefore not smooth as they would be in theory. $Q_G(t, \underline{q}, \lambda)$ are illustrated by the grey quantile lines and those of $\underline{p}$ by the black lines. The chosen confidence level for rejection is chosen to be 90% (i.e. $\lambda = 90$), the desired probability of rejection is chosen to be 75% (i.e. $\gamma = 75$). The long run Skill Gap is positive in this example because the grey lines converge above the x-axis. Therefore we look for the intersection of the 90th quantile of the black lines with the 25th quantile of the grey lines, illustrated by a red cross - and vertical line in figure. The time at which the lines cross (i.e. the Rejection Time) occurs at the 102nd observation.

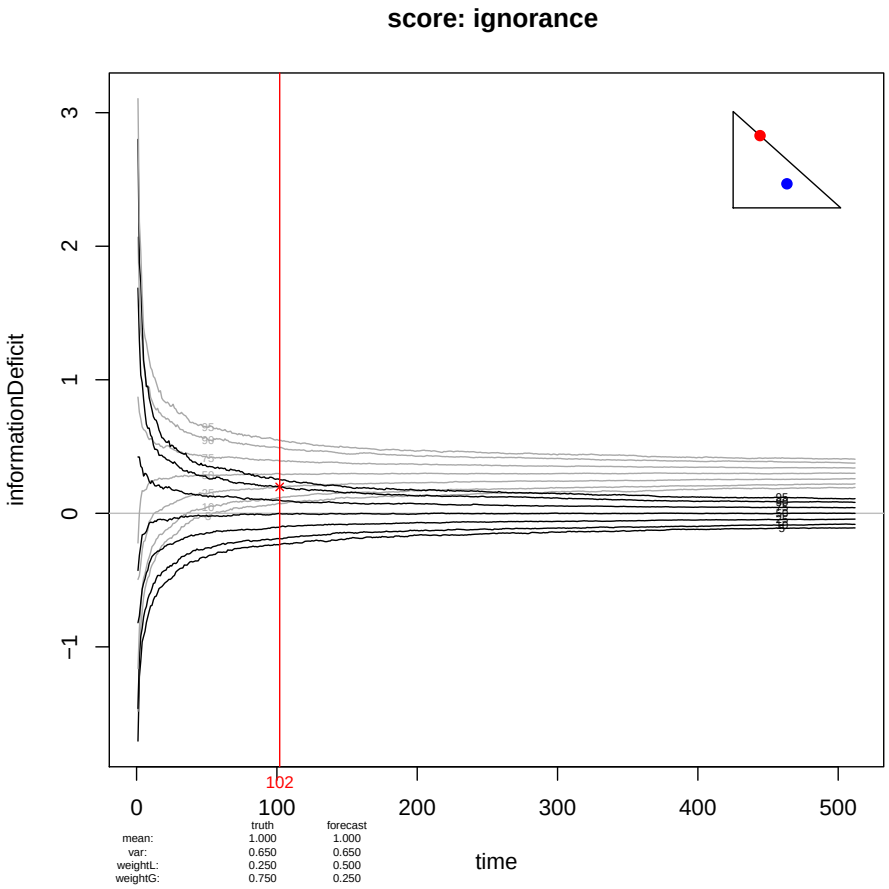


Figure 2.12: Rejection Time diagram: The Rejection Time is illustrated by red vertical line. Grey quantile lines show the observed Skill Gap, black lines show the expected Skill Gap if the forecast is correct. By time 102 we will have rejected the forecast system 75% of the time (at a 90% confidence level). The top right triangle graphically illustrates the chosen truth and forecast distributions.

The following algorithm (see also figure 2.13) uses an empirical approximation of $Q_G(t, q, p, \lambda)$ by sampling M values - sorting them and taking the 100λth largest value.

Experiment 2.4: Algorithm to calculate the Rejection Time

- **Choose confidence level λ and probability γ**
- **Choose forecast and truth** Let p' and q' be points in the triangle of allowable distributions
- For $i \in \{1...M\}$
  - **Sample observations** Let $X_{i,1}, ...X_{i,N_{obs}}$ be sampled from a distribution with PDF q' and let $Y_{i,1}, ...Y_{i,N_{obs}}$ a sample from PDF p'
  - **Calculate Skill Gap sequences** Let $S_X = \{G_i(j, q', p')\}_{j=1}^{N_{obs}}$ and let $S_Y = \{G_i(j, p', p')\}_{j=1}^{N_{obs}}$
  - If $i = M$ stop, otherwise continue for next value
- **Calculate proxy for** $G_{\infty(p',q')}$ Let the median $Q_G(M, p', q', 0.5)$ be a proxy for $G_{\infty(p',q')}$
- **If** $G_{\infty(p',q')} > 0$
  - **Estimate** Q_G For each $j \in \{1, ...N_{obs}\}$ estimate Q_G using the empirical approximation. Let $Q_X(j) = Q_G(j, q', p', 1-\gamma)$ and $Q_Y(j) = Q_G(j, p', p', \lambda)$
- **If** $G_{\infty(p',q')} < 0$
  - **Estimate** Q_G For each $j \in \{1, ...N_{obs}\}$ estimate Q_G using the empirical approximation. Let $Q_X(j) = Q_G(j, q', p', \gamma)$ and $Q_Y(j) = Q_G(j, p', p', 1-\lambda)$
- **Estimate Rejection Time** Let $RT(q', p', \lambda, \gamma)$ be the smallest value j' at which $Q_X(j') = Q_Y(j')$

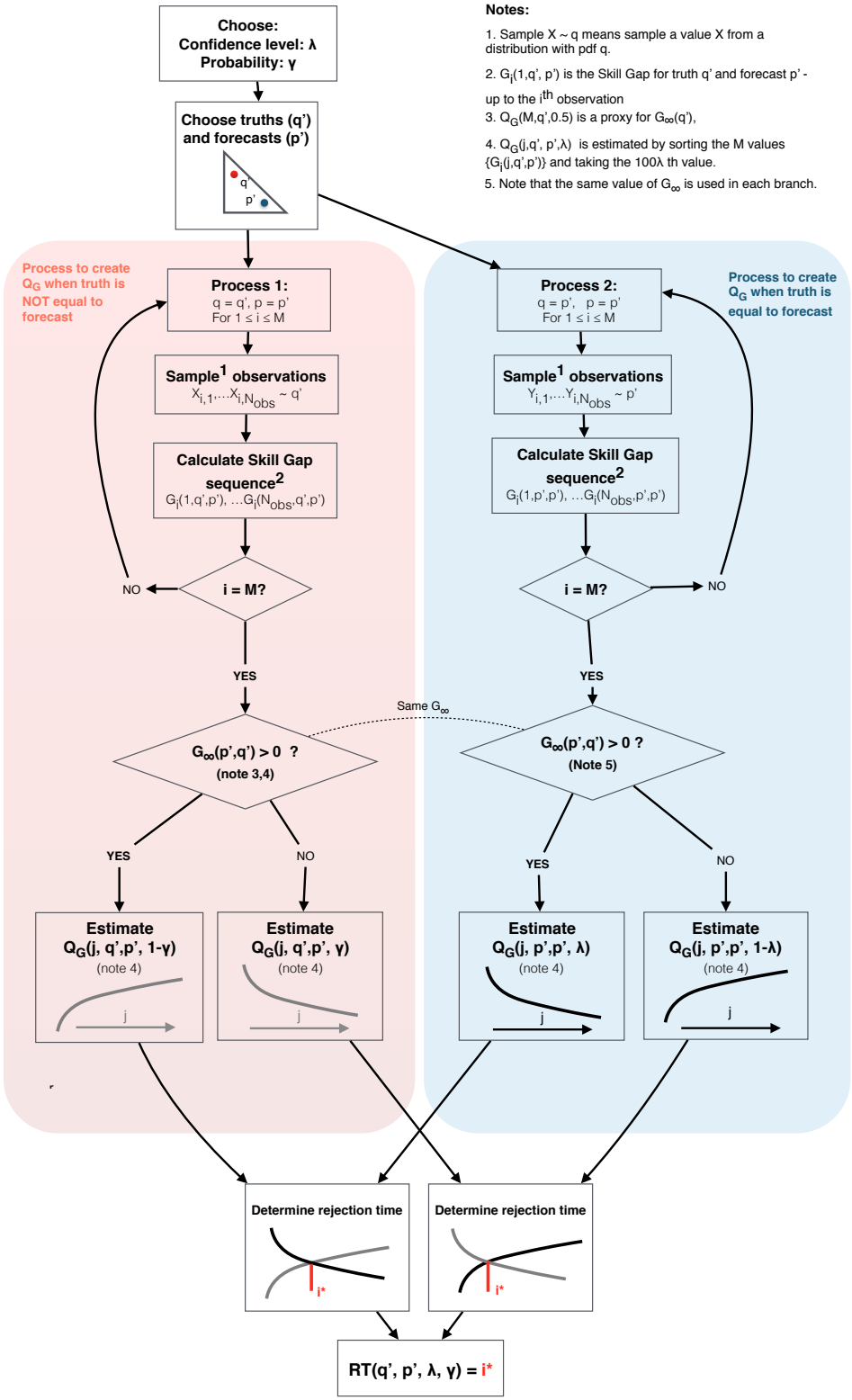


Figure 2.13: Experiment 2.4 Flowchart illustrating the algorithm to estimate the Rejection Time

Experiment C2.4.1 Find Rejection Times for specified forecast/truth pairs using Ignorance.

$\lambda = 0.75$, $\gamma = 0.75$ (these are chosen to keep run times lower)

$\mu = 1$, $\sigma^2 = 0.65$

The available distributions are constrained to (1) lie inside the triangle defined above and (2) be defined by grid points of the form $(\frac{a}{4}, \frac{b}{4})$, where $a, b \in \{0, 1, 2, 3, 4\}$. The exception to this is the point $(0, 0)$ which represents a Pareto distribution. The Pareto distribution requires that values less than its defining parameter be impossible; which leads to infinite Ignorance scores for some observations since the other distributions considered allow any value greater than or equal to zero. To avoid this the Pareto was not tested and a **'HybridPareto'** , defined as having the PDF $f_{(0.025,0.025,0.95)}$, is used in its place. Hence the point $(0, 0)$ is replaced by $(0.025, 0.025)$ in the triangle. The forecast is always denoted by a blue dot. All possible truth distributions within the available distribution space are tested (except where the forecast and truth are the same).

$N_{obs} = 2^9$, $M = 2^{10}$

Skill Scores = {Ignorance}

Experiment C.2.4.2 Find Rejection Times for multiple skill scores.

As for C2.4.1 except for the following:

$\underline{p}$ restricted to only consider a Gamma forecast (i.e. $f_{(0,1,0)}$).

Skill Scores = {Ignorance, Naive Linear, Proper Linear, Spherical}

Results for experiment C.2.4.1 Figure 2.14 shows the Rejection Times plotted at the coordinates of the two dimensional weight combination that determines the true underlying distribution. For example, in the triangle in column two and row four of the graphic, the forecast is $p = f_{(0.25,0.75,0.00)}$. Consider the truth $q = f_{0.25,0.25,0.50}$, this has Rejection Time $RT(\underline{q}, \underline{p}, 0.75, 0.75) = 127$. The bottom graphic in figure 2.14 shows the results for this combination of truth and forecast for 11 different

seeds this level of uncertainty would not alter the key conclusions.

Some key findings from the graphics are:

- The further the truth is from the forecast the shorter the time it takes to reject the forecast system.
- If the forecast is close to a HybridPareto and the outcome is drawn from a different distribution - then the method rejects it quickly whereas it takes longer to reject Gamma-like forecasts.
- Some truth distributions did not lead to a rejection at all within the timeframes considered (2048 time steps, shown as 9999 in the graphic).

Results for experiment C.2.4.2 Figure 2.15 shows Rejection Times. Some key observations are:

- An initially surprising result, all the scores apart from the Ignorance score have the same Rejection Times. This is because the extensions of the Naive Linear score all involve integral terms that are constant for a given forecast, these cause the Skill Gap of these related scores to be scalar multiples of one another and hence the same Rejection Times arise. This is proved on page 120 below.
- For all scores, the further truth is from the forecast the faster the forecast system will be rejected.
- Ignorance gives a shorter Rejection Time when the truth is HybridPareto - but the Naive Linear, ProperLinear and Spherical do better (in some cases much better) when the truth is Lognormal like. This leads to a key conclusion that **using multiple scores would be useful in some contexts**.

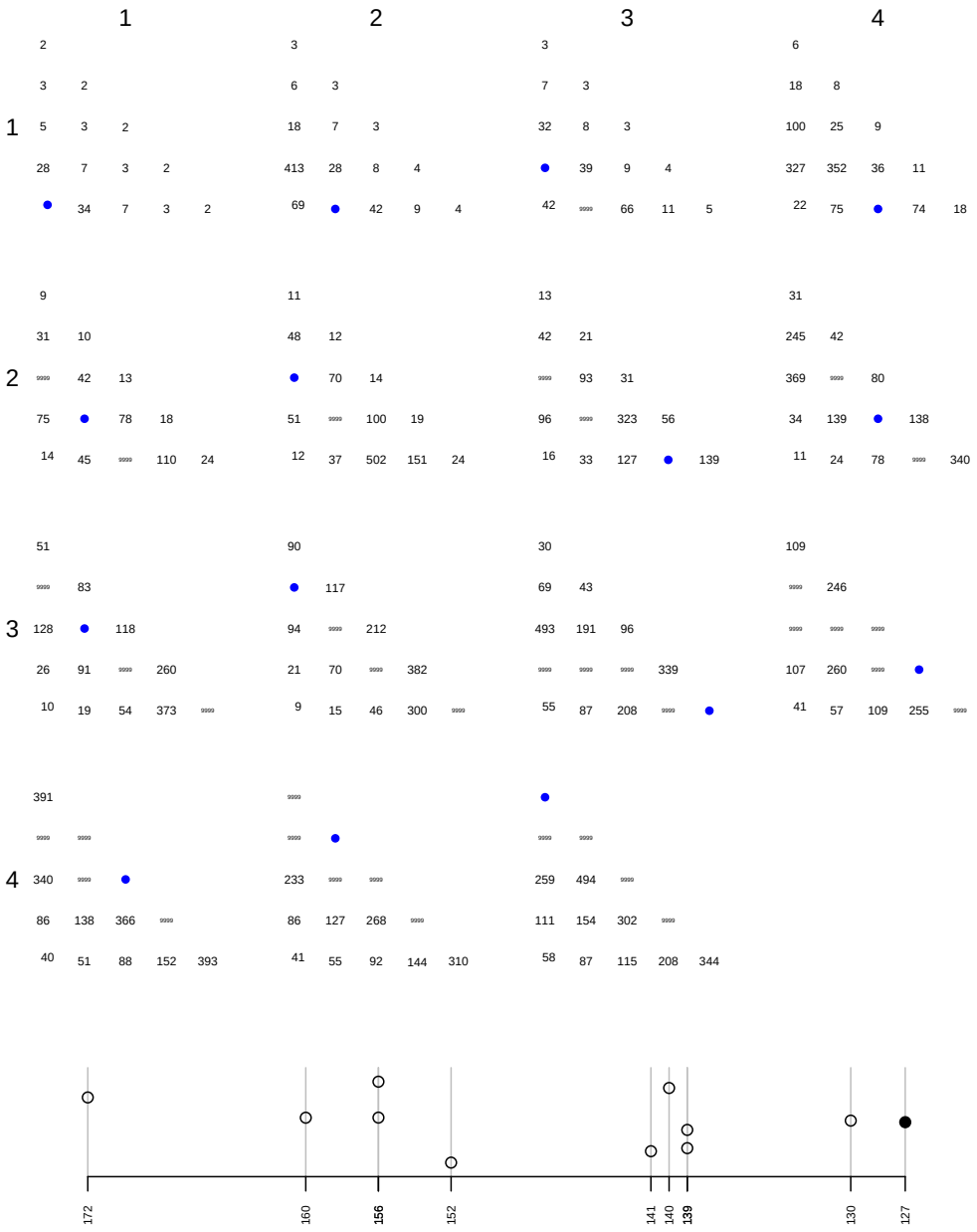


Figure 2.14: Rejection times as truth and forecast (blue dot) vary over the available weights - for fixed mean ($\mu = 1$) and variance ($\sigma^2 = 0.65$). 9999 denotes non-convergence within 2048 observations. The bottom plot illustrates the degree of sampling error by considering 10 different seeds when the forecast is (0.25,0.75,0) and truth is (0.25, 0.25, 0.5) (from the triangle in column 2 and row 4 of the top graphic), the black filled dot shows the results from the seed used in the top graphic, the hollow plot characters show 10 other seeds - the vertical height allows duplicate cases to be shown without overlap.

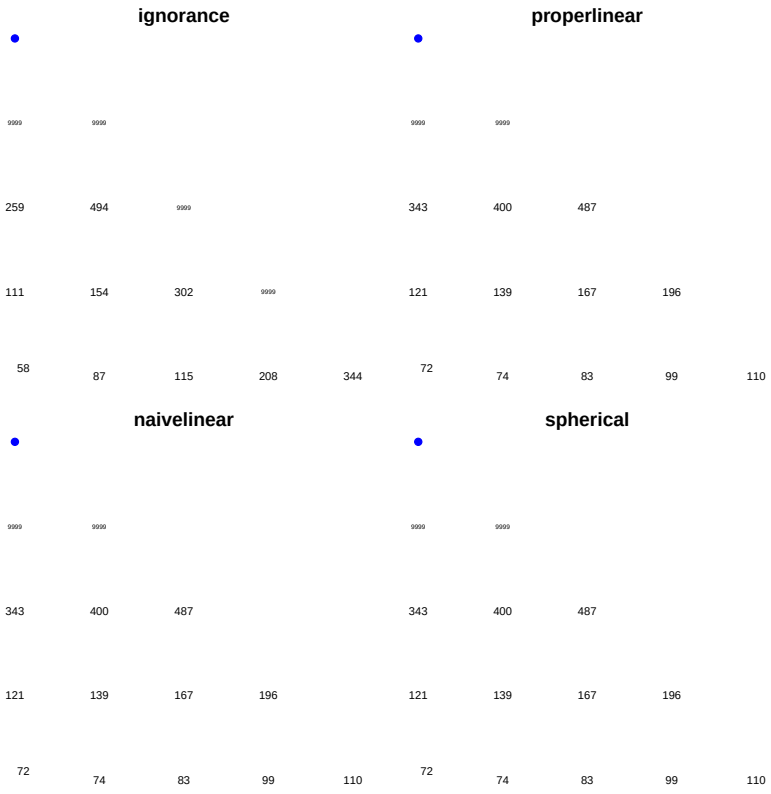


Figure 2.15: Rejection times for different score types, for forecast $f_{(0,1,0)}$ (Gamma distribution, denoted by a blue dot). Note that the Rejection Times for the Proper Linear, Naive Linear and Spherical scores are all the same. When observations are drawn from a Hybrid Pareto distribution (bottom left vertex of triangle) the Ignorance score rejects the forecast after 58 observations compared to (and faster than) 72 for the other scores. When the observations are drawn from a Lognormal (bottom right vertex) distribution, however, the Ignorance score required 344 observations to reject the forecast compared to 110 for the other scores. **This illustrates that there are situations where using multiple proper skill scores will be informative.** Values of 9999 (in small font) show cases where the forecast is not rejected within the maximum number (2048) of observations tested.

Demonstration why the Rejection Times are the same for Naive Linear, Proper Linear and Spherical scores In the following let G_S denote the Skill Gap for score type S. Then G_{NL} refers to the Naive Linear score, G_{PL} to Proper Linear and G_{SP} to the Spherical score. When the Naive Linear score (equation 5.1) is substituted, equation 2.54 becomes:

$$G_{NL}(t) = \frac{-1}{t}\sum_{i=1}^{t} q(X_i) + \int q(x)q(x)dx \tag{2.56}$$

Lemma 1: If $G_S = \alpha G_R$ for two scores S and R then the Rejection Time from S will be the same as R. This is because the Rejection Time is calculated as the intersection point of two quantile lines for the Skill Gap. Since, for every sequence of observations, the Skill Gap for S is a scalar multiple of R its Rejection Time diagram will simply be a stretch in the y-axis direction - this doesn't affect where the lines cross and so Rejection Times will be the same.

Lemma 2: Proper linear $G_{PL} = 2G_{NL}$. When the Proper Linear score (equation 2.6) is substituted into equation 2.54 the equation becomes:

$$G_{PL}(t) = \frac{1}{t}\sum_{i=1}^{t}(\int q^2(z)dz - 2q(X_i)) - \int q(x)\{\int q^2(z)dz - 2q(x)\}dx \tag{2.57}$$

Now $\int q^2(z)dz$ is just a constant so:

$$G_{PL}(t) = \frac{t}{t}\int q^2(z)dz + \frac{-2}{t}\sum_{i=1}^{t}(q(X_i)) - \int q^2(z)dz \int q(x)dx + 2\int q^2(x)dx \tag{2.58}$$

Because q is a PDF $\int q(x) = 1$, hence the two $\int q^2(z)dz$ terms cancel out. So the Skill Gap reduces to:

$$G_{PL}(t) = 2(\frac{-1}{t}\sum_{i=1}^{t} q(X_i) + \int q(x)q(x)dx) \tag{2.59}$$

This is exactly double the expression for the naive linear score stated in equation 2.56. So $G_{PL} = 2G_{NL}$.

Lemma 3: Spherical $G_{SP} = \frac{1}{\kappa}G_{NL}$, where $\kappa = (\int_{-\infty}^{\infty} q^2(z)dz)^{\frac{1}{2}}$. When the spherical score (equation 2.12) is substituted into the Skill Gap equation we see:

$$G_{SP}(t) = \frac{-1}{t}\sum_{i=1}^{t}\frac{q(X_i)}{(\int_{-\infty}^{\infty} q^2(z)dz)^{\frac{1}{2}}} + \int q(x)\frac{q(x)}{(\int_{-\infty}^{\infty} q^2(z)dz)^{\frac{1}{2}}}dx \tag{2.60}$$

κ, defined above, is a constant so:

$$G_{SP}(t) = \frac{1}{\kappa}\{\frac{-1}{t}\sum_{i=1}^{t} q(X_i) + \int q(x)q(x)dx\} \tag{2.61}$$

This is a scalar multiple of the Naive score, so $G_{SP} = \frac{1}{\kappa}G_{NL}$.

Corollary: The Rejection times are always the same for Naive Linear, Proper Linear and Spherical scores Let $R = NL$ then lemma 2 shows that the condition of lemma 1 is met for $S = PL$ where $\alpha = 2$ and lemma 3 shows the condition of lemma 1 is met for $S = SP$ where $\alpha = \frac{1}{\kappa}$. So the Rejection Times for NL,SP and PL are the same.

2.7 Improving skill with climatology blending

The intention of this section is to test the extent to which Climatology Blending, introduced by Brocker and Smith [30], can improve forecast skill for various levels of forecast quality (defined on page 123 below). A climatology [30] is an empirical distribution based on past observations over a defined period of time. Let a forecasted variable be denoted y, let $f_u(y)$ be the climatology of the variable and $p(y)$ a forecast. The Blended Forecast is defined as a weighted average $r(y)$ of these two PDFs as follows:

$$r(y) := \alpha p(y) + (1 - \alpha) f_u(y) \tag{2.62}$$

Where $0 \leq \alpha \leq 1$. The weighting variable α, as well as any parameters[5] used to create the forecast $p(x)$, are found simultaneously to minimise a chosen average skill score, given multiple observations. Brocker and Smith's method ensures that $r(y)$ will not be zero for any observations unless they fall outside of the union of climatology values and the support of $p(y)$; this reduces the chance of very large or even infinite scores for some score types. If infinite scores do arise an observation has occurred that is both outside of the historical record and also beyond what the forecast thought possible. It is arguable that an infinite penalty is appropriate in such a case; it is surely important to draw attention to such anomalies.

[5]For example, if Kernel Dressing, as described in equation 2.37, is used to generate the forecast PDF then the kernel width parameter σ_m is one of the parameters of $p(y)$. Note, however, that parameters used to generate the ensemble are not included in this process.

In an insurance context it is common to use Climatology (f_u) as a forecast when the history of past insurance claims, and averages arising, are used to calculate expected claims in the future. This assumes the claims process is stationary and that the data contains sufficiently many observations to produce an appropriate estimate of the average. Often the risk environment in future is not expected to be identical to the past and so an adjustment is made. The purpose of any forecast, insurance or otherwise, is to describe the distribution of potential outcomes in a way that is more skillful than relying on climatology alone. This additional skill is measured by the **'Relative Skill Score'** which is the difference between the score for a given forecast/observation pair and the score that arises using climatology. Specifically:

Definition of Relative Score For a given score S with a forecast p, climatology f_u and observation v:

$$\text{Relative Score} = S(p, v) - S(f_u, v) \tag{2.63}$$

Definition of Discrete Forecast A discrete forecast is a single real number $\hat{y}$ that is an estimate of the value of an observed variable y in future. I.e. a forecast ensemble with one member only.

The algorithms below are designed to create a forecast which has a controllable level of 'quality'. Let $u \sim U(a, b)$ denote a sample from a uniform random variable with lower limit a and upper limit b, the character Q is used below to denote a random real number between zero and 1 (i.e. $Q \sim U(0, 1)$).

Algorithm to create observations and discrete forecasts of desired quality (See figure 2.16)

- Choose a real number $\beta \in [0, 1]$
- Let $i \in \{1, ...N_{obs}\}$
- **Sample potential observations** Let $v_{i,1}, v_{i,2} \sim f_u$ be two sample values from the climatology distribution f_u
- **Create discrete forecast** A discrete forecast ($\hat{y}_i$) is created as follows:
  - let $X \sim \text{Lognormal}(\mu_L, \sigma_L)$, so that $E(X) = 1$

- let $\hat{y}_i = \frac{v_{i,1}}{X}$ (note the use of $v_{i,1}$ specifically here)

- **Create observation**: The observation v_i is controlled by a parameter β, as follows:
  - If $Q < \beta$ then set $v_i = v_{i,1}$;
  - if $Q \geq \beta$ then $v_i = v_{i,2}$. This defines a 'forecast bust' where the forecast and observation are likely to be very different.

- Repeat to Generate multiple pairs (v_i and $\hat{y}_i$) for $i \in 1, ...N$

Definition of Forecast Quality In this experiment a 'high quality' forecast is one with β close to 1; a 'low quality' forecast is one with β close to zero.

Algorithm to create Kernel Dressed and Blended forecasts (See figure 2.17)

- **Kernel dressed forecast** Define $p_{i,1}(y) \sim N(\hat{y}_i, \sigma_1^2)$ where σ_1 is the optimal score estimate given the observations $\{v\}$.
- **Blended forecast** Define $p_{i,2}(y) \sim N(\hat{y}_i, \sigma_2^2)$, and define r_i by equation 2.62. Iterate to find the minimum of the average score $S_{avg} = \sum_{i=1}^{N_{obs}} S(r_i, v_i)$ over all observations using constrained optimisation on the variables σ_2 and α simultaneously;

In words, the two algorithms, in sequence, result in forecasts $r_i(y)$ and $p_{i,1}(y)$ that are, with probability β, closely related to the observation v_i, or, with probability $1 - \beta$, independent from the observation. The three forecasts (f_u, $p_{i,2}$ and r_i) are illustrated in figure 2.18.

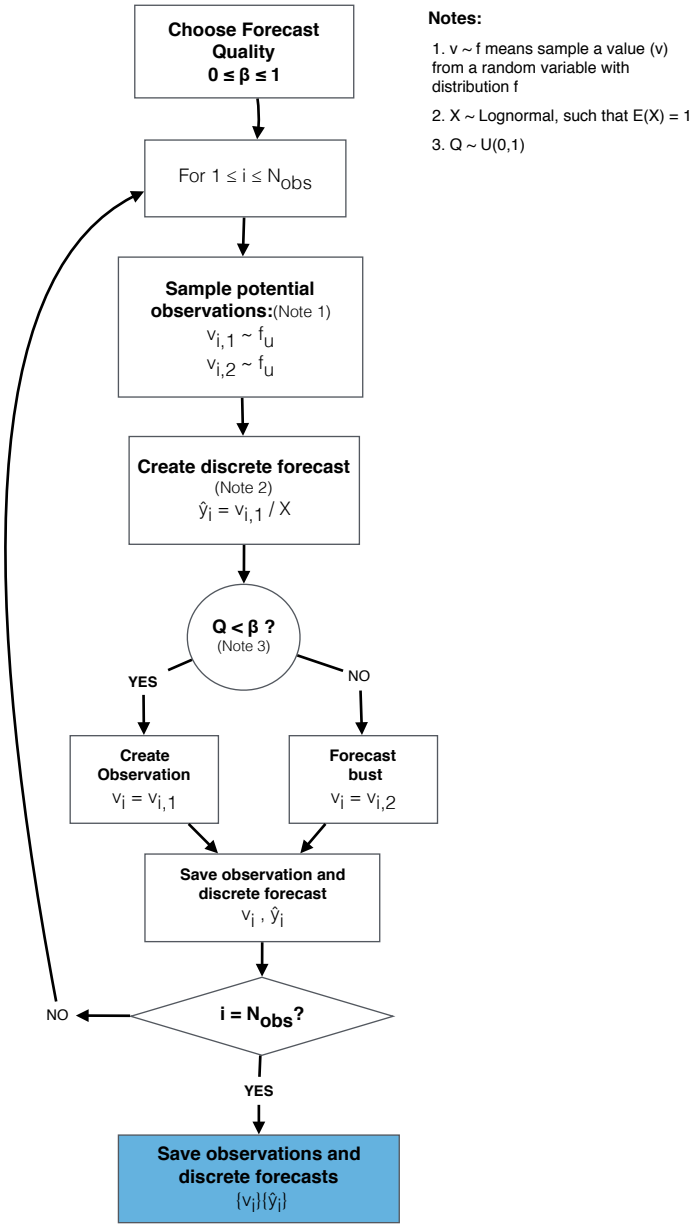


Figure 2.16: Example 2.5: Flowchart for process to create observations and discrete forecasts.

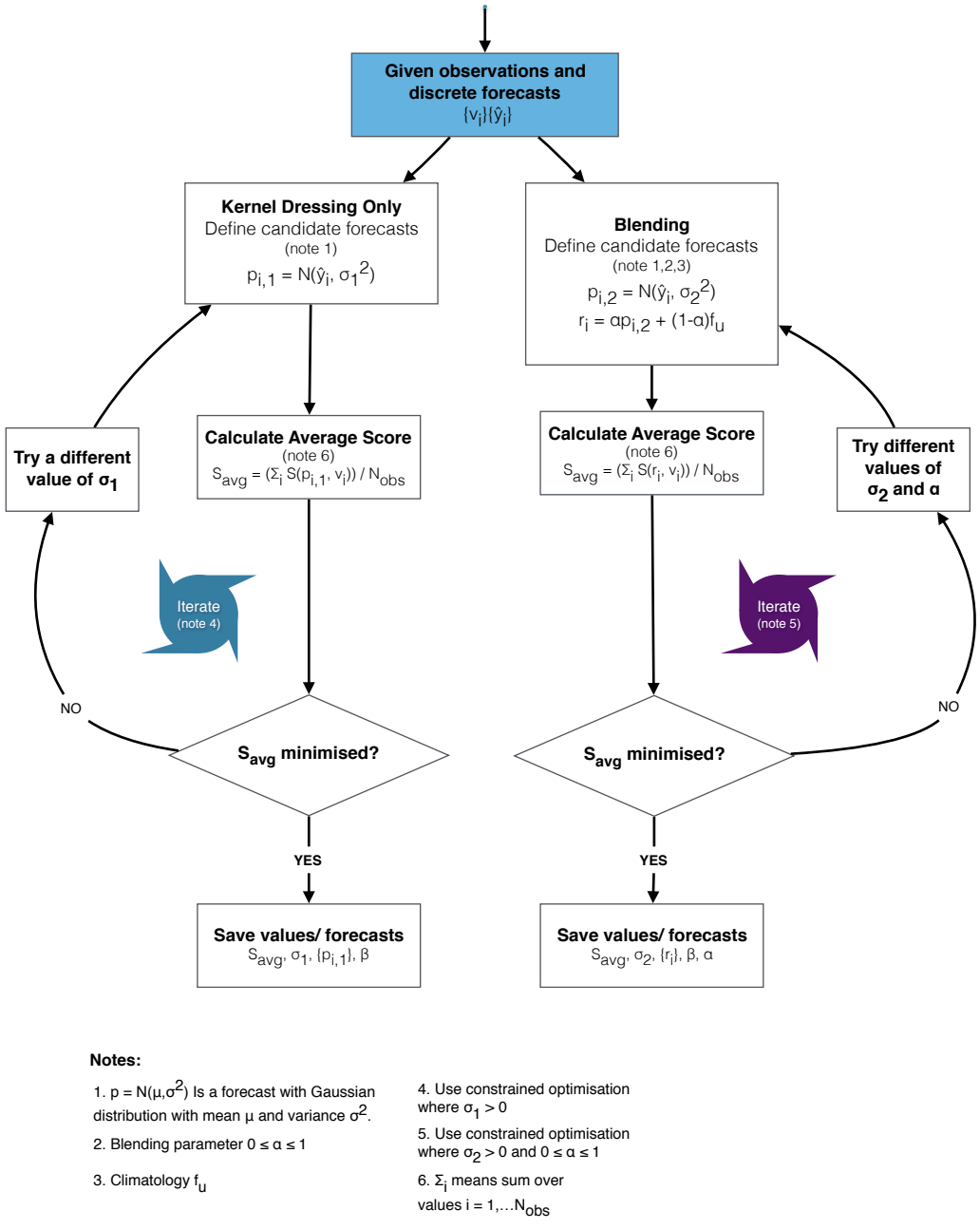


Figure 2.17: Example 2.5: Flowchart for process to create kernel dressed and blended forecasts from observations v_i and discrete forecasts $\hat{y}_i$.

The following experiment is carried out to test whether blending adds skill over kernel dressing alone and whether this occurs for various levels of forecast quality. The experiment will also test whether such forecasts do better than climatology.

Experiment C2.5 Testing Blending.

The following values of forecast quality are tested: $\beta \in \{0, 0.1, ...1\}$

Climatology f_u is defined using equation 2.37 with the following variables: $S_u = \{X_1, ...X_M\}$ where $X \sim N(\mu_S, \sigma_S^2)$ and $M = 2^{13}$. $\mu_S = 5$ and $\sigma_S = 5$. The kernel width $\sigma_u = 1.3$.

The Lognormal variable (X) is defined with $\sigma_L = 0.1$ and $\mu_L = -\frac{1}{2}\sigma_L^2$.

$N_{obs} = 2^7$

Skill Score = Ignorance (chosen because it has performed well in the experiments described in this chapter)

In the calculation of the blended forecasts r_i the values of α and σ_m are chosen using a constrained Nelder Mead algorithm (where $0 \leq \alpha \leq 1$ and $\sigma_m > 0$)

Results for experiment C2.5 The key results from the Blending Experiment are:

- Figure 2.19(c) shows that as the forecast quality increases ($\beta \to 1$) the skill of both r_i and $p_{i,1}$ increases (since the Relative Ignorance becomes more negative). The blended forecast $r_i(y)$ is more skillful than the kernel dressed forecast $p_{i,1}$ for all values of β and is more skillful than climatology (because the Relative Ignorance is negative for all values of β). The kernel dressed forecast is only more skillful than climatology for $\beta > 0.6$. **Hence in this experiment Blending improves the skill of forecasts at all quality levels (β).**
- Figure 2.19(b) shows that as the forecast quality decreases the kernel width σ_1 widens. This is intuitive as a wider bandwidth will give more probability further from the mean value $\hat{y}_i$ and so will not score so poorly when the observation v_i is far from the mean. Such cases can arise randomly but are likely (with probability $1 - \beta$) when $Q > \beta$ in the Blending Experiment Algorithm

above. The behaviour of σ_2 (the kernel width in the blended forecast) is more complex; for very poor forecast quality ($\beta \leq 0.2$) the kernel width widens to compensate for forecast busts; but otherwise this parameter is quite stable around a value $\sigma_2 \approx 2$ though it increases slightly for larger values of β;

- Figure 2.19(a) shows the value of α against β. As expected, as the forecast quality increases, the weight put on the forecast (α) increases. The increase in α is monotonic which helps to explain the behaviour of σ_2 discussed above: the climatology has a wide distribution and so the effect of putting more weight on the forecast $p_{i,2}$ is to narrow the blended distribution (i.e. put more probability near the mean). This has broadly the same effect as narrowing σ_2 and appears to provide better improvements in skill score - this has not been explored further.

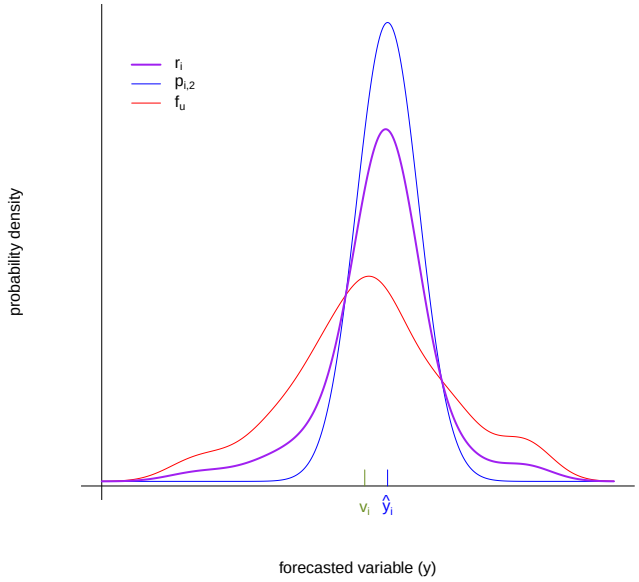


Figure 2.18: Example 2.5: Climatology f_u (red) along with a particular observation v_i (green) a kernel dressed forecast ($p_{i,2}(y) \sim N(\hat{y}_i, \sigma_2^2)$) is shown in blue. The blended forecast $r_i(y)$, for an illustrative value of $\alpha = 0.6$, is shown in purple. Note that the blended forecast assigns greater probability to the forecast variable in the left and right hand tails of the distribution.

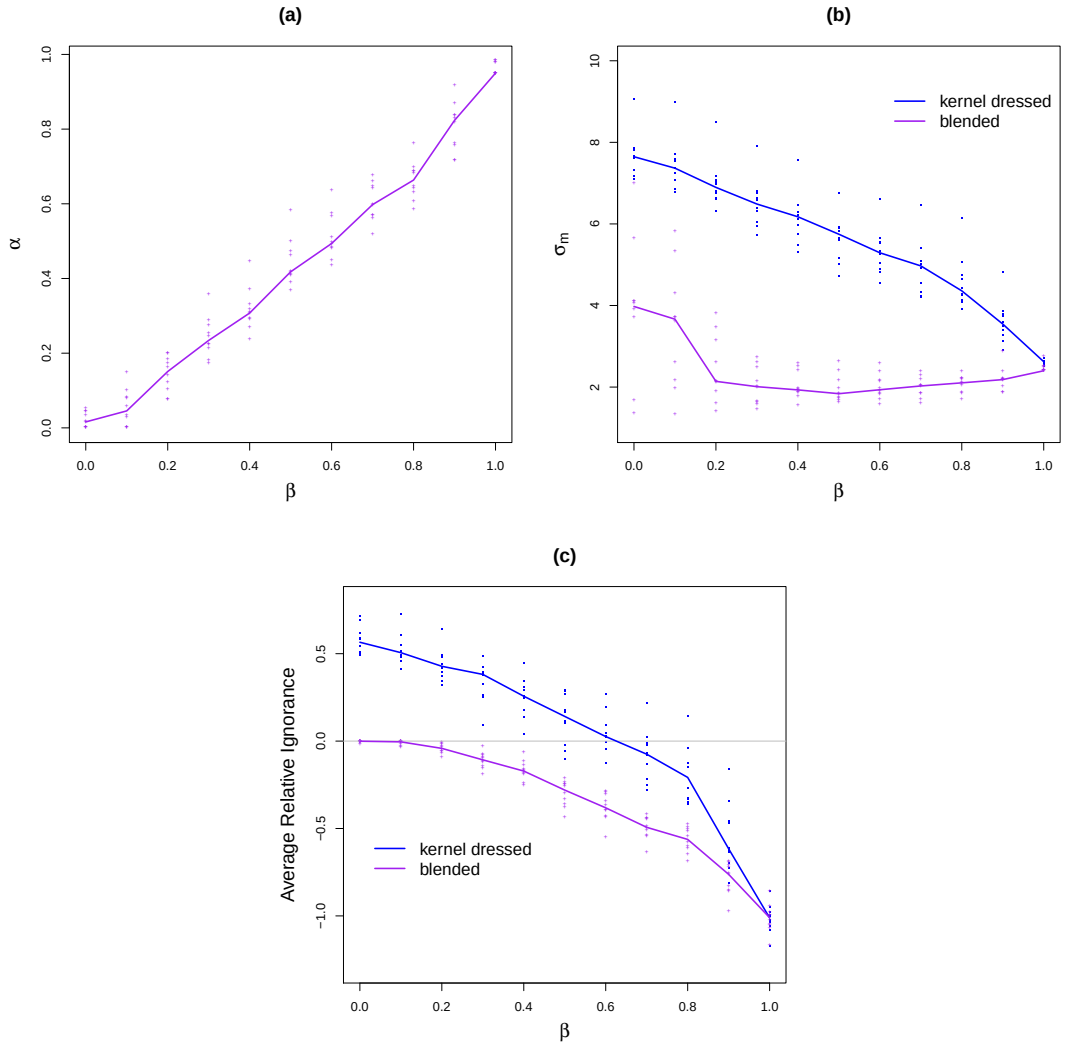


Figure 2.19: Blending example: Figure (a): Relationship between the blending parameter α and the quality of the forecast β; as the forecast quality increases ($\beta \to 1$), then the weight put on the forecast increases ($\alpha \to 1$). Figure (b) shows the size of the kernel bandwidth (σ_m) against β for a blended forecast (purple) and kernel dressed forecast (blue). For the kernel dressed forecast as the quality of forecast improves the bandwidth narrows; for the blended forecast the bandwidth initially narrows but then slightly widens again. Figure (c) again compares the blended and kernel dressed cases, showing the relative Ignorance versus forecast quality β; the blended forecast shows better skill than climatology (Relative Ignorance negative) for all values of β; the kernel dressed forecast only shows skill for $\beta > 0.6$. In each experiment results are produced for 10 different random seeds and the resulting values are plotted using points; for each value of β the median values are joined together to form a line plot.

2.8 Conclusions

This chapter has investigated the properties of various skill scores most relevant in the insurance sector. A new property has been introduced called 'Feasibility'. A score is not Feasible if it evaluates highly a forecast that places a high probability on an event that has a low frequency of occurrence in practice. The CRPS and MSE scores are shown to not be Feasible. The Ignorance score performs very well in two ranking tests. A third ranking test uses the Skill Gap to test how quickly a chosen skill score will reject a forecast system. The Rejection Time of the Ignorance score can be more or less than the Proper Linear, Spherical and Naive Linear scores, suggesting that the use of multiple proper[6] scores will be helpful in practice. Finally Climatology Blending is explored using the Ignorance Score and is shown to improve forecast skill in one particular experiment. The next chapter makes use of the well known Lorenz 96 dynamical system [154] and uses the results in this chapter to improve various forecasts of this system.

[6] noting that the Naive Linear score is not proper and should be passed over in favour of proper scores

Chapter 3

Exploring Lorenz 96

'***The relevance of mathematically defined systems cannot be too strongly emphasised; much of what we know, or believe that we know, about real systems has come from the study of models.***'

Ed Lorenz 1996 [155]

This chapter explores a well known dynamical system of differential equations introduced by Ed Lorenz [155] in 1995[1] as a simple yet rich system to examine predictability. In this chapter both systems and models of these systems will be introduced. The models will be run with various parameterisations to produce forecasts of the systems. Five models are developed for each system, to test whether different modelling approaches do better.

Climatology Blending (described in Chapter 2) is used to create best scoring forecasts. One system (80001[2]) is explored in detail. The process is repeated for five different systems (80002, to 80006) to assess whether the results are particular to system 80001 or more general. The results of this chapter are then used to guide further exploration and applications to insurance in Chapter 4.

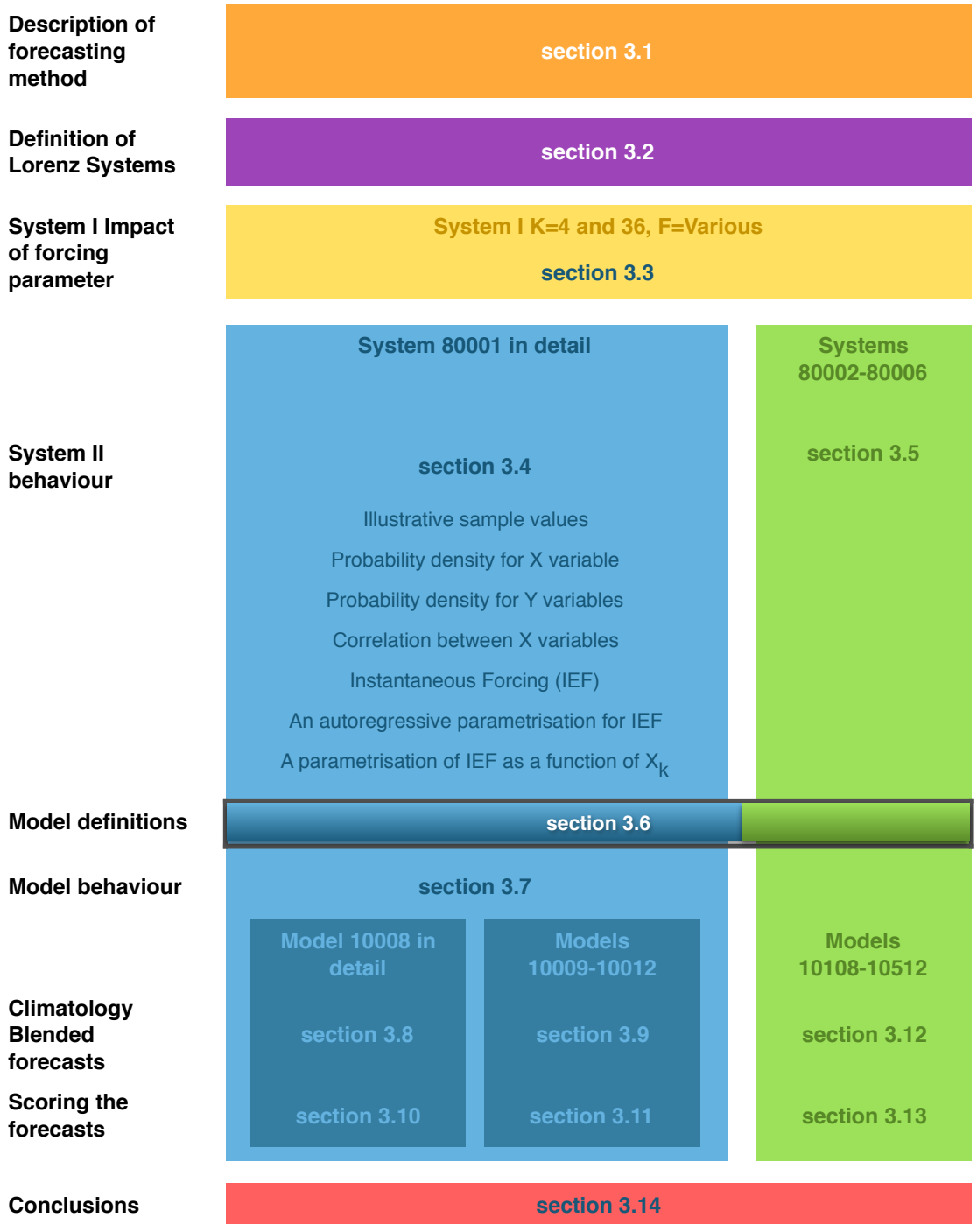
[1]Ed Lorenz presented the two systems of ODEs discussed in this chapter at a Seminar Held at ECMWF on Predictability in 1995 but the proceedings of the meeting were published in 1996. For this reason the systems are known as both Lorenz 95 and Lorenz 96 in the literature; here the latter is used.

[2]To enable future extensibility of this work and to ensure unique IDs each system and model is given a 5 digit ID where the leading digit is different to emphasise systems (leading digit 8) and models (leading digit 1), these are defined in sections 3.2 and 3.6.

The original elements of this chapter are believed to be:

- Illustration of the impact of an increasing forcing parameter on various quantiles of the X variables including a comparison with the quantile relationship for lower forcing levels and also comparison with a Gaussian distribution with the same mean and standard deviation at each level of forcing;
- Exploration of a model using kernel smoothers to derive a functional relationship between the effective forcing in the system and the value of the corresponding X variable, this model has the best skill scores;
- Exploration of a model which uses an AR(4) process to parameterise one of the key variables, this model performs poorly, leading to the conclusion that the lack of conformity with the dynamics of the system is a larger factor in model quality than the closer adherence of one of its parameters to the 'correct' statistical behaviour;
- Use of the derived blending parameter values to: compare and discuss different systems (each with five models); and to explore the effect of the forcing parameter in the systems.

Chapter Structure The figure below sets out the structure of this chapter. The method used to create forecasts is described first because it introduces terms used thereafter. Next the two Lorenz systems are defined. Section 3.3 explores the impact of a key parameter in Lorenz System I. Six parameterisations of Lorenz System II are described and given IDs 80001, ...80006. System 80001 is used throughout the chapter to illustrate detailed methods and results; then results are shown for the other systems 80002-80006. Five models for each of these systems are defined in section 3.6, model behaviour is then illustrated for the five models of system 80001. Climatology blending (Chapter 2, equation 2.62) is carried out for each forecast and skill scores are calculated in each case. This is first derived in detail for system 80001 and one of the models (ID 10008), is then compared with the other models of this system (10009... 10012). Finally the blending parameters and skill scores are compared for the remaining systems and models.


Description of forecasting method
section 3.1
Definition of Lorenz Systems
section 3.2
System I Impact of forcing parameter
System I K=4 and 36, F=Various
section 3.3
System II behaviour
System 80001 in detail
section 3.4
Illustrative sample values
Probability density for X variable
Probability density for Y variables
Correlation between X variables
Instantaneous Forcing (IEF)
An autoregressive parametrisation for IEF
A parametrisation of IEF as a function of X_k
Systems 80002-80006
section 3.5
Model definitions
section 3.6
Model behaviour
section 3.7
Model 10008 in detail
Models 10009-10012
Models 10108-10512
Climatology
Blended forecasts
section 3.8
section 3.9
section 3.12
Scoring the forecasts
section 3.10
section 3.11
section 3.13
Conclusions
section 3.14

3.1 Description of forecasting method

Motivating analogy The following example will motivate the general description below. Consider an atmospheric variable such as temperature at a given location. This variable changes continuously but is observed discretely (say twice a month) over a given year. The atmosphere is the 'system' using the terminology above. Models are initialised at the start of the year using the observed temperature and values are sampled with the same frequency so that values can be compared. The observed temperatures are subject to observational error (such as instrument failures, human errors etc) and to allow for this the initial conditions for the models are sampled from a distribution around the observed temperature so that a range of predictions are produced. The outputs from these different initial conditions comprise an 'ensemble' whose values at each observation time can be compared to the system.

Definition of Initialisation Time Consider a continuous system whose variables are indexed by time t, let estimates of variables from the system be observed at time t_0. If a model is run with initial conditions based on the observed variables then the **'initialisation time'** is t_0.

Definition of Initial Condition The **'initial conditions'** are the starting values for a given model at time t_0.

Definition of Period A **'period'** is the length of time τ between model initialisation times. Equivalently the length of time for which a model is evolved forward before being reinitialised.

Definition of Lead Time Given model outputs at time t, within a given period, with initialisation time t_0 then the **'lead time'** is defined as $t - t_0$.

Table 3.1: Concrete example to illustrate forecast definitions

Definition	Example
Initialisation time	00:00 on 1 January 2015
Period	2015 (one year)
Lead time	4 months (i.e. forecasts of variables at 00:00 on 1 May 2015)

Definition of model ensemble Let the variables of interest from the system be denoted $X_k(t)$ for $k \in 1, ...K$. Consider a model of the system, and N_{ens} sets of initial conditions $\{\{x_k\}_{k=1,...K}\}^{i=1,..N_{ens}}$ with initialisation time 0. Let the values from the model with initial condition i be denoted $\hat{X}(t)^i_k$. Then the set of values $\{\hat{X}(t)^i_k\}$ at a given lead time t is known as an **'ensemble'**.

Process to create initial conditions The initial conditions used in this chapter are created as follows. This will be referred to as a **'Gaussian Initial Condition Ensemble'**.

- Extract the true positions of the X variables from the system at the initialisation time t_0, $(x_{k,0} = X_k(t_0))$;
- Perturb these using dynamical noise (IID Gaussian) - to create an **'observed initial condition'** $(\hat{x}_{k,0} = x_{k,0} + \epsilon)$, where $\epsilon \sim N(0, \sigma^2)$;
- Create an ensemble of model initial conditions also using a Gaussian dynamical noise centred around the observed initial condition with standard deviation σ^2 (i.e. the same as used to create the observed initial condition), specifically: $\hat{x}^j_{k,0} = \hat{x}_{k,0} + \epsilon^j$, where $\epsilon^j \sim N(0, \sigma^2)$ and $\text{cov}(\epsilon^j_a, \epsilon^j_b) = 0 \ \forall a, b$.

Note that the true positions $x_{k,0}$ are unknown in practice but are used as an interim step here to create observations. This method ensures that the mean observed initial condition will be centred on the true initial condition. Also, the ensemble values will be centred on the observed initial condition and hence also on the true initial condition on average, so there is no inherent bias in the starting point. The use of the same standard deviation in both perturbation processes guarantees this for a Gaussian distribution. With this 'inverse noise method' the initial conditions created may not lie on the attractor of the model [69] which can lead to less informative

results. Methods to correct for this are currently being explored [268] but these are not considered here.

3.2 Lorenz 96 systems I and II

A **'system'** of Ordinary Differential Equations (ODE) is formed by specifying two or more ODEs with some dynamical variables in common, each being a function of an independent variable (such as 'time'). Lorenz introduced the following two systems in 1995 at a conference followed by a paper [155] in 1996.

Lorenz System I Lorenz defined the following system in equation 3.1 of his 1996 paper.

$$\frac{dX_k}{dt} = (X_{k+1} - X_{k-2})X_{k-1} - X_k + F \tag{3.1}$$

Where, $k \in 1, ...K$ for some integer K which determines the number of X variables within the system. k is defined mod K so that $X_k = X_{k-K}$. The **'forcing'** F is an exogenous parameter which prevents the solution tending to zero. In Lorenz's original paper F is constant, but can be extended (e.g. [11, 129, 239, 271–273]) to be a function of time or other variables. An Initial Condition for this system is a set of real numbers $\{x_1, ...x_K\}$ which are the starting values for the X_k variables.

Lorenz System II Lorenz defined the following system in equations 3.2 and 3.3. of his original paper.

$$\frac{dX_k}{dt} = (X_{k+1} - X_{k-2})X_{k-1} - X_k + F - \frac{h_x c}{b}\sum_{j=1}^{J} Y_{jk} \tag{3.2}$$

$$\frac{dY_{j,k}}{dt} = cb(Y_{j-1,k} - Y_{j+2,k})Y_{j+1,k} - cY_{j,k} + \frac{h_y c}{b}X_k$$

where, again F is a forcing term which prevents the solution tending to zero. h_x and h_y determine the strength of **'coupling'** between the X and Y variables, i.e. the extent to which the value of the Y variables influence the X variables and vice versa. In Lorenz's original paper the coupling parameter $h = h_x = h_y$ was not split into its x and y components, the presentation here follows Smith [239] and allows the influence of the Y's on the X variables to be controlled without affecting the influence of $X's$ on the Y variables (except through second order effects). b and

c both affect the strength of coupling but also affect the Y dynamics in different ways. Lower values of b reduce the number of oscillations of the X series materially but less so for Y. Lower values of c slightly increase the number of oscillations of the X variables whilst materially reducing those of the Y variables. The total number K, of X variables can be any integer value and the indices k are calculated mod K, so that $X_{K+1} = X_1$. The variables X_k form a system and can be thought of as sitting on a circle. The values $Y_{j,k}$ also form a system and sit on a circle so that $Y_{1,1} = Y_{JK+1,K}$ and in general $Y_{j-J,k} = Y_{j,k-1}$ and $Y_{j+J,k} = Y_{j,k+1}$. The system is illustrated in figure 3.3 which is an actual realisation of system 80001 after Smith [239]. The X values are shown as perturbations from their mean value (red circle) and the Y values from their mean value (blue circle). The X and Y values are scaled to have unit standard deviation in the graphic. The Ys are thought of as small scale dynamics which feed into the larger scale X dynamics. Note that for each X_k there are J variables $Y_{1,k}, ...Y_{J,k}$ which have an influence. This is referred to as the **'J-block'** below. An 'Initial Condition' for this system is two sets of real numbers $\{x_1, ...x_K\}$ and $\{y_{1,1}, ...y_{J,K}\}$ which are the starting points for the X_k and $Y_{j,k}$ variables respectively. The following describes the various parameterisations of Lorenz System II that are used in this and the next chapter.

Systems 80001 - 80006 Lorenz System II parameterisations

The following parameterisations are for Lorenz System II (equation 3.2).

Each parameterisation is given a unique ID code of the form 8000[x] where [x] $\in \{1, ..6\}$. These IDs will be used in descriptions and various plots.

In all cases $K = 36$, $J = 10$, $c = b = 10$ and $h_y = 1$. **80001:** $F = 10$, $h_x = 1$

80002: $F = 20$, $h_x = 1$

80003: $F = 20$, $h_x = 0.1$

80004: $F = 10$, $h_x = 0.1$

80005: $F = 9.1$, $h_x = 0.1$

80006: $F = 11$, $h_x = 1$

System 80001 has the same parameterisation used by Lorenz [155].

Values are produced using method 'lsoda' under the R statistical language 'ode' function. Outputs are captured in 0.1 time increments.

Definition of Instantaneous forcing In Lorenz System I the X values are forced only by the term F. In Lorenz System II, however, the Y variables provide additional forcing and the total forcing varies at each point in time. For later use the total forcing is defined below as the **'Instantaneous Effective Forcing'** (IEF):

$$IEF_k := F - \frac{h_x c}{b} \sum_{j=1}^{J} Y_{jk} \tag{3.3}$$

3.3 System I - Impact of forcing parameter

This section explores the impact of the forcing parameter in Lorenz System I, defined in equation 3.1. Figures are created by evolving the system forward from a point on the attractor.

Probability density of X_k Two examples are illustrated in figure 3.1 with $K = 4$ and $K = 36$. The figure illustrates the normalised probability density for each value of F (x-axis) and value of X_1 (y-axis). Densities are normalised, for a given value of K, by dividing each density by the maximum density, this ensures the two plots are on the same colour scale. These reflect the behaviour described in Orrel and Smith [189] but are ordinal in that they show normalised probability densities. The colour represents the probability density of a given point. Red indicates high density and cyan low density. Dark blue shows zero density.

- As F increases the range of X also increases;
- The density initially clusters around a few values suggestive of periodic behaviour - but after a threshold ($F \approx 11.7$ in figure 3.1(a) and $F \approx 4.4$ in figure 3.1(b)) the density plot evens out and X_1 then takes a broad range of values;
- The threshold of periodic behaviour is larger when $K = 4$ - suggesting that systems with larger values of K stop displaying periodic behaviour for lower forcing values.

Impact of F on high quantiles of X_k Figure 3.2 illustrates how the mean and four different quantiles of the values taken by X_1 change as the forcing increases from $F = 5$ to $F = 11$. All figures illustrate the median, 90th, 95th and 99th

isopleths as well as the mean. To create these statistics the system is evolved from a point on the attractor over a time period of length 36.5, the time series is observed in time increments of 0.1. Figure (a) shows the 10th and 90th quantiles of each statistic estimated from 128 different initialisations of the system. Figure (b) shows the average value of the five statistics only (as dots); the solid lines define the higher quantiles for F relative to the values when $F = 5$. Specifically if $Q_p(F)$ denotes the pth quantile value for a forcing value of F, and if $M(F)$ denotes the mean value, then the solid lines are defined as $y_p(F) = \frac{Q_p(5)}{M(5)} M(F)$. Figure (c) shows the average statistic values (as dots) - against the quantiles from a Gaussian distribution with same mean (red dot) and standard deviation as the observed X_1 variables for a given level of forcing. The key findings are:

- Figure 3.2(a) - the mean and median are similar to each other for all values of F and increase by $\approx 60\%$ amount as F increases from 5 to 11, the 99th quantile increases by $\approx 100\%$ as the forcing increases over this range;
- Figure 3.2(b) also shows that the higher quantiles increase faster than the mean, as F grows;
- Figure 3.2(c) - the relationship to a normal distribution with same mean and standard deviation at each F value, remains broadly stable. The 90th %ile is larger in the Lorenz System I and the 99th %ile is smaller;

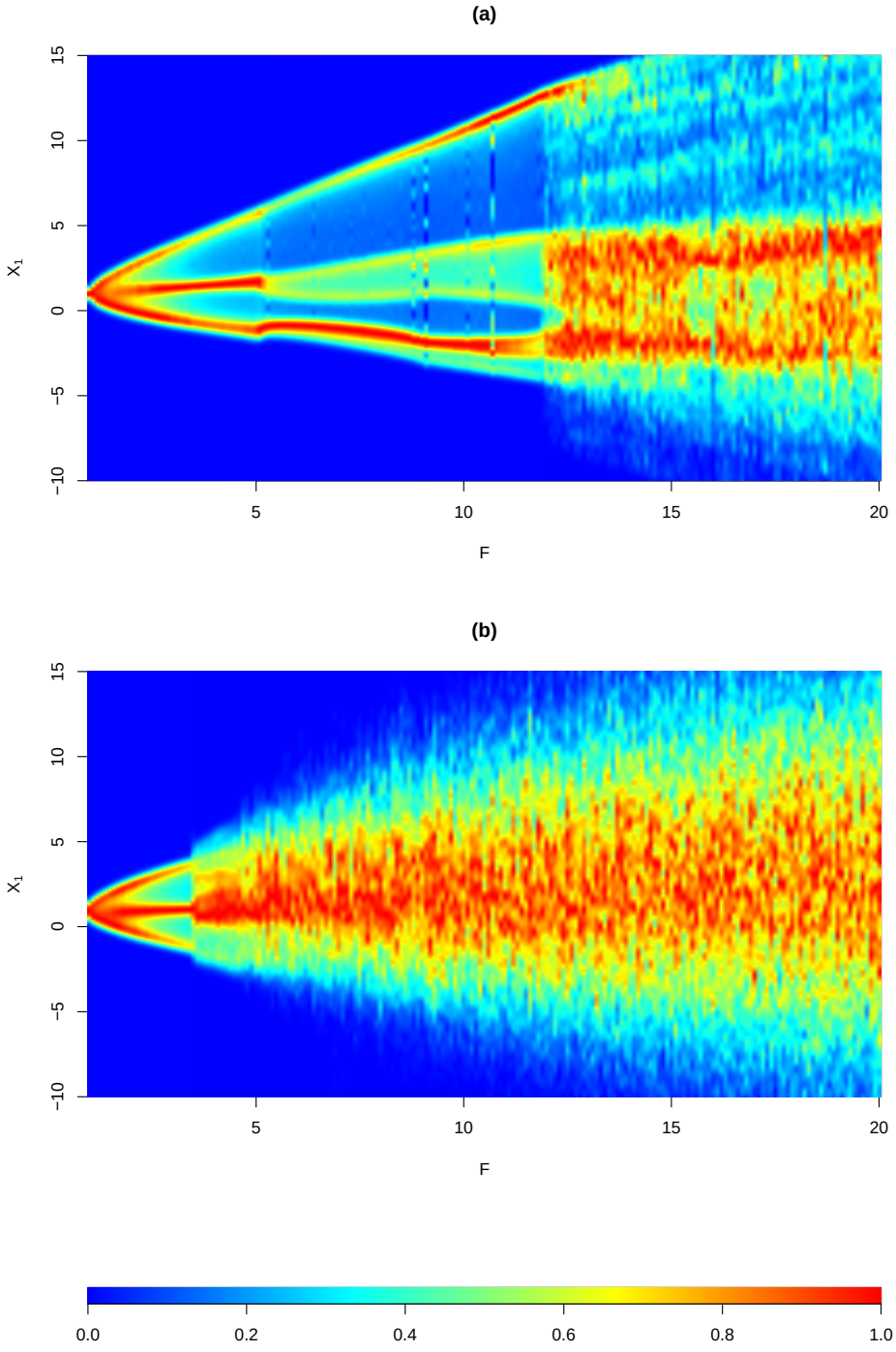


Figure 3.1: Probability density plots for two different values of K (Figure (a) K=4 and Figure (b) K=36). The x-axis shows the value of the forcing parameter (F) and the y-axis shows the value of the X_1 variable in Lorenz System I and the colour denotes the normalised density of observations taking that value, shown in the colour key. Red indicates high density, cyan low density and blue zero density. The density is normalised by dividing by the maximum density for each value of K - this ensures the plots are both on the same colour scale.

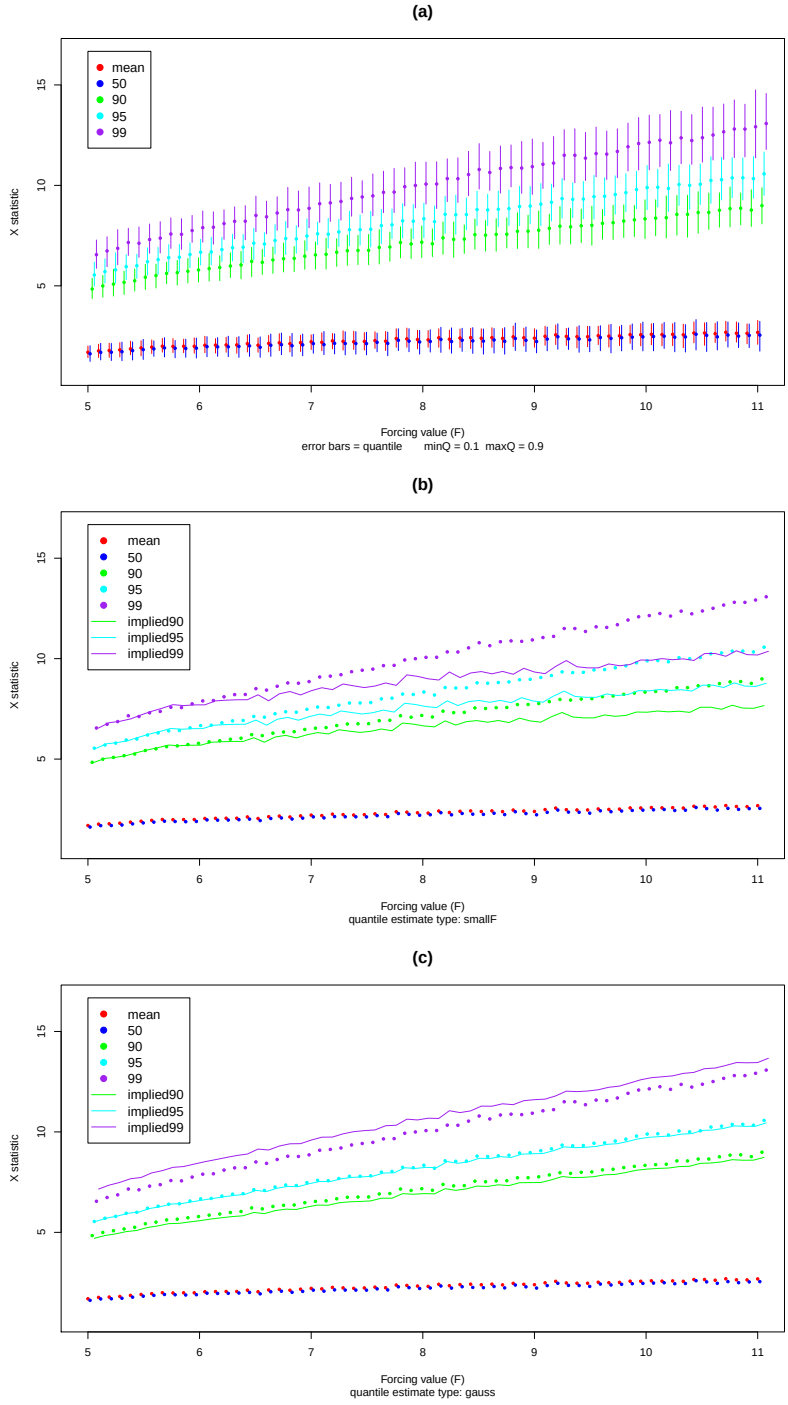


Figure 3.2: Lorenz System I, K=36. F is variable and shown on the x-axis in each plot. In Figure (a) the y-axis shows the mean of a chosen statistic of X_1 (dots) and the bar indicates the 10th and 90th quantiles of the statistic (these are estimated over 128 blocks of data). Each block contains 365 observations every 0.1. Figure (b) shows just the mean dots compared with a stable quantile/mean ratio estimated when $F = 5$. Figure (c) again shows mean dots - against the quantiles from a Gaussian distribution with same mean (red dot) and standard deviation as the observed X_1 variables.

3.4 System II behaviour: Parameterisation 80001

The following defines the sample of system values used in this section.

Sample S3.1 Sample from system 80001 used in section 3.4

System: Lorenz System II

System: 80001

Number of observations: $N_{obs} = 2^{14}$

Observations of each X_k variable and each $Y_{j,k}$ variable are taken in time increments of 0.1. Values of IEF_k are calculated from these values using equation 3.3 and the chosen parameters.

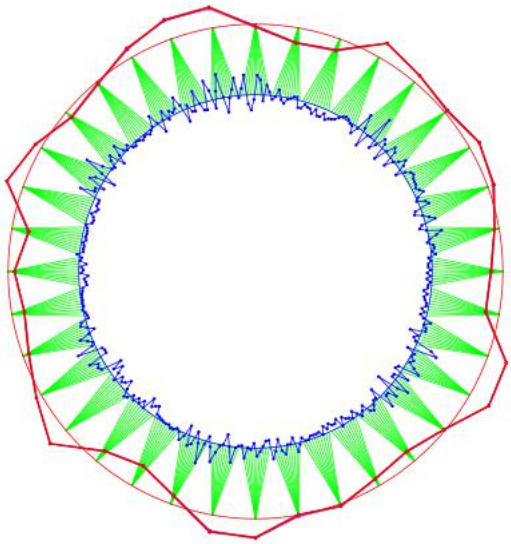


Figure 3.3: Illustration of Lorenz 96 system II, parameterisation 80001, where K=36 and J = 10. X values shown in red and Y values shown in blue. The J-block of the 10 Y values that relate to each X value are shown at the foot of the green lines emanating from that X value.

Illustrative sample values Using system 80001, for illustration, figure 3.4 shows a sample of $Y_{2,1}$ compared with X_1 up to time 20 with observations every 0.1. Note that the Y values oscillate more quickly (there are 57 local maxima in the Y variables

over this period compared to 22 in the X variables) - for this reason the Y variables are sometimes referred to as the 'fast' variables and X is described as 'slow'. The Y values are approximately $\frac{1}{10}$th the size of the X variables.

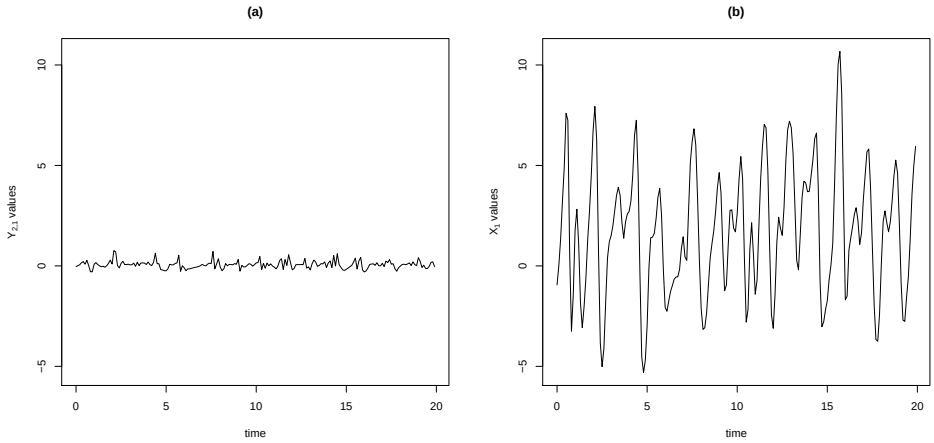


Figure 3.4: Illustrative time series plots from the Lorenz System II, parameterisation 80001. $Y_{2,1}$ values are shown in the graphic (a), X_1 in graphic (b). The y-axis scale of graphic (a) is chosen to be equal to that of (b) to highlight the difference between the Y and X variables.

Probability density for X variable Figure 3.5 shows a kernel smoothed probability density (in black) the X_{20} variable from sample S3.1. The kernel bandwidth is 0.4584. The grey lines are the density plots for the other X variables. By symmetry each X variable has the same asymptotic distribution; despite the density being created over a large sample, however, there remains a degree of spread between them in this case.

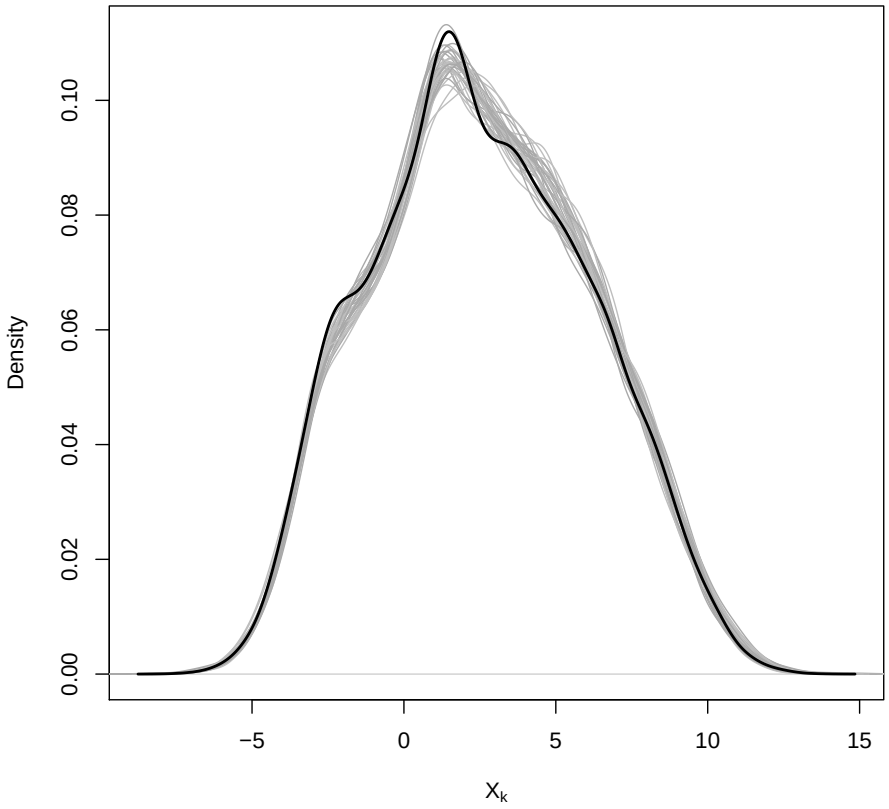


Figure 3.5: Probability density plot of X_{20} from System 80001 (black line). Density plots for X_k $k \neq 20$ are also shown in grey. Density produced using Gaussian kernel with bandwidth 0.4584 over 2^{14} observations in time increments of 0.1.

Probability density for Y variables Figure 3.6 shows ten kernel smoothed density plots of the $Y_{j,k}$ variables in sample S3.1, one for each j position. The kernel bandwidth is derived using Silverman's 'rule of thumb' [229] and takes values between 0.02086 and 0.02232, for such a large sample size this makes little difference. Each plot shows the density of values taken by each of the 36 Y variables that take value j (i.e. $Y_{j,1}, ...Y_{j,36}$). The plot shows that the different Y variables ($Y_{j,k}$) have slightly different distributions. This is because the Y_j in the same J-block are not defined symmetrically, and so need not share the same distribution even asymptotically. In each J-block the first Y variable is influenced by other Y variables that are themselves influenced by the the *previous* X variable. Similarly, the last two Y's are influenced the following X variable (counting round the circle). Those in the centre of the J-block are not so directly influenced by other X variables in this

way. It is clear that, for each fixed j, the density plots appear similar and, indeed by definition, should be asymptotically equal. Hence when the Y variables are summed in the forcing term for the X values they will on average provide the same influence.

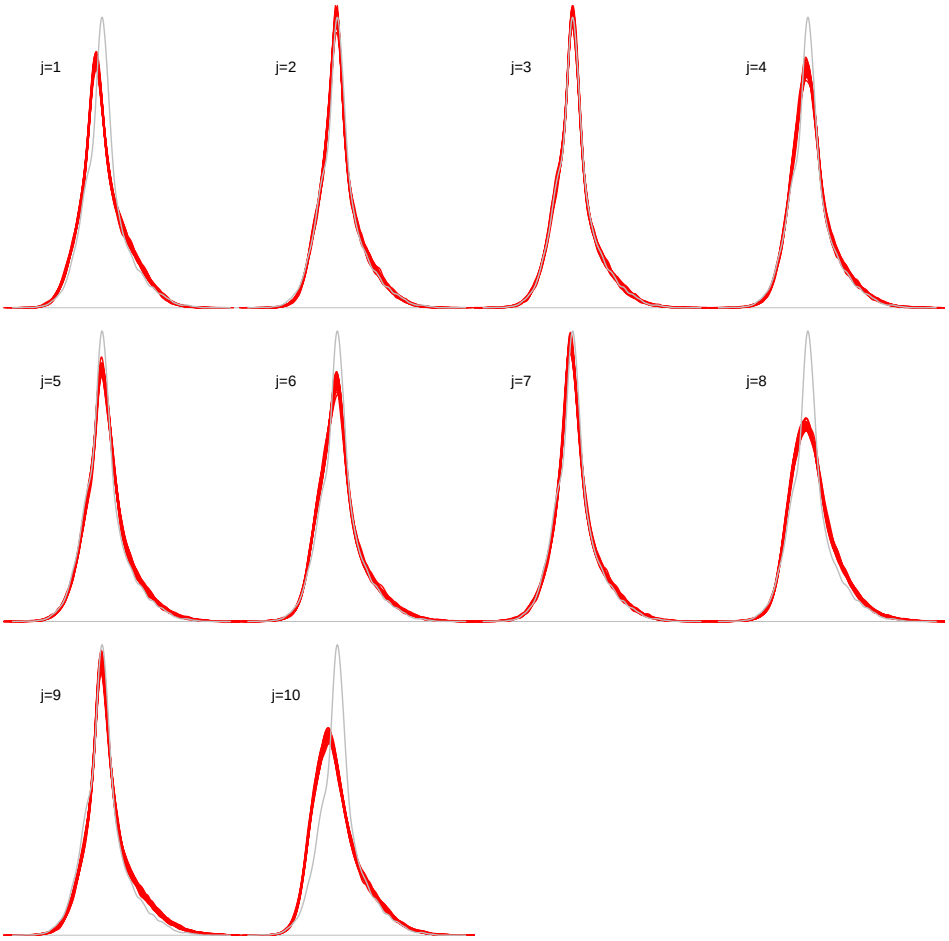


Figure 3.6: Sample probability density for each j value (1,2,...10), for System 80001 over 2^{14} sample values. A each pane shows multiple red lines, i.e. the probability density of the 36 Y variables with value j (i.e. $Y_{j,1}, ...Y_{j,36}$). The sample probability density for $Y_{3,1}$ is shown (in grey) to ease comparison between the plots and to ensue the y-axis scales are the same in each plot.

Correlation between X variables Figure 3.7 compares the correlation coefficient (within sample S3.1) between the X variables. Figure (a) shows the correlation of X_1 with X_k for all values of k, Figure (b) plot is divided into squares (i, j), where i refers to the row number labelled on the left hand side and j denotes to the column labelled at the top. Each square is coloured according to the correlation between X_i and X_j. Shades of blue denote negative correlation and red positive, as shown in the colour chart in figure (c). The diagonal is shown in black to emphasise the correlation of a variable with itself is 100%. The graphics show that the correlation between X_i and X_j decays as $|i - j|$ increases. X_k and X_{k+1} are slightly positively correlated for each k; X_k and X_{k+2} show the strongest correlation for each k with an average coefficient of -0.44 in this sample. It is clear from the colour plot that the correlation decay seen in the top plot is repeated for each X variable as as required by the symmetry of equations 3.2. The red line of figure 3.3 shows series of peaks and troughs around the circle. These correspond to waves which are easily seen in an animation of the system over time. These explain the negative correlation between X_k and $X_{k\pm2}$ since when one variable takes a high value (the peak of the wave) the others are typically take low values (nearer the trough). Since this correlation applies for any k it explains why there are waves of negative and positive correlation though this does not occur exactly in k-steps of 2 (this was not explored further). The key conclusion is that (for any k, j) X_k and X_j are not independent, but the strength of dependency is weak, especially when $|k-j| > 2$. This will be exploited later when certain parameters are derived using all X values simultaneously as though they were independent.

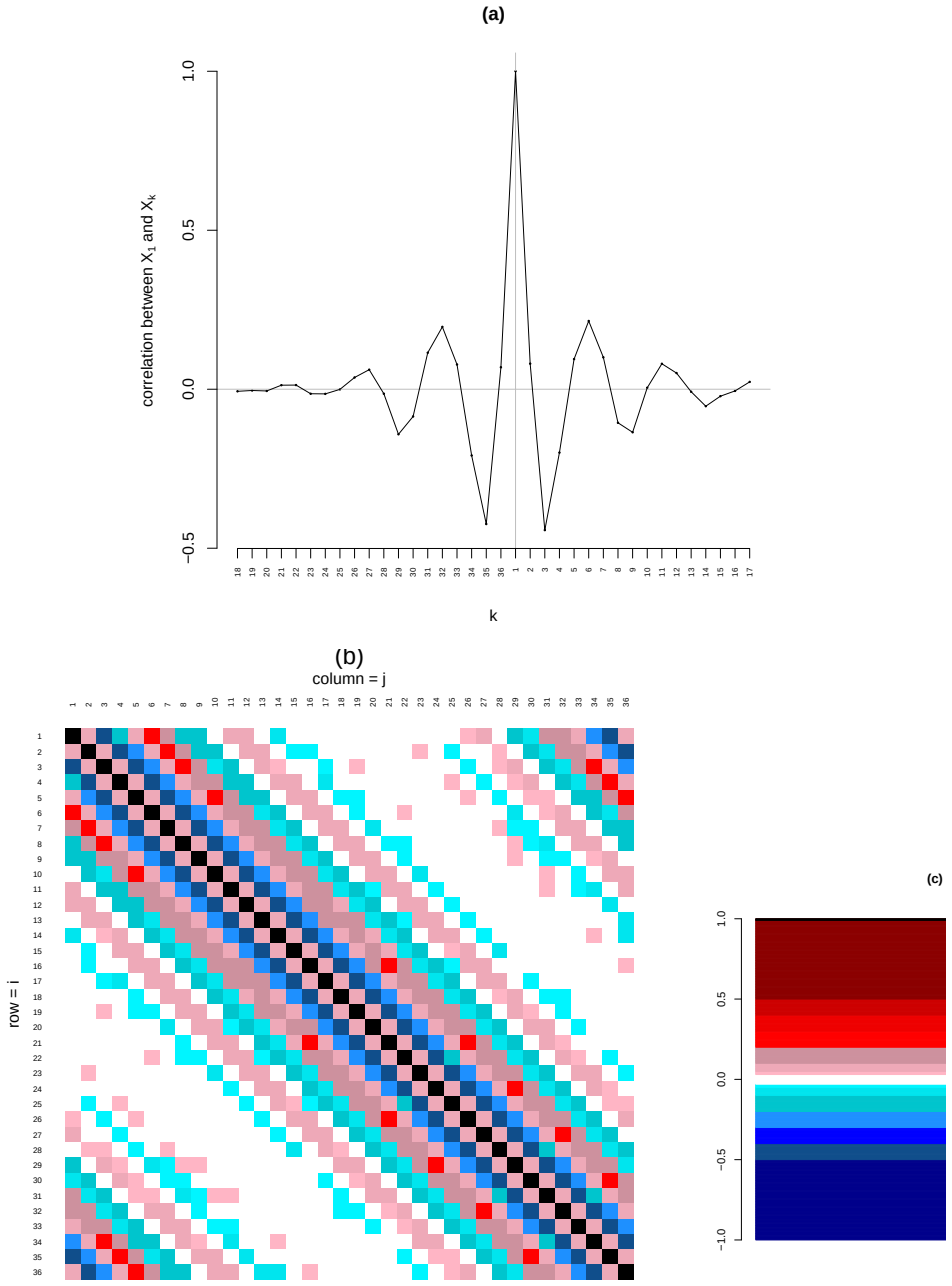


Figure 3.7: Figure (a) shows correlation coefficient value (y-axis) against k for X_1 and X_k. Figure (b) shows correlations between X_{k_1} and X_{k_2} for all pairs of variables indexed by i=k_1 and j=k_2. The strength of correlation is indicated by the colour. The colour key (c) shows that blue shades are used to denote negative correlation and red shades, positive. Black is reserved for 100% correlation.

Instantaneous forcing Figure 3.8 illustrates a short time series of IEF_1 (as defined in equation 3.3) in system 80001; the mean (calculated over values in S3.1) is 9.01. Note that the average IEF is less than the fixed forcing term ($F = 10$). Hence the average contribution from the Y variables is negative. Figure 3.9 illustrates IEF_1 (on the y-axis) compared to X_k, for $k \in \{33, 34, 35, 36, \mathbf{1}, 2, 3, 4, 5\}$. There is clearly a strong relationship between X_1 and IEF_1, with a high R^2 value suggesting that the value of X_1 at a given time explains much of the variance in IEF_1. In the other cases the R^2 value is low suggesting that the influence of X_k on IEF_1 where $k \neq 1$ is not significant. Figure 3.10 illustrates how the value of the IEF_k is related quite strongly to the value time 0.1 previously. The strength of R^2 between times t and s where $|s - t| > 0.1$ diminishes quickly (not shown). Figure 3.11 shows the correlation between pairs of IEF_k variables, blue shades for negative and red for positive. As with figure 3.7 there are decaying waves of correlation between IEF_i and IEF_j as $|i - j|$ increases.

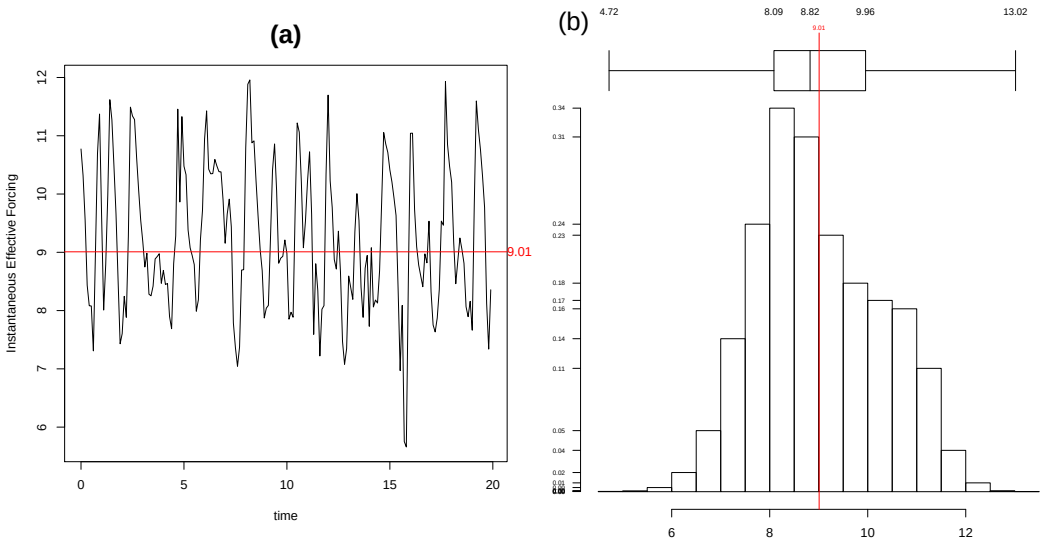


Figure 3.8: Figure (a) short time series of Instantaneous Effective Forcing (IEF_1) with time-mean value shown as red line. Figure (b) histogram of IEF_1 values with box plot above (calculated over 2^{14} sample values).

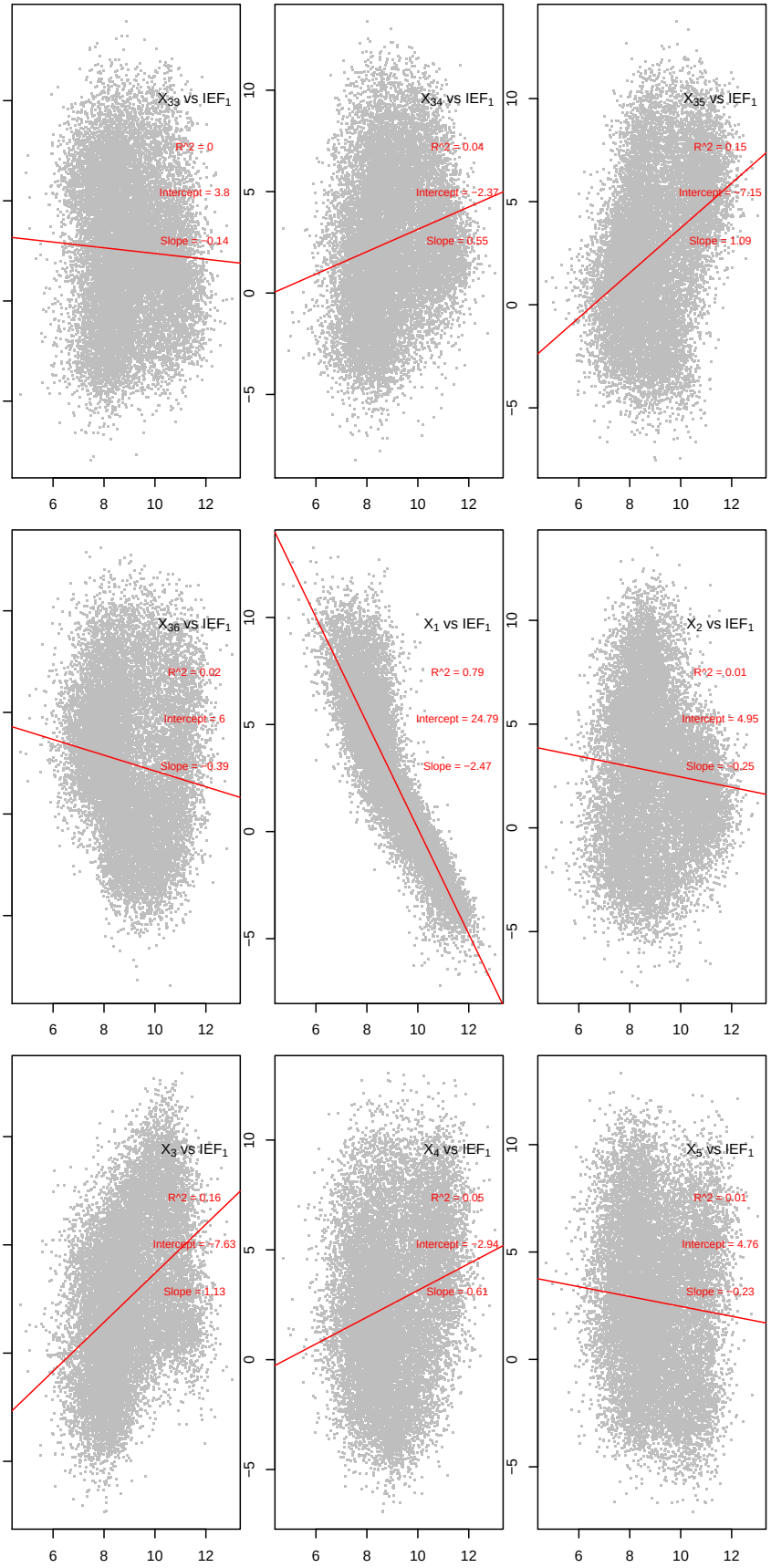


Figure 3.9: X_k values for $k \in \{33, 34, 35, 36, 1, 2, 3, 4, 5\}$ (y-axis) versus IEF_1 (x-axis). The middle graphic shows that X_1 is strongly related to IEF_1 with an R^2 value of 0.79. The relationship between X_k and IEF_1 for $k \neq 1$ is much weaker with low R^2 values in all cases.

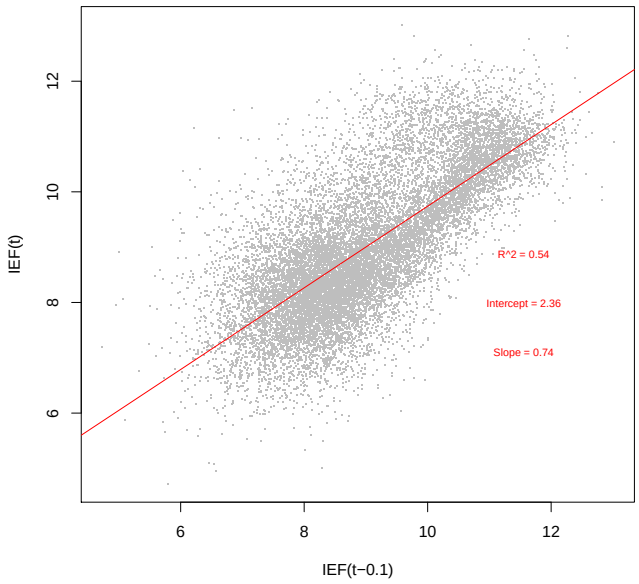


Figure 3.10: $IEF(t)$ versus $IEF(t-0.1)$ showing that the value of the Instantaneous forcing at time t is conditionally related to its value at time $t-1$.

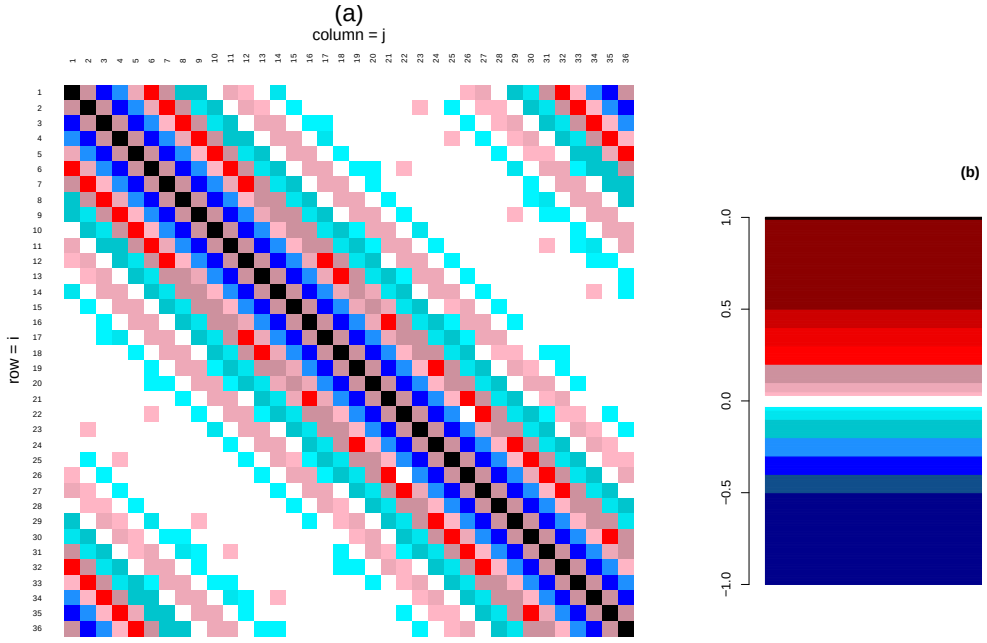


Figure 3.11: Figure (a) correlation colour plot for IEF_{k_1} and IEF_{k_2} for all pairs of variables i=k_1, j=k_2. The strength of correlation is indicated by the colour map shown in figure (b), blue shades are reserved for negative and red for positive. White is used for correlation in the range (-3%, 3%) and black for 100%.

An autoregressive parameterisation for IEF An autoregressive process ($\hat{I}$) is presented in the following as a parameterisation for the IEF. Such a process takes the form:

$$\hat{I}_t = C + \sum_{i=1}^{p} \phi_i \hat{I}_{t-i} + \xi_t \tag{3.4}$$

Where, C is a constant, ϕ_i determines the contribution of past values of $\hat{I}$ and ξ_t is Additive Observational Noise at each time t with zero mean. ξ_t is assumed to be normally distributed with standard deviation σ.

Given a sample of the process the parameters can be estimated using the Yule Walker method [32]. Taking expectations if follows that $C = E(\hat{I})(1 - \sum \phi_i)$. The Yule Walker approach (after first deducting the mean so that the resultant process has zero mean) requires us to multiply the equation defining $\hat{I}_t$ by $\hat{I}_{t-j}$ and then to take expectations so that these cross terms become covariance terms. These operations result in a linear equation in ϕ_i which can be solved. Finally the variance of the noise term can be determined by multiplying the equation defining $\hat{I}_t$ by itself and taking expectations. This gives:

$$\sigma^2 = Var(\hat{I}_t) - \sum_{i=1}^{p} \phi_i E(\hat{I}_t \hat{I}_{t-i}) \tag{3.5}$$

The first step is to determine a suitable value for p. Figure 3.12 shows that the partial autocorrelation function achieves value zero reliably after around $p = 12$. Figure 3.13 shows a sequence of three plots: (a) shows the Akaike Information Criterion (AIC [3]) that is derived from an $AR(p)$ fitting process as p increases (values for IEF_1 are shown in black and the other 35 variables shown in grey). These are equal in distribution so the variation in the grey lines represents sampling error and is quite high. They all show a similar pattern. Figure 3.13(b) shows the p values at which the AIC is minimised for each of the 36 variables - they are all greater than or equal to 12 but there is significant scatter. Finally figure (c) shows (on a log scale) that considerable improvement in AIC is achieved even when $p = 4$.

[3] AIC $:= 2p - 2\ln(L)$, where L is the maximum value of the likelihood function for the model.

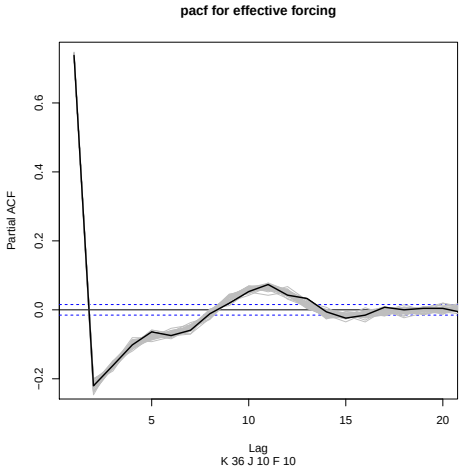


Figure 3.12: Partial autocorrelation function for IEF_1

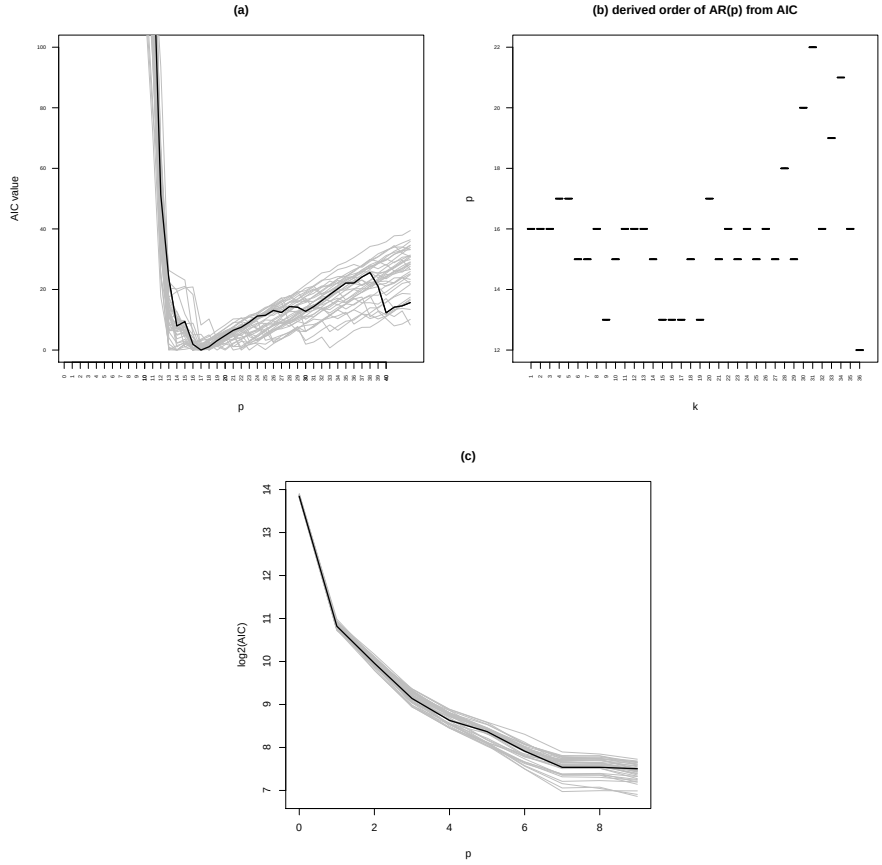


Figure 3.13: Akaike Information Criterion (AIC) graphics for IEF. Figure (a) shows the AIC value for different $AR(p)$ processes where p is shown on the x-axis - the black line uses the IEF_1 variable to derive the fitted parameters, the grey lines illustrate the results for the remaining k variables. Figure (b) shows the value of p (y-axis) that gives the minimum AIC value for each k. Figure (c) shows $\log_2(\text{AIC})$ on the y-axis against p to illustrate how the value is significantly reduced by the time $p = 4$.

Parameters of $AR(4)$ model $\hat{I}$ Based on the previous analysis $AR(4)$ processes $\hat{I}_k$ are fitted to the IEF_k data in sample S3.1. The derived values for $\hat{I}_1$ are: $\phi_1 = 0.8483$, $\phi_2 = -0.0825$, $\phi_3 = -0.0736$, $\phi_4 = -0.1018$, $C = 3.689$ and $\sigma^2 = 0.6710$. Figure 3.14 shows box plots for the coefficients for each of the 36 IEF variables, it is clear that these parameters are quite stable. Note that no correlation has been included between the AR(4) processes for each k whereas figure 3.11 suggests there should be some weak correlation; it would be possible to build such correlation in but this has not been explored further here.

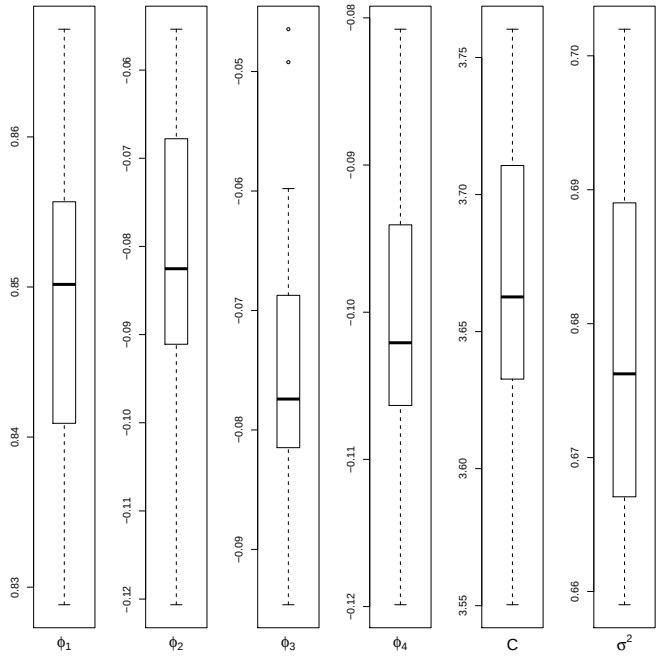


Figure 3.14: Results of fitting an $AR(4)$ processes $\hat{I}_k$ to each IEF_k. Boxplot of coefficients arising when each of the 36 X variables is used separately to fit the model. Since the IEF_k are equal in distribution these parameters should converge in the limit, the boxplots show that there is little scatter in values and hence using the same parameter set for each k, as an approximation to the true parameters, is appropriate.

Gaussian model $\hat{G}$ for IEF A sample of values is drawn from a Gaussian distribution with the same mean and variance as IEF_1. This will be shown for comparison with outputs from the $AR(4)$ process $\hat{I}_1$ and also the true IEF_1.

Comparison of IEF with $AR(4)$ and Gaussian models Figure 3.15 shows histograms of the IEF_1 values (a) from sample S3.1 , AR(4) model $\hat{I}$ (b) and Gaussian model $\hat{G}$ (c) . Both the AR(4) and Gaussian models understate the interquartile

range (slightly) and overstate the full range of values. The Gaussian distribution produces extremes from the model that are larger than the AR(4) (which is in turn larger than the system). Figure 3.16 shows histograms of the *differences* between successive values from the system and models. In this case the AR(4) and Gaussian models tend to overstate the interquartile range. The AR(4) model, however, understates the largest jumps (considerably) and the Gaussian model overstates them considerably.

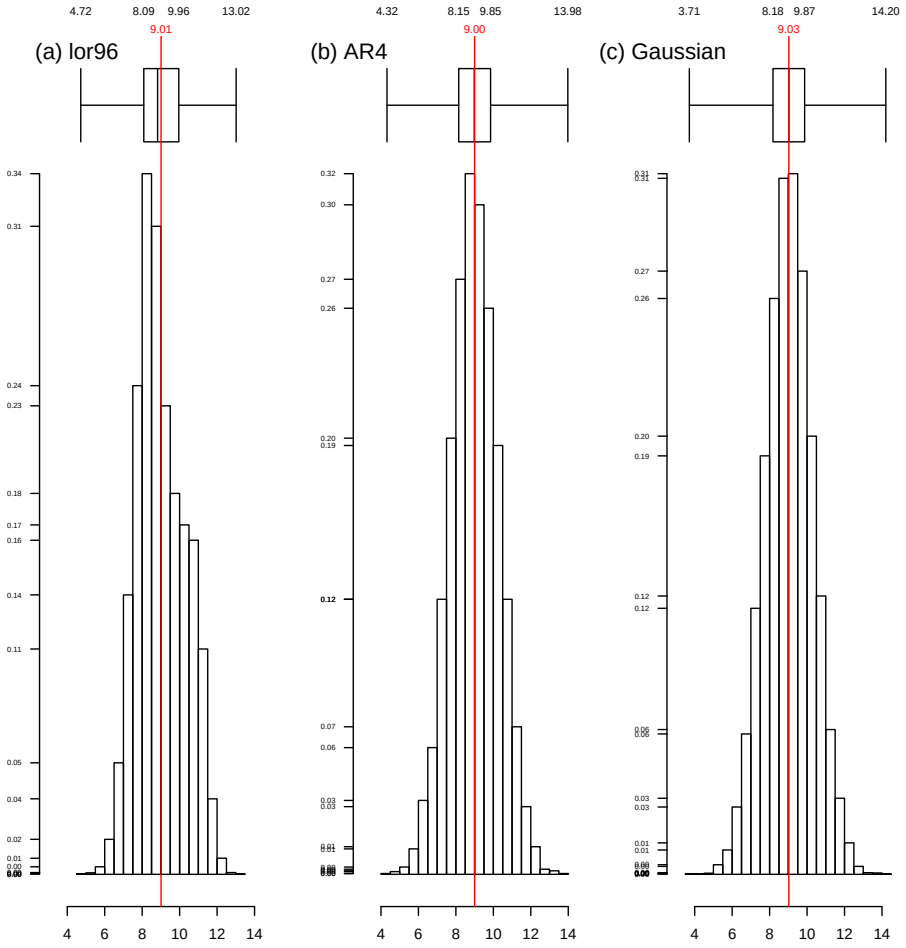


Figure 3.15: Histograms of true IEF values (a) , values derived from an AR(4) model (b) and from a simple Gaussian model with the same mean and variance as the observed IEF_1 (c). Box plots are shown above the histograms illustrating the mean, interquartile range and data extremes.

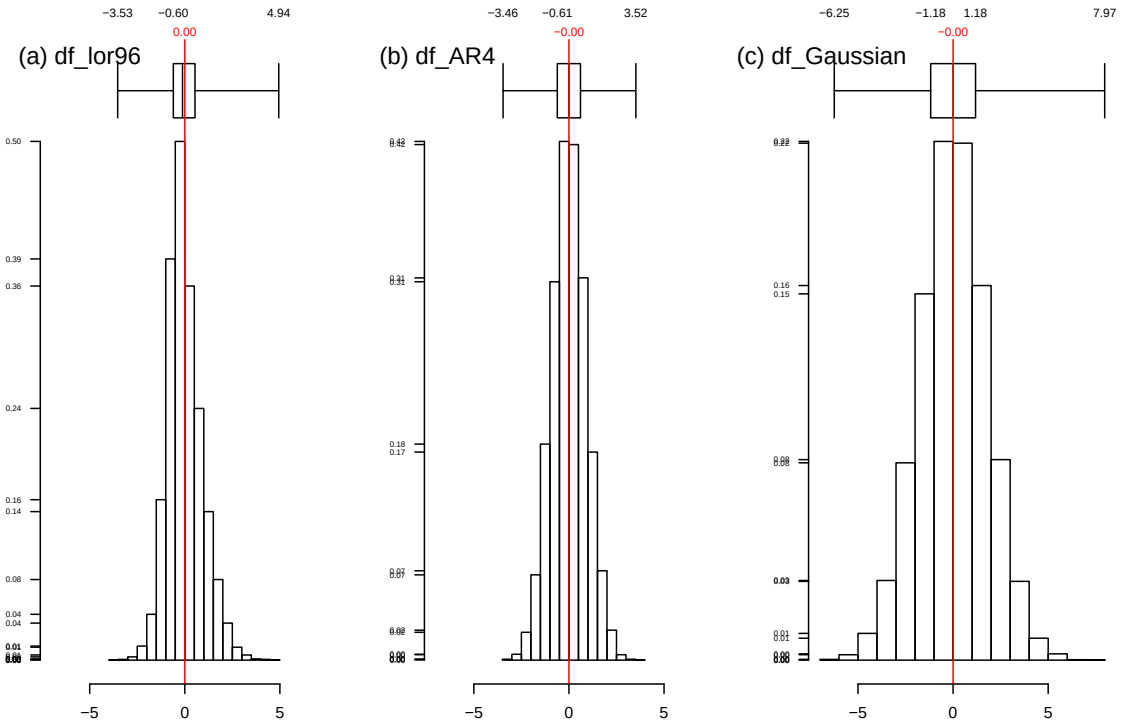


Figure 3.16: Histograms of the differences between successive values of the IEF_1 (a) , AR(4) model (b) and a simple Gaussian model (c).

A parameterisation of IEF_k as a function of X_k Figure 3.17(a) shows a scatter plot from sample S3.1 of IEF_1 compared to the value of X_1 at the same time. A one dimensional functional relationship between the IEF and X is created by averaging (truncated) kernel smoothed lines through the data; for data extremes (the far left and far right of the plot) the relationship is restricted to a constant; the points at which the kernel smoother is replaced by these horizontal lines is chosen manually. This process is first carried out when $k = 1$ and the resulting line is shown as the black line on the left hand graphic. The process is then repeated for each of the values of k to derive the relationship given those data points. As the the variables IEF_k are the same in distribution and similarly X_k the relationship should be identical in the limit. Differences will be caused because the system observations are finite. All 36 relationships are shown in Figure 3.17(b) as grey lines. It is clear that there is only a small amount of difference between the lines. Finally a central line is chosen (shown as red) by averaging the values of the 36 grey lines. It is this average line that is taken to define the relationship between IEF and X for later use. The relationship between X and IEF suggests that using a constant forcing term in any model of the system will cause systematic errors. This is explored

later. The general idea here is not new (see Smith [239] and has been developed further by Wilks [271–273], Kwasniok [129] and Arnold [11]). The development and exploration of this and the other models in this chapter has been carried out to enable consideration of model use in the context of insurance in chapter 4.

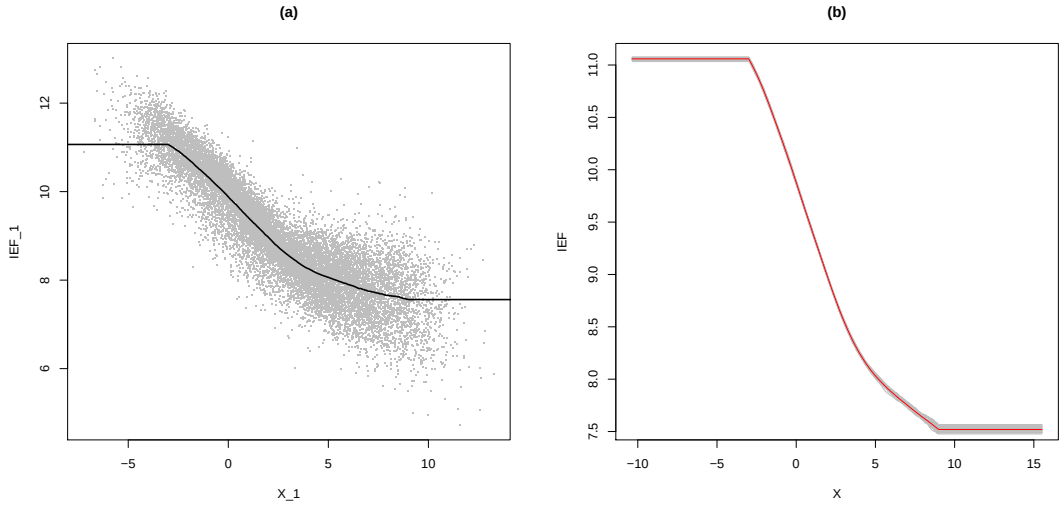


Figure 3.17: Figure (a) IEF_1 versus X_1 scatter plot with smooth line through the data, Figure (b) truncated smooth lines for all K variables (grey) and chosen average relationship $F = f_1(X)$ (red); this illustrates the impact of sampling error is small and the relationship is stable.

Definition f(X) Let the functional relationship defined by the red line in figure 3.17 be denoted $f_1(X)$. The subscript 1 refers to the last digit of the system ID 80001 to which the relationship applies. The relationship f_s for other systems $8000s$ is found similarly.

3.5 System II behaviour: 80002 - 80006

This section briefly discusses the differences in behaviour between the different parameterisations defined in section 3.2. In Chapter 4 it will be important to have systems that behave in different ways to illustrate various features. Specifically it will be interesting to consider systems that are more or less predictable. For this reason four forcing parameters are considered ($F = 9.1$, $F = 10$, $F = 11$ and $F = 20$) and two different coupling parameters are considered ($h_x = 1$ and $h_x = 0.1$). Several combinations of these are used, giving six systems in all, as previously defined.

Parameterisation 80004 was chosen to be a more predictable version of parameterisation 80001. The reason predictability was expected to increase is that the coupling parameter is lower in 80004, hence the influence of the fast Y variables will be much less and the effective forcing much closer on average to the fixed forcing parameter. Therefore it was initially thought that the models would do better in this system; this is not correct as shown in section 3.12 below. System 80002 has a much higher forcing term and is a-priori expected to be less predictable. System 80003 was chosen as a pair for this system that was expected to be more predictable for the same reasons just described and, for the same reasons, this turns out to be false. Given the unexpected results for predictability two further systems were introduced: System 80005 has a lower forcing term but also lower coupling than 80001, in fact these parameters are chosen so that the IEF, defined in equation 3.3, is approximately equal in both systems. System 80006 is chosen so that the IEF is approximately 10.

The following results are based on samples defined as follows:

Samples S3.2 - S3.6 Sample from parameterisations 80002 - 80006 used in section 3.5

Sample names: S3.x relates to system 8000x

System: Lorenz System II

Systems: 80002 - 80006

Number of observations: $N_{obs} = 2^{14}$

Observations of each X_k variable and each $Y_{j,k}$ variable are taken it times each 0.1 apart. Values of IEF_k are calculated from these values using equation 3.3 and the chosen parameters.

Comparison of systems The following comments summarise differences in the behaviour of the six systems:

- System 80001 is discussed in section 3.4 and is shown as a familiar baseline against which the others can be compared;

- System 80004 has the same forcing as 80001, but lower coupling. The range of values arising is slightly wider at the interquartile range and full range. The median is also slightly higher. The IEF of system 80004 is much narrower and much closer to 10. This is expected because the coupling is much smaller, so the Y variables have much less influence. The median IEF is higher as a consequence. Recall the mean IEF (not shown) of system 80001 is 9.01; for 80004 it is 9.91. This explains why the range of X_1 values is higher for system 80004, because on average it has larger forcing and section 3.3 shows that this leads to a wider range of values.
- Similarly 80003 has a slightly larger range of values than 80002 because the coupling is lower and the IEF is consequently higher on average. Specifically the mean IEF for system 80002 is 18.9 and for 80003 is 19.90
- Systems 80002 and 80003 have a much wider range of values than 80001 - this is due to the larger forcing value $F = 20$ vs $F = 10$ in these systems.
- System 80005 is designed to have very similar average effective forcing as 80001 (in the sample taken it is 9.0061 compared to 9.0065 respectively) however because it has $h_x = 0.1$ the effective forcing has a much lower standard deviation (0.136 compared with 1.27).
- System 80006 is designed to have a slightly larger range than 80001. The lower end of the range of X_k values for system 80006 is on average 15.7% lower than system 80001 (but with a high standard deviation of 13.5%), and the upper end of values is on average 11% higher (with standard deviation 6.8%).

3.6 Model specifications

This section defines the specific models and parameterisations that will be used to produce forecasts for testing later. First consider the following variant to Lorenz System I.

Lorenz System I - model variant A Consider the following variant to Lorenz System I.

$$\frac{dX_k}{dt} = (X_{k+1} - X_{k-2})X_{k-1} - X_k + F_k \tag{3.6}$$

Where F_k is constant over time but varies with k.

Model L1A1 - $AR(4)$ method The following model uses Lorenz system I, variant A with the following 'time step' parameterisation of F_k.

- Let $\{\hat{x}_{k,0}\}$ be a set of initial conditions for the system s.
- For period $\pi_j \in \{\pi_1, ...\pi_{N_{periods}}\}$, where $\pi_j = \{t_{j,0}, ...t_{j,P}\}$
- For $t_{j,r} \in \pi_j$,
  - Define $F_{k,t_{j,r}} = C_s + \sum_{i=1}^{4} \phi_{i,s} F_{k,t_{j,r-i}} + \xi_t$, where $\xi_t \sim N(0, \sigma_A^2)$ (This is an AR(4) process as defined in equation 3.4). Values $F_{k,t_{j,r-i}} := F_{k,t_{j-1,P-i+1}}$ for $i \in \{1, 2, 3, 4\}$ come from the prior period. (Note that the $AR(4)$ parameters C_s and $\phi_{i,s}$ depend on the system s.)
  - Using equation 3.6 evolve the system forward using $F_k = F_{k,t_{j,r}}$ to time $t_{j,r+1}$
- Loop through values of $t_{j,r} \in \pi_j$ redefining the forcing terms each time.

Note that F_k is constant for each k between observation times $t_{j,r-1}$ and $t_{j,r}$ but different for each value of k. As such F_k does not describe a continuous time autoregressive process [33] rather it is derived from a discrete time AR(4) process and held constant between observations.

Model L1A2 - $f_s(X)$ method The following model uses Lorenz system I, variant A with an alternative 'time step' parameterisation of F_k.

- Let $\{\hat{x}_{k,0}\}$ be a set of initial conditions for the system s.
- For period $\pi_j \in \{\pi_1, ...\pi_{N_{periods}}\}$, where $\pi_j = \{t_{j,0}, ...t_{j,P}\}$
- For $t_{j,r} \in \pi_j$,
  - Define $F_{k,t_{j,r}} = f_s(X_k(t_{j,r}))$, for some functional relationship f_s
  - Using equation 3.6 evolve the system forward using $F_k = F_{k,t_{j,r}}$ to time $t_{j,r+1}$
  - Loop through values of $t_{j,r}$ redefining the forcing terms each time.

Note that F_k is constant for each k between observation times $t_{j,r-1}$ and $t_{j,r}$ but different for each value of k.

Description of models Five models have been set up for each of the systems 80001 to 80006, three use Lorenz System I, one uses model L1A1 and the other L1A2. These models are each given a unique ID which is made up of 5 digits which have the structure [xx][y][zz]. The leading two digits have [xx] = [10] in each case and was used to allow exploration of different models in future, [y] relates to the system in increasing order such that y=0 relates to system 80001 and y=5 to 80006. [zz] relates to the model type. In summary:

zz = 08 implies a model where the constant forcing variable is set equal to the forcing in the system;

zz = 09 where the constant forcing variable is set equal to the average Instantaneous Effective Forcing in the system;

zz = 10 when the constant forcing term is set equal to the same quantile of the corresponding system's IEF in each case (defined on page 162);

zz = 11 when the model forcing is time dependent and equal to a function of the value of the X variable chosen as described above; and finally

zz = 12 when the model forcing is an AR(4) process related to prior values of the forcing

The value of [zz] therefore determines a 'class' of model types across the systems and for brevity a shorthand notation is used below where, for example, *8 relates to all models where [zz]=08. From now on, rather than say 'System II, parameterisation 80001' the short description 'system 80001' will be used unless the longer description is clearer. The following specify the models in detail and also state the numbers of observations used in various forecasts.

Table 3.2: Constant forcing parameters for models *8, *9 and *10

Model class	80001	80002	80003	80004	80005	80006
*8	10	20	20	10	9.1	11
*9	9.008	18.957	19.903	9.905	9.005	9.989
*10	8	17.547	19.752	9.794	8.896	8.925

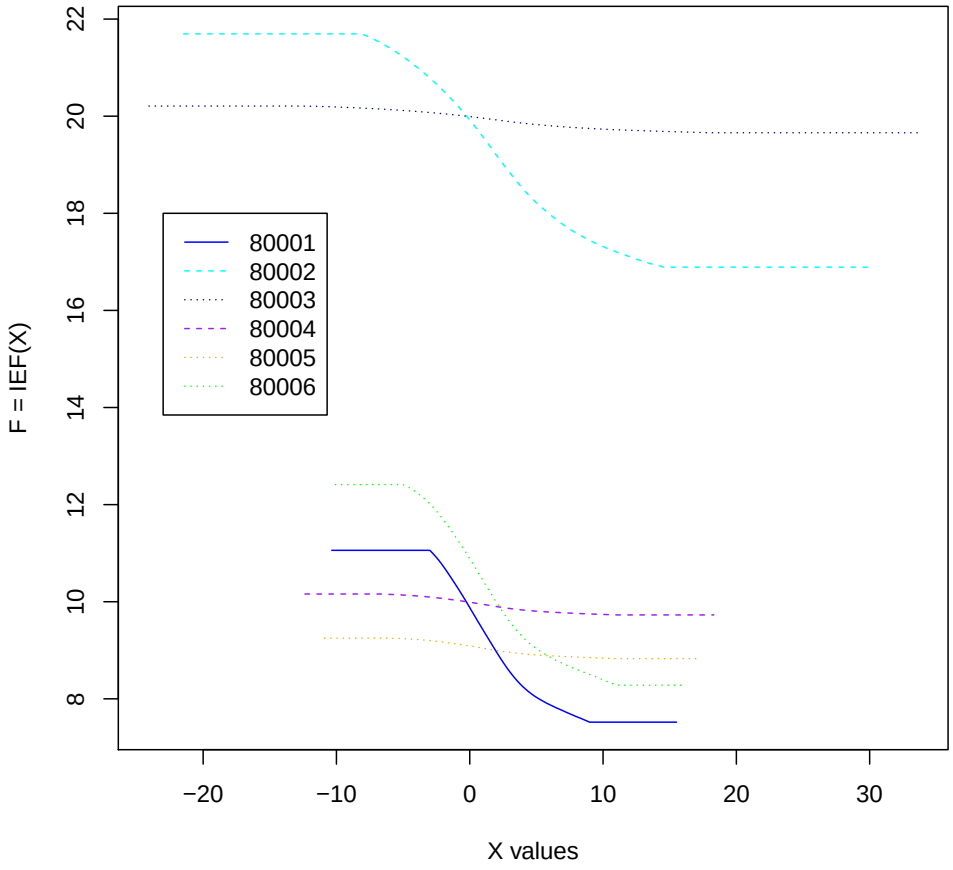


Figure 3.18: Comparison of functional relationship $F = IEF(X)$ for systems 80001, ...80006, used in definition of model class *11

Table 3.3: Parameters of AR(4) processes for IEF in systems 80001-80006

Parameter	80001	80002	80003	80004	80005	80006
ϕ_1	0.8483	0.5633	0.5182	0.8007	0.8334	0.8086
ϕ_2	-0.0825	-0.1236	-0.1145	-0.1019	-0.0985	-0.0828
ϕ_3	-0.0736	-0.0285	-0.0087	-0.0793	-0.0778	-0.0773
ϕ_4	-0.1018	-0.0456	-0.0358	-0.0347	-0.0491	-0.0843
C	3.689	12.025	12.754	4.113	3.530	4.351
σ_A^2	0.6710	2.3526	0.0279	0.0099	0.0084	0.8398

Common parameters Parameters that are common to all models *8-*12

Systems: 80001 - 80006

Observation time step: Increments of 0.1

Period: 24 observation time steps

Skill score: Relative Ignorance

Ensemble Size: $N_{ens} = 2^4$

Training observations: Blending parameters are based on $N_{train} = 2^5$ observations at each lead time for each of the $K = 36$, X_k variables. (see Note A below)

Skill Observations: $N_{periods} = 1184$, values available for each of the 24 lead times.

Climatology: Created from $N_{clim} = 2^{10}$ consecutive observations on the attractor of variable X_1 in each system. Probability forecast created by kernel dressing these observations using equation 2.37 with a kernel width of $\sigma_u = 1.0$.

Note A: The number of training observations is quite low due to run time constraints and this is why the scores for variables other than X_1 are included in the optimisation process. For a given lead time, the skill score is therefore minimised over the aggregate of the $N_{train} = 2^5$ observations and also the $K = 36$ variables. Since these variables are equal in distribution this gives 1152 observations at each lead time. These are not independent due to the slight correlation between X variables.

Model class *8 Fixed forcing equal to system value

Model ID: Summarised in *8 row of table 3.4

Model: Lorenz System I (equation 3.1)

Forcing parameter - F: Summarised for each model and system in *8 row of table 3.2

Model class *9 Fixed forcing equal to average IEF

Model ID: Summarised in *9 row of table 3.4

Model: Lorenz System I (equation 3.1)

Forcing parameter - F: Summarised for each model and system in *9 row of table 3.2

Model class *10 Fixed forcing equal to consistent low quantile of IEF

Model ID: Summarised in *10 row of table 3.4

Model: Lorenz System I (equation 3.1)

Forcing parameter - F: Summarised for each model and system in *10 row of table 3.2

Choice of fixed forcing term: Let Φ_x denote the CDF of the IEF from system 8000x[a]. In model 10010 the constant forcing parameter $F = 8$ is chosen arbitrarily to explore the resulting behaviour. Let $q_1 = \Phi_1^{-1}(8)$ be the quantile of the distribution of IEF in system 80001 that gives a value $IEF = 8$. For a given system 8000x, let the fixed forcing term in model *10 for system x (denoted F_x) be defined as

$$F_x = \Phi_x(q_1) \tag{3.7}$$

In words, the constant forcing term used in models *10 is chosen so that it is the same quantile of the IEF in each system.

[a] $\Phi(x) = P(IEF_k < x)$, where $P(A)$ is the probability of A

Model class *11 Forcing with functional relationship to X variables

---

Model ID: Summarised in table 3.4

Model: Model L1A2 (equation 3.6, variant 2)

Forcing parameter - F: Functional relationship f_s illustrated for each system in figure 3.18

Model class *12 Forcing with AR(4) process

Model ID: Summarised in table 3.4

Model: Model L1A1 (equation 3.6, variant 1)

Forcing parameter - F: AR(P) process - where $P = 4$, Parameters defined for each system in table 3.3

Table 3.4: Summary of Model IDs defined by System ID (column) and treatment of forcing (row)

		System					
Definition	shorthand	80001	80002	80003	80004	80005	80006
$F_{model} = F_{system}$	$*8$	10008	10108	10208	10308	10408	10508
$F_{model} = E(IEF_{system})$	$*9$	10009	10109	10209	10309	10409	10509
$F_{model} = quantile(IEF_{system})$	$*10$	10010	10110	10210	10310	10409	10509
$F_{model} = f(IEF_{system})$	$*11$	10011	10111	10211	10311	10409	10509
$F_{model} = AR(4)$	$*12$	10012	10112	10212	10312	10409	10509

3.7 Model behaviour for System 80001

This section explores and compares the large scale statistical behaviour of the X_k variables from the models of system 80001. No assessment is made at this stage of whether the model values at a given point in time are close to those of the system. Insurers are interested in probabilities of extreme values and a model which produces *behaviour* that is similar to the system can still be useful even if forecast skill is low. The following samples were created:

Samples S3.7.x Sample from models of system 80001
Sample names: S3.7.x relates to model 1000x
Models: 10008 - 10012
Number of observations: $N_{obs} = 10241$
Observations of each X_k variable are taken in time increments of 0.1. The model is run from a single initialisation to produce all observations (i.e. these are not split into periods and reinitialised).

Comparison of system and models under constant forcing Figure 3.19 illustrates the results of using different constant forcing values. The plot shows four groups of four box plots. Within each group the four boxplots reading from left to right show (1) system 80001, and then three models: (2) Model 10008 (3) Model 10009, and (4) Model 10010. The four groups show different percentiles of observed time series illustrated by boxplots of the value arising from each of the 36 X variables. The percentiles are: Top left: 50%, Top right 60%, Bottom left 90% and Bottom right 99.5%. The y-axis is truncated to be on the same scale for each of the models. The figure illustrates that with any constant forcing parameter the model is not able to match all quantiles of the system. For example at the 50%ile a forcing value of $F = 10$ produces median model outputs that are close to those of the system. At the 60%ile the average forcing value of $F = 9.008$ has a median that is closest to the system and at the higher quantiles (90% and 99.5%) the lower value of $F = 8$ has a closer median. This suggests that different values of forcing parameter F could be chosen depending on which part of the distribution a user is interested in.

Comparison of system and models under time dependent forcing Figure 3.20 shows similar plots for parameterisations 10011 and 10012 where the value of F varies with time. It is clear that these models do much better in capturing the observed quantiles from the system. Model 10012 (AR(4) process for F) doesn't perform quite as well at the higher quantiles, however. Model 10011, where F varies using the functional relationship with X, does very well at all quantiles.

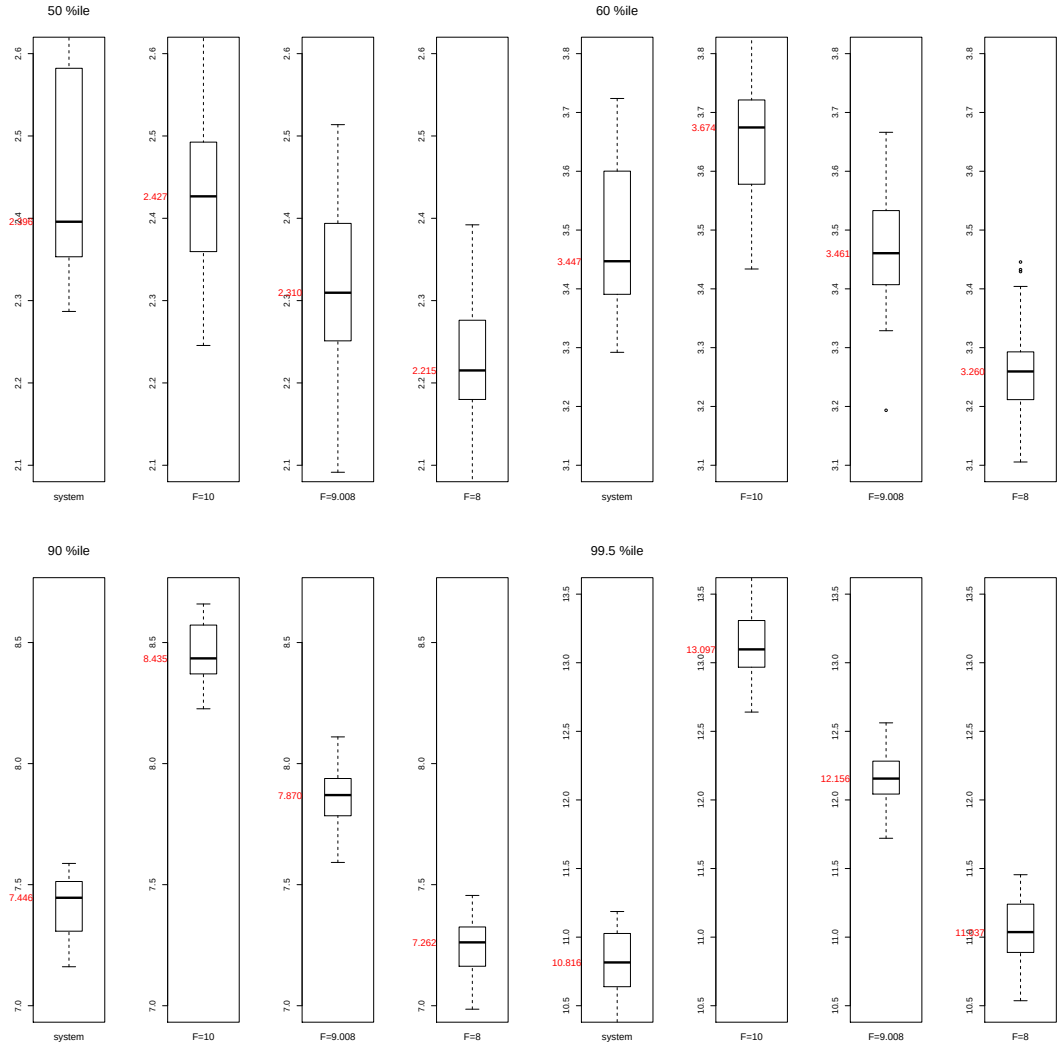


Figure 3.19: Boxplots of time series quantiles of one system and three models of that system (boxes show interquartile range). Quantiles of X_k are shown in four blocks of four graphics for values 50%, 60%, 90% and 99.5%. Each box plot shows the range of quantile values arising for every value of k. Each block of 4 graphics shows four cases,left to right: (1) system 80001 (2) model 10008 (3) Model 10009 (4) Model 10010.

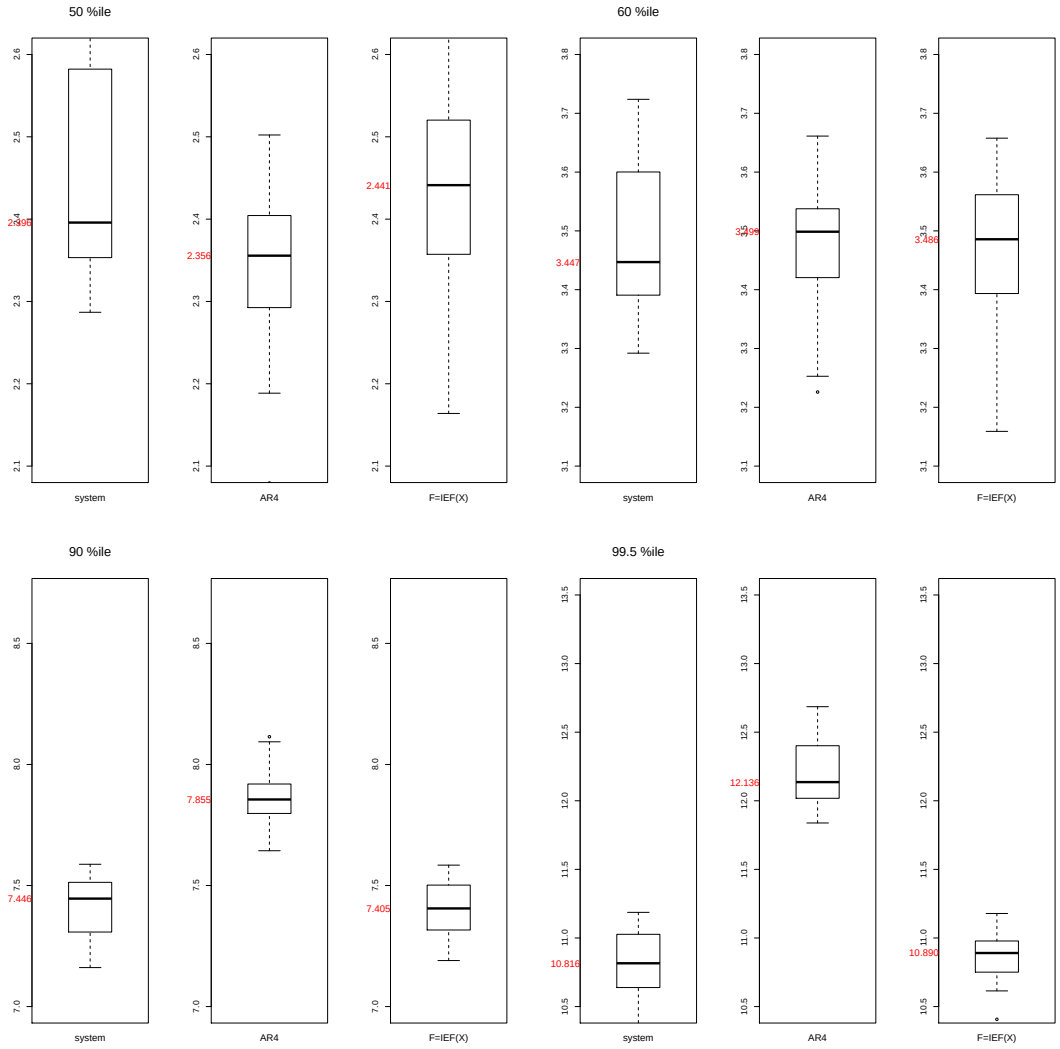


Figure 3.20: Boxplots of time series quantiles. Similar to figure 3.19. Comparison of system quantiles with varying-F models (10012 AR(4) and 10011 $F = IEF(X)$). Quantiles of X_k are shown for values 50%, 60%, 90% and 99.5%

3.8 Climatology Blended Forecasts: 10008

This section describes how climatology blended forecasts are created at each lead time for system 80001 and model 10008. Chapter 2 showed that under many circumstances the Ignorance skill score performs better than others. Therefore the Ignorance will be used as the sole skill score in this and the next chapter. The following uses a climatology blending as defined in Chapter 2, equation 2.62 where $y = X_1$ is being forecast and the initial forecast $p(y)$ is created by kernel dressing the ensemble $\hat{X}_1^1, ...\hat{X}_1^{N_{ens}}$ using the method described in equation 2.37.

Illustration of ensemble forecasts in model 10008 Figure 3.21 illustrates forecast 10008 over four different periods. Time is shown on the x-axis and labelled as $n.q$ where n is the period number and q is the proportion through the period at which the observation is made. As described above there are 24 observations per period and so the values of q are $\frac{1}{24}, \frac{2}{24}, ...\frac{24}{24}$. Period 0 (for example) illustrates a situation where the error in the observed initial condition is sufficiently large that the ensemble initial conditions fail to include the true system value. Around a third of the way through Period 0 the ensemble range is wide and in the last quarter of the period it is almost as wide as the entire range of possible X values. Nevertheless in period 0 the ensemble range does include the system values throughout; this is not the case for period 1 where the system values fall outside of ensemble range for an interval. Period 2 illustrates that the model ensemble can stay close to the system for almost half the period. Finally period 3 illustrates relatively close agreement between the model and system through to time 3.7. Taken as a group the graphics illustrate the wide range of behaviours an ensemble can show relative to the system. Figure 3.22 illustrates the five different models described above for the period 0.

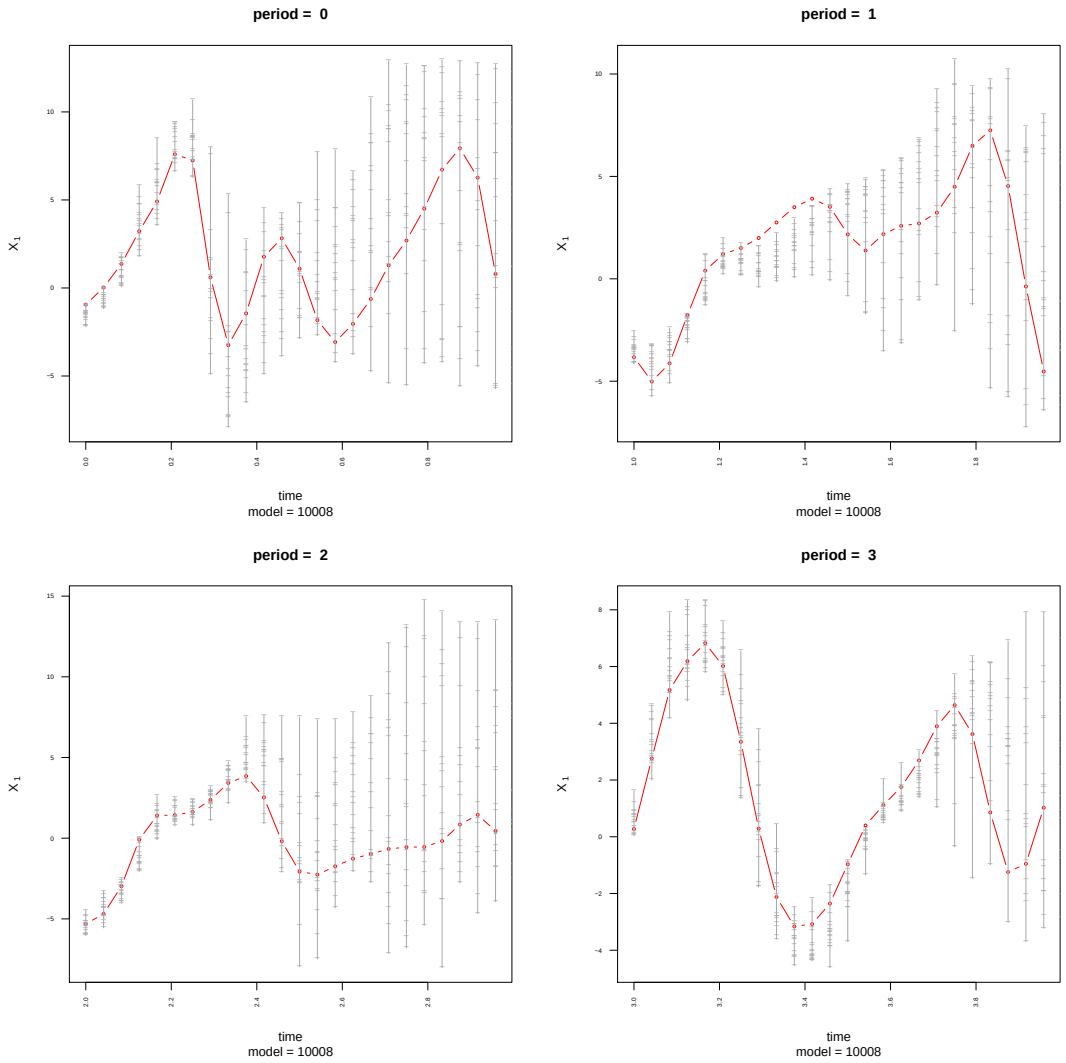


Figure 3.21: Comparison of ensemble values (grey) from model ID=10008, with system values (red) for various illustrative periods, each containing 24 timesteps of length 0.1

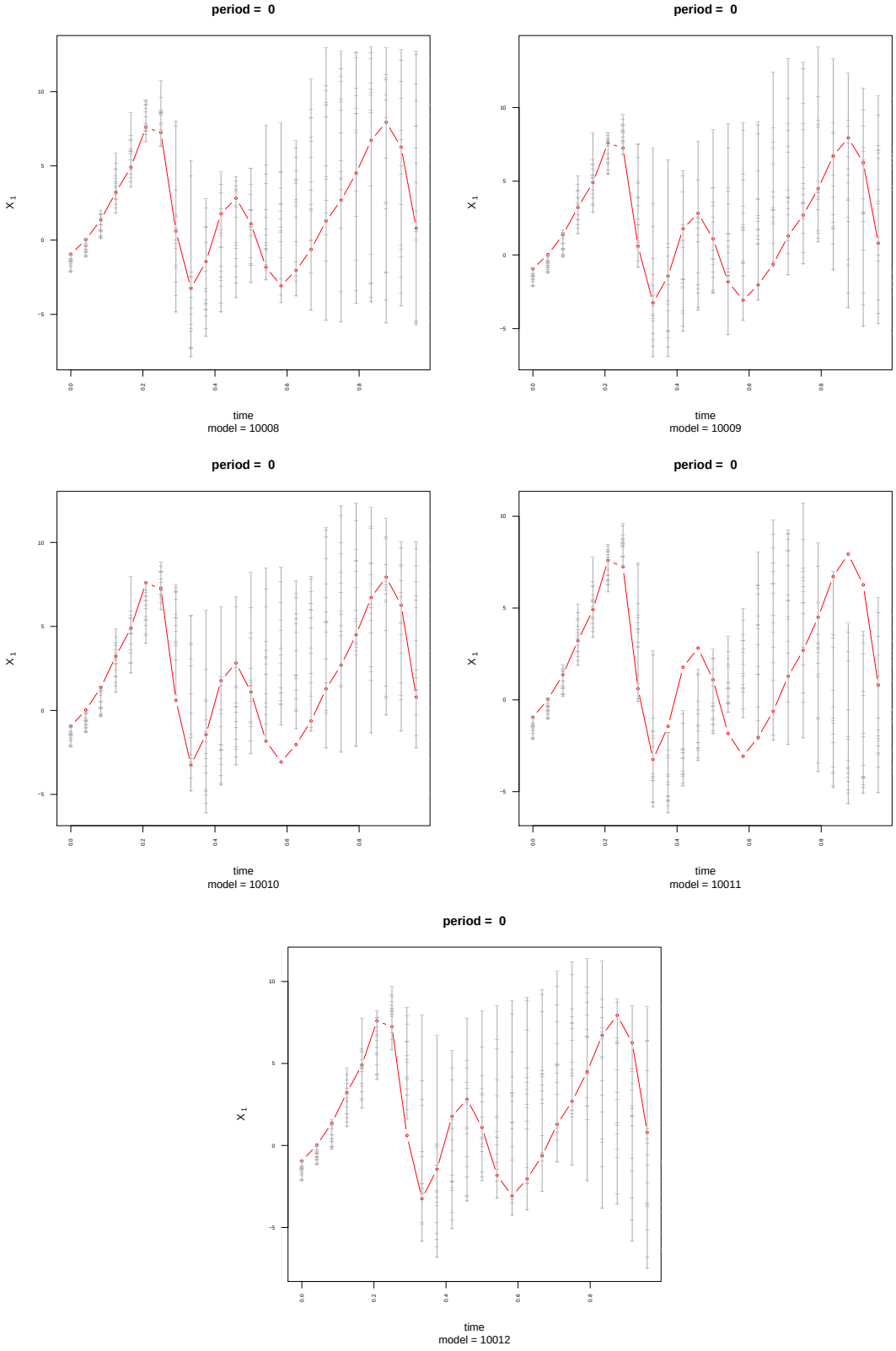


Figure 3.22: Comparison of system (red) with ensemble values (grey) from models 10008, 10009, 10010, 10011 and 10012

Determination of climatology blending parameters for forecast 10008 The determination of the blending parameters is carried out in three stages described below. The parameters are chosen to be score optimal but restricted to a discrete set of trial values[4]. As such the true score optimal values may not be found but this is not thought to be a constraint given the density of the grid used. When a parameter is chosen to give the lowest Ignorance score from a one dimensional set of trial values it will be referred to as the **'discrete-best'** parameter; when two parameters are chosen simultaneously from a two dimensional grid of discrete values they will be referred to as **'grid-best'** parameters.

- **Fixed σ, varying α:** Initially choose a fixed value for the kernel dressing bandwidth σ and calculate the average relative ignorance for different values of α in equation 2.62. Use this to calculate the discrete-best values $\alpha_0(t)$, which produces blended forecasts $r(X)$ with the lowest Ignorance at each lead time t.
- **Calculating best σ values given $\alpha_0(t)$:** At each lead time t, use the value $\alpha_0(t)$ defined above in the blending equation 2.62 and try different values for the kernel dressing bandwidth to find discrete-best values $\sigma_0(t)$ that give the lowest Ignorance forecast r.
- **Jointly best values:** Finally, consider a grid of blending parameters centred on $\alpha_0(t)$ and $\sigma_0(t)$ to test whether a better score can be found using different parameters. Define the **'blending parameters'** to be the grid-best values α and σ that therefore have the lowest Ignorance on the grid.

Fixed σ, varying α As an initial exploration the kernel width σ was fixed at 1.0 and α was allowed to vary between zero and one. This was done seperately for each of the 24 observation times during the period. Figure 3.23 shows an illustrative plot of the resulting average Ignorance score against α (at observation time 0.833). Note the discrete-best value of α where the average Ignorance score is lowest. Given the smoothness of the relationship between α and the average Ignorance score this is not believed to be a constraint.

The average Ignorance score was created by evaluating 2^5 periods and also each

[4]Due to run time constraints optimisation routines were not used in this chapter or the next

of the (36) X variables in the Lorenz 96 system - giving an average over 1152 scores for each time value. Due to the slight correlation between the X variables this may tend to include runs of high or low scores. This feature is not considered further.

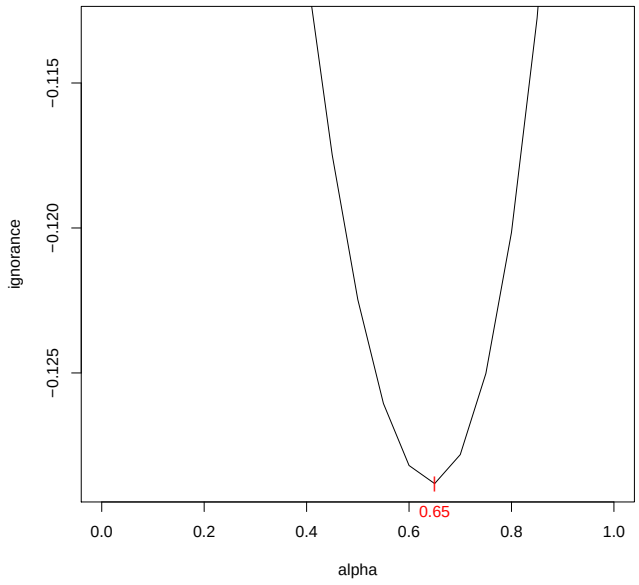


Figure 3.23: Illustration of process to determining the discrete-best blending α parameter for a fixed kernel bandwidth ($\sigma = 1.0$ in this case). This process was repeated for each of the observation times during the period. The graphic illustrates the relationship between the value of α and the average Ignorance score when the observation time is 0.833 through the period. The graph is piecewise linear illustrating the discrete values at which α was tested. The value of α that minimises the average Ignorance score is 0.65 in this case.

The discrete-best α variables are calculated for each observation time through the period and these are shown in figure 3.24. Figure 3.21 illustrates that typically the model ensemble values stay close to the system trajectory quite well initially but then they diverge in the second part of the period such that they nearly span the whole range of potential X values. This suggests that in the second half of the period the ensemble may be little better than a climatology forecast. Figure 3.24 is consistent with this interpretation. For observation times less than 0.5 the value of α that optimises the Ignorance score is close to 1 meaning that almost full weight is given to the kernel dressed ensemble. After this time the value of α falls to a minimum value of 0.55 at the end of the period - i.e. more weight is given to climatology at this time. Since different values of α are derived at each observation

time during the period the resulting values will be denoted below as $\alpha_0(t)$.

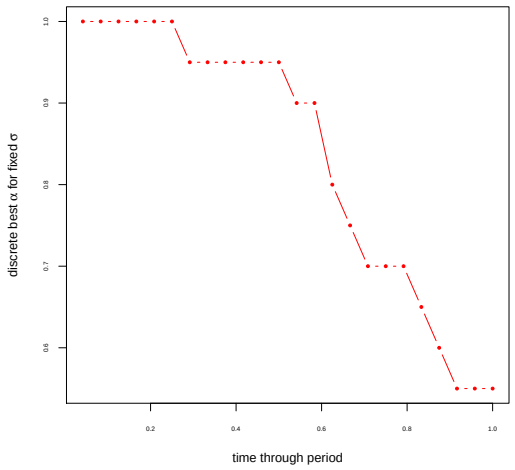


Figure 3.24: Discrete-best α_0 values for each observation time through the period

Calculating best σ values given $\alpha_0(t)$ In the previous sub-section σ was held fixed at a value of 1.0. This sub-section explores whether the average Ignorance score can be improved further by allowing σ to vary and using the values $\alpha_0(t)$ that were calculated above. Two examples of this are shown in figure 3.25 the left hand plot shows the average Ignorance when $t = \frac{3}{24}$ and the right hand plot looks at a later time in the period when $t = \frac{19}{24}$. The figure shows that better average scores can indeed be found when σ is allowed to vary. In the early part of the period a value for the kernel bandwidth (σ) of less than 1 puts more weight on the ensemble values leading to a narrower distribution. Conversely, later in the period a better average score is achieved by letting the kernel bandwidth be greater than 1; this leads to a wider distribution which avoids poor Ignorance scores when the ensembles have failed to shadow the system. The discrete-best σ values are therefore determined given each value of $\alpha_0(t)$ these are shown in figure 3.26. Note that for early parts of the period the values are less than 1 and these are times when the α value is high (i.e. significant weight given to the ensemble). Figure 3.21 makes this clear - initially the forecasted values are close to the observed value and so a small bandwidth (σ) for the kernel smoother gives a lot of weight to the actual ensemble values - and as the observation is close to these the score will be high.

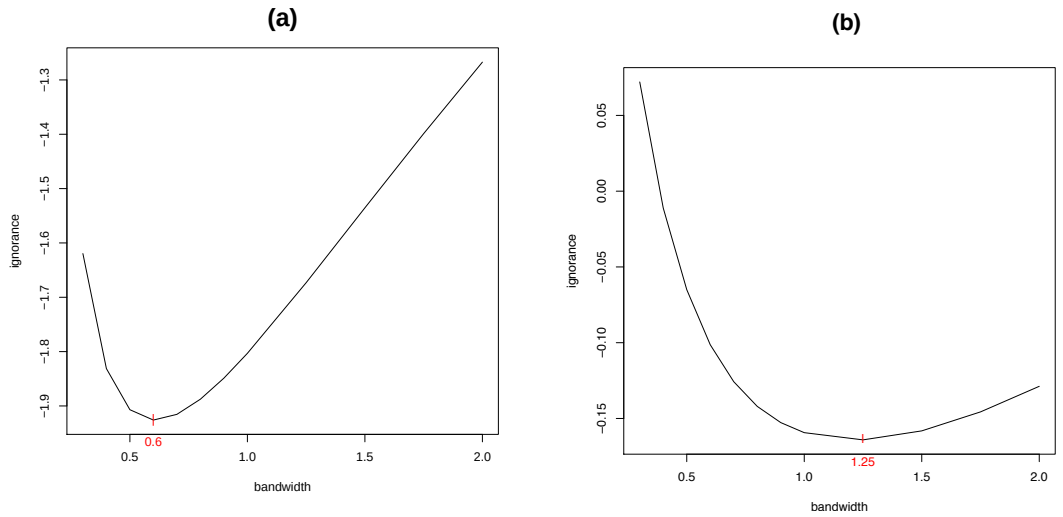


Figure 3.25: Discrete-best σ against observation time for locally discrete best $\alpha_0(t)$. The left hand graphic shows the value when the time is $\frac{3}{24}$ through a period and the right when it is $\frac{19}{24}$ through. The left hand plot shows that a kernel bandwidth of less than 1 gives a lower average score in the early part of the period; the right hand plot shows that a value greater than 1 is optimal nearer the end of the period.

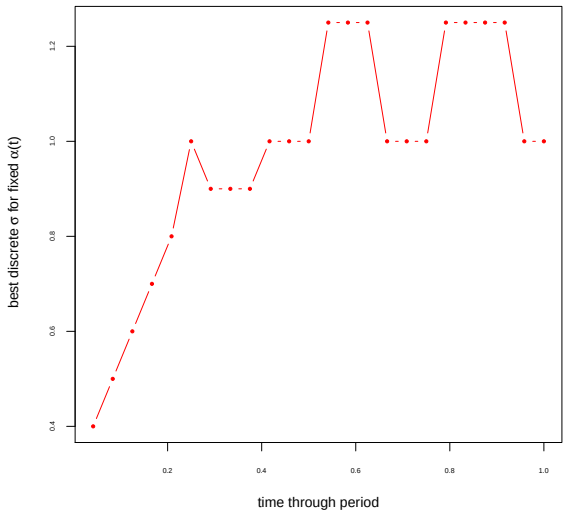


Figure 3.26: Discrete best σ_0 values for each value of $\alpha_0(t)$

Grid-best values The process above keeps one factor fixed for each step of the optimisation. Better values still may be found if the two parameters are permitted to vary simultaneously. This is explored in this section. The discrete-best values $\alpha_0(t)$ and $\sigma_0(t)$ values determined above are placed at the centre of a grid of trial values in each variable. The average score is calculated at each of these trial values (some higher some lower than the central values). In several cases the best pair is not at the centre of the grid indicating better value pairs are available than derived by the one factor approach above. Figure 3.27 shows an example of a contour plot of α against σ with the grid-best value highlighted.

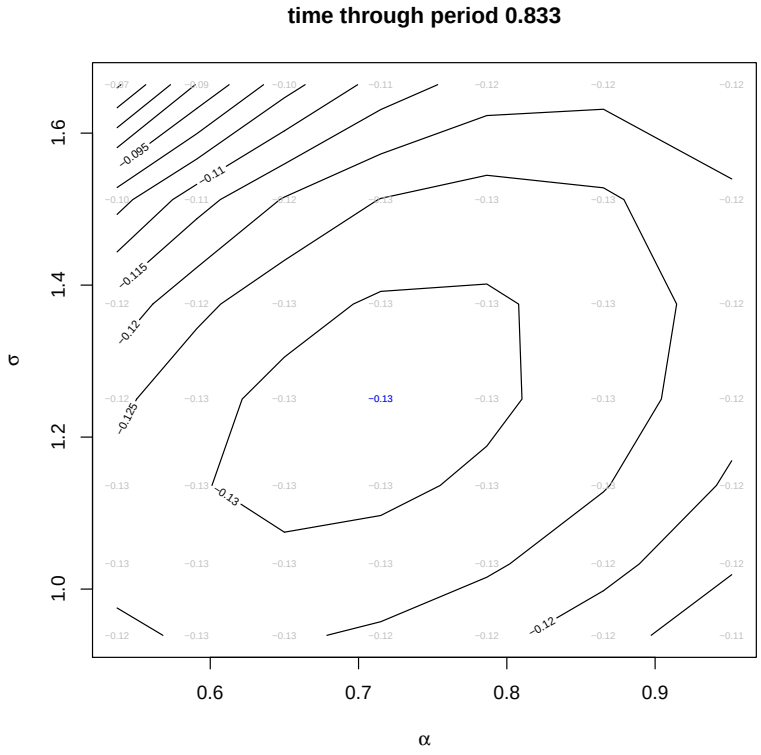


Figure 3.27: Contour plot of average Ignorance score for different values of α and σ. Average score values are calculated at 49 grid points with the one factor best $\sigma(t)$ and $\alpha(t)$ taken as the centre of the grid. The values of the average score are shown in grey and plotted at the intersection of the grid lines. This example illustrates the picture for observations 0.833 through the period. In this case the grid-best (minimum Ignorance) parameters are in the centre of the grid.

The joint optimisation process described above was carried for each observation time. Figure 3.29 shows the resulting grid-best parameter values as black circles with the chosen (smoothed) parameter values as blue lines and dots. Since the forecast skill gets progressively worse over the period it is appropriate that α (figure (b))

should monotonically decrease and similarly σ (figure (a)) should increase over the period. Note, α has been constrained to take a maximum value 97.5% to ensure that every forecast contains an element of climatology which, in turn, ensures that no forecast ascribes zero probability to an observed value (unless that value is outside of the observed climatology). The value of 97.5% is chosen arbitrarily to be very close to 100%. Figure 3.28 shows the σ and α parameters with the times they apply to shown as text, this shows an (almost) right angular effect where at the start of the period the value of σ gradually increases to reflect the widening funnel of doubt in the forecast - but the value of α remains close to unity because the forecast still contains lots of information. Roughly half way through the period the opposite occurs - the kernel bandwidth remains constant and the value of α falls away consistently as the forecast gets progressively worse and climatology does better. As noted previously α does not decrease to zero indicating that the kernel dressed ensemble retains some value later in the period.

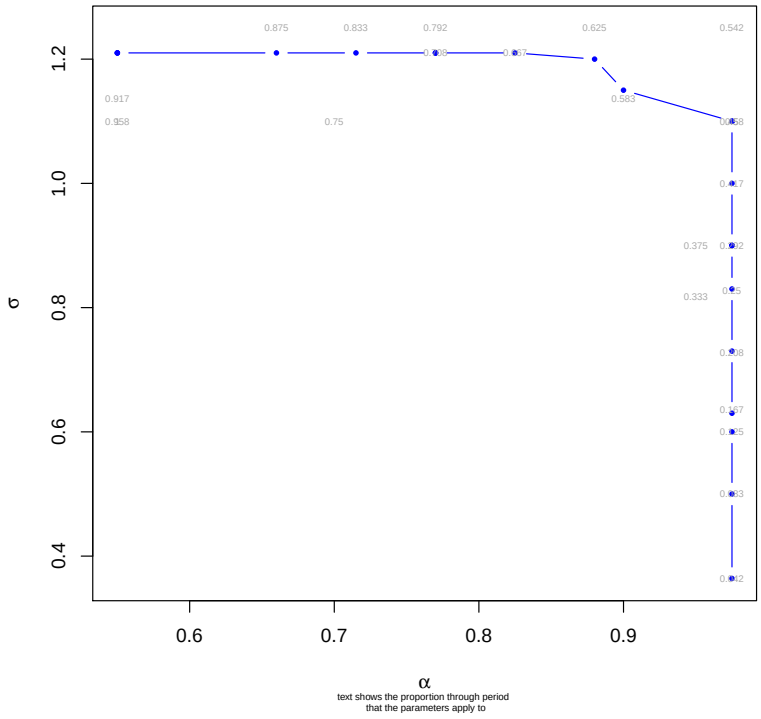


Figure 3.28: σ vs α - where the time that the parameter pairing occurs is shown in the text - the blue dots show the chosen parameters which fit through the grid-best parameters.

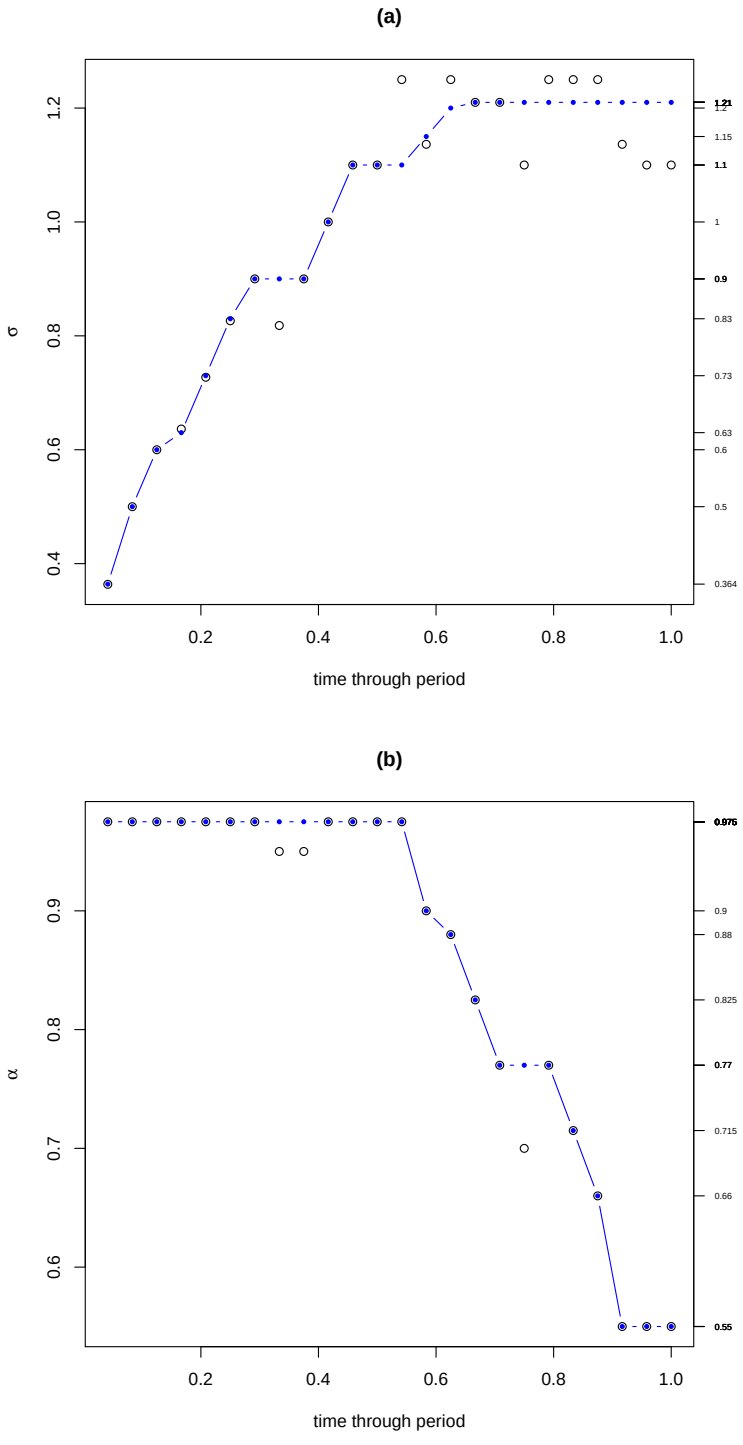


Figure 3.29: Blending parameters: chosen (manually smoothed) σ (figure (a)) and α (figure (b)) values (blue lines and dots) for different times through the period; grid-best values (before smoothing) are shown as black circles.

3.9 Climatology Blended Forecasts: 10009-12

Figure 3.30 compares the blending parameters for the other models of system 80001. These were chosen by using the parameters used for forecast 10008 as a starting point and then creating a grid of values around these to test for scoring improvements. The grey shading shows the range of α and σ values that were tested at each time during the period.

Some conclusions are:

- All models follow similar patterns (α tends to diminish over the period and σ to increase);
- The grid-best parameters at a given time (dots) show some scatter due, to sampling error, which has been smoothed away in the final manual selection;
- Models 10008 and 10009 have similar parameters although the weight given to the kernel dressed ensemble is lower for the latter model;
- Under Model 10010 the value placed on the kernel dressed ensemble falls away quickly and a high kernel width is applied. This suggests that the model has low skill and section 3.11 shows that this is indeed the case;
- Model 10011 retains the maximum permissible level of α (i.e. 0.975) suggesting that the kernel dressed ensemble retains considerable value late into the period. The kernel width σ also remains much lower than the others but does widen over the period;
- Model 10012 ($F = AR(4)$) has a lower value for the α parameter and a higher value of σ than forecasts 10008 and 10009 - this is despite recreating the climatology of the system better as shown in figure 3.20. The σ value for this model has been allowed to reduce below a high point in the middle of the period because the there were no higher values to support the monotonic rule in this case.

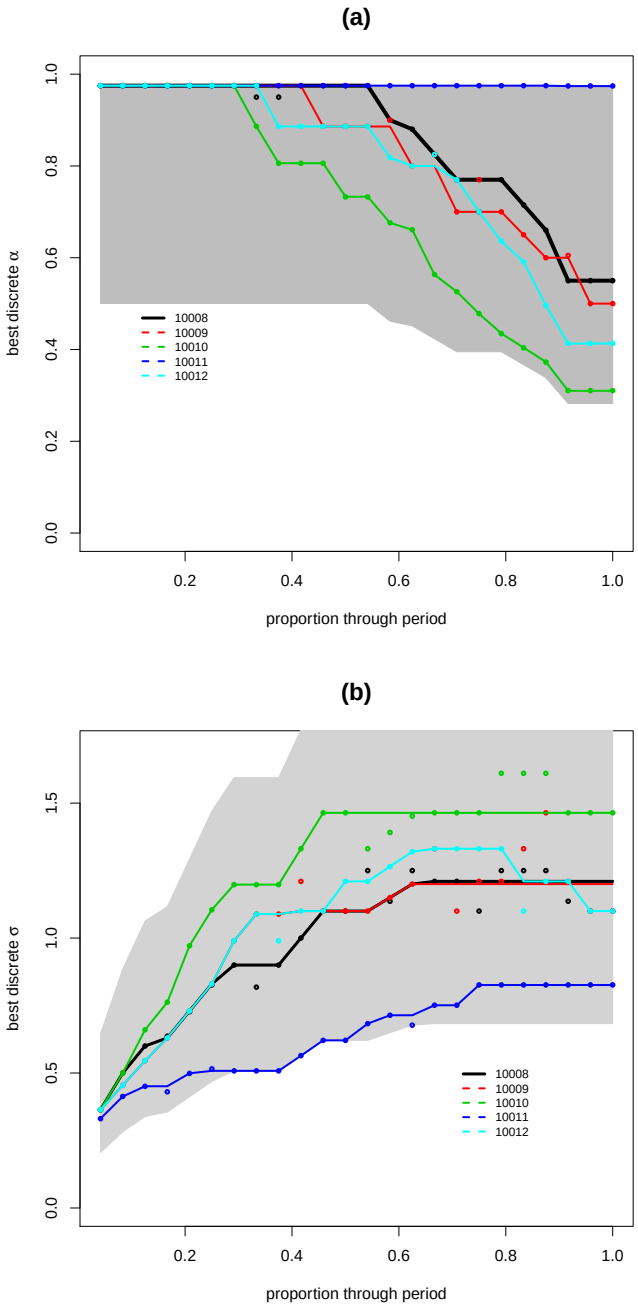


Figure 3.30: Chosen blending parameters for the other forecasts. Black line shows the parameters for forecast 10008. Grey shading shows the range of parameter values tested in the grid. Dots show the gird-best parameters (i.e. those giving the best score on the grid a particular proportion through the period); lines show the chosen manually smoothed values.

3.10 Scoring forecasts - Model 10008

This section illustrates how forecasts produced using model 10008 perform when scored against out of sample observations with the Ignorance score. As with earlier time series graphics, figure 3.31 displays the System observations in red; the forecast ensemble values are shown in grey. The Ignorance score values for the given observation and climatology blended forecast are shown as a black line relative to a secondary y-axis at the right of each plot. Various periods are illustrated which were chosen to illustrate periods in which a particularly bad period average score arises. There are different ways this can happen:

- In period 6 the observed values in the middle of the period fall outside of the forecast ensemble - on the high side, climatology blending ensures the score is finite but nevertheless it spikes. A poor score occurs for several values in succession, this is typical;
- period 316 shows a poor score at time $\frac{9}{24}$ - this is interesting because the observed value falls within the range of ensemble values but mid-way between two groups of values - the blended score will still be low between the two probability peaks - the CRPS would have been less sensitive to this forecast bust based on the work in Chapter 2;
- period 361 shows the opposite case as period 6, i.e. the observed value falls below the ensemble range;
- period 461 shows a long run of observations that are higher than the ensemble range.

Figure 3.32 shows how the scores (for forecasts of the variable X_1) vary over the period. The graphic shows that the median score rises from less than -2 to approximately zero. Scores are shown relative to climatology and so a score of zero means that the forecast is adding no additional skill at that stage. The coloured lines show quantiles of the score values and the dots show all values outside of the interquartile range.

Figure 3.33 shows that the pattern of change in average score for each of the 36 X variables. The behaviour of these variables in the system and for any one model

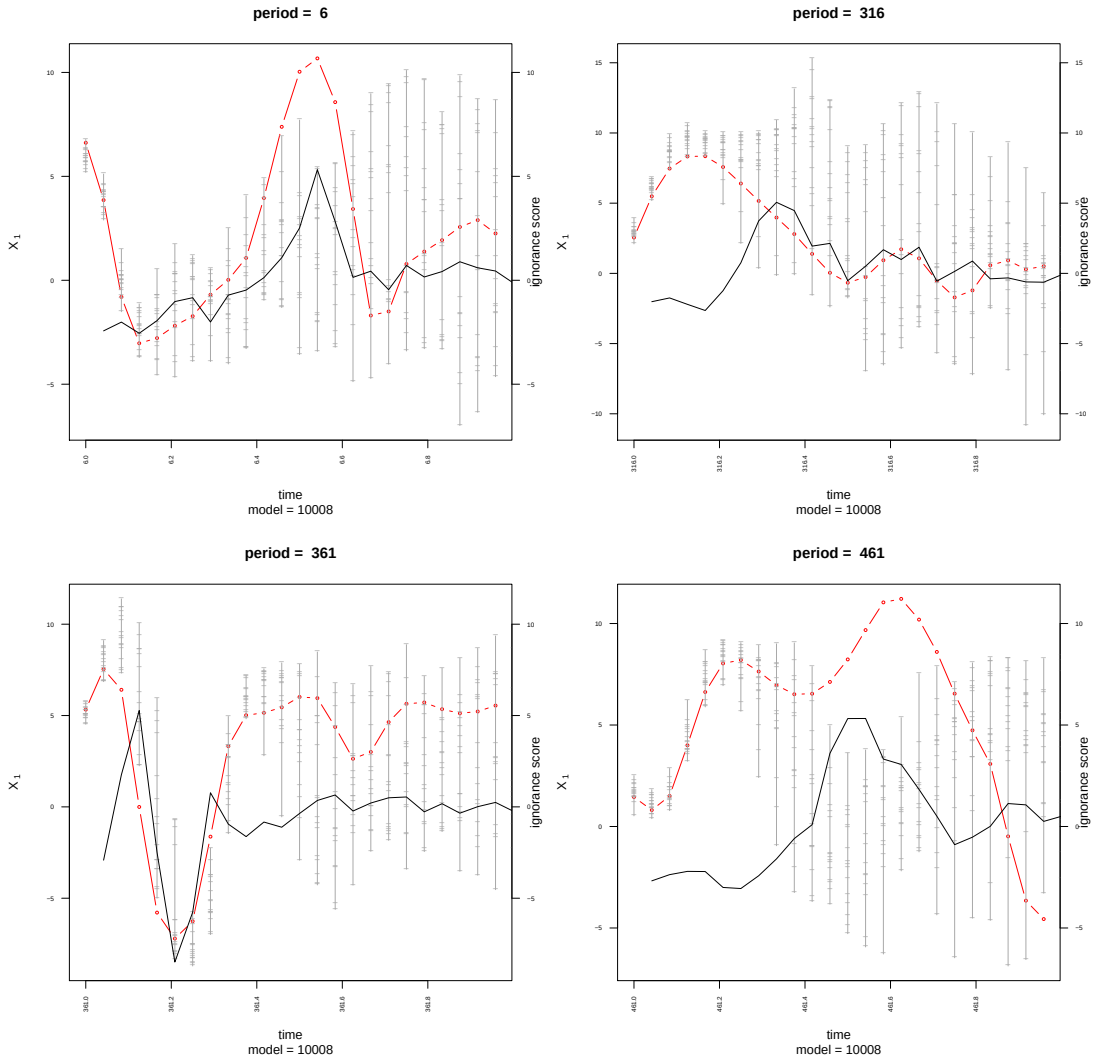


Figure 3.31: Comparison of system observations (red) and forecast ensemble for model 10008 (grey dashes and grey bar). The Ignorance score for the given observation and climatology blended forecast is illustrated with a black line whose values are shown on a secondary y-axis to the right of the plot.

is equal in distribution. This remains true for the more complex models 10011 and 10012. In the case of 10011 the functional relationship between X_k and F_k is the same for each k. In the case of 10012 the parameters of the AR(4) process used is the same for each k. Given the scoring process is also the same for each X_k the plot of average Ignorance over the period should converge in the limit. Therefore a plot of the sample average Ignorance for each of the variables illustrates the degree of uncertainty in the conclusions. The average Ignorance for variable X_1 is shown in black in the plot and the equivalent lines for the other 35 variables are all shown in grey. Since these all display the same behaviour and are close to the black line it is

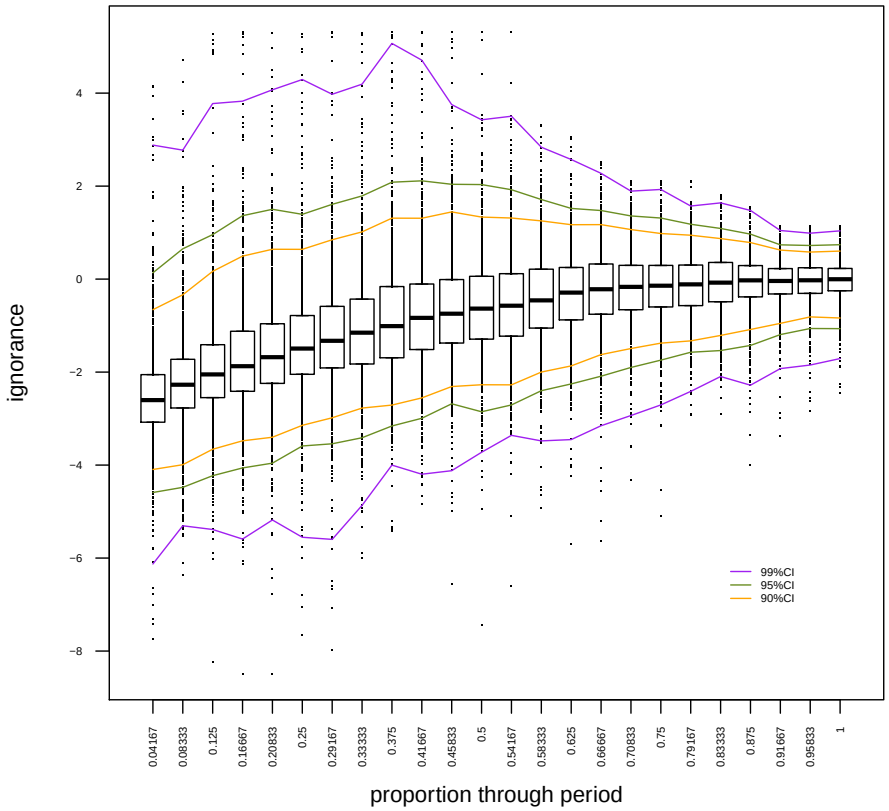


Figure 3.32: Box plots of Ignorance values for forecast 10008 at different times during the period. Coloured lines show the quantiles of the distributions. Dots show all values outside of interquartile range.

safe to conclude that the relationship over the period is well described by the black line and free from material sampling error.

The value of climatology blending for forecast 10008 was explored by comparing the blended forecast with a kernel dressed forecast using the same kernel width (but no blending). In the early part of the period the two approaches give very similar scores, this is to be expected because the α parameter is close to 1 at this time and the climatology is given very little weight. Towards the end of the period, however, two things occur (1) the ensemble members typically drift apart and (2) more weight is put on climatology. This additional climatology weight does lead to a better scoring forecast (as intended) but the improvement in the Ignorance score is less than 0.2.

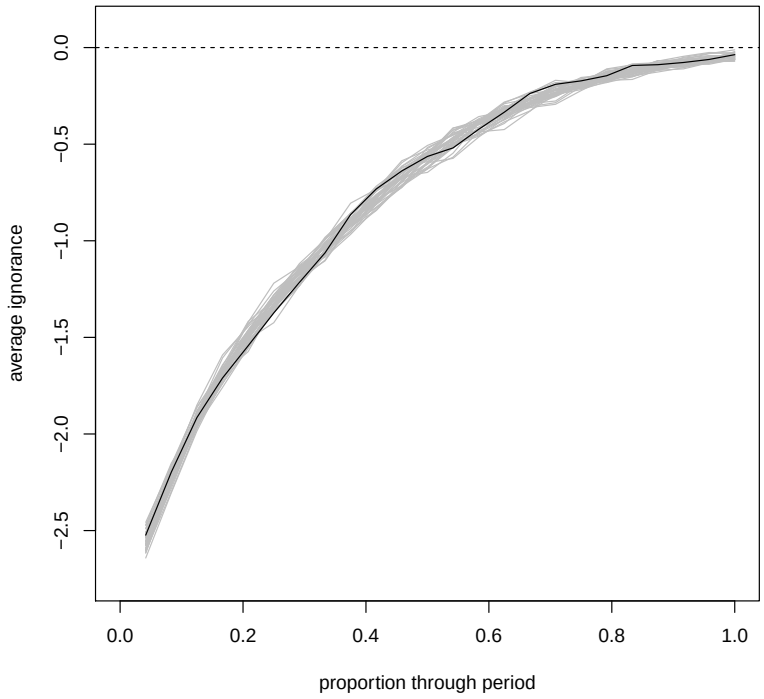


Figure 3.33: Black line shows average Ignorance at different proportions though the period for the climatology blended forecast, for forecasts of the variable X_1. Grey lines show the average score for the other 35 variables $(X_2, ...X_{36})$

The average period score can be created by taking the scores at each discrete time step during a period and averaging them to create a single score measure for the period. Whilst the average of proper scores may not be proper this measure does allow two forecast methods to be compared over a whole period. The top left graphic in figure 3.34 illustrates (for X_1) the probability density of the observed period scores over the 1184 periods tested. As above the figure shows the climatology blended case in red and the kernel dressed (only) case in blue. Blending narrows the distribution leading to less periods with bad forecasts (positive values) but also less chance of a very good forecast (negative values). The top right graphic shows that the period-average Ignorance is better (lower) in the blended case. The box plots are taken over forecasts of all 36 X variables. The grey lines represent the change in value for each variable between the kernel dressed and blended case and it should be noted that, whilst the box plots do overlap, each of the grey slopes is negative so the reduction in mean occurs in every case and the result is robust. The

bottom graphic in the figure shows the standard deviation of period scores over all 36 variables - again showing that the narrowing of the density illustrated in the top left plot is a general result.

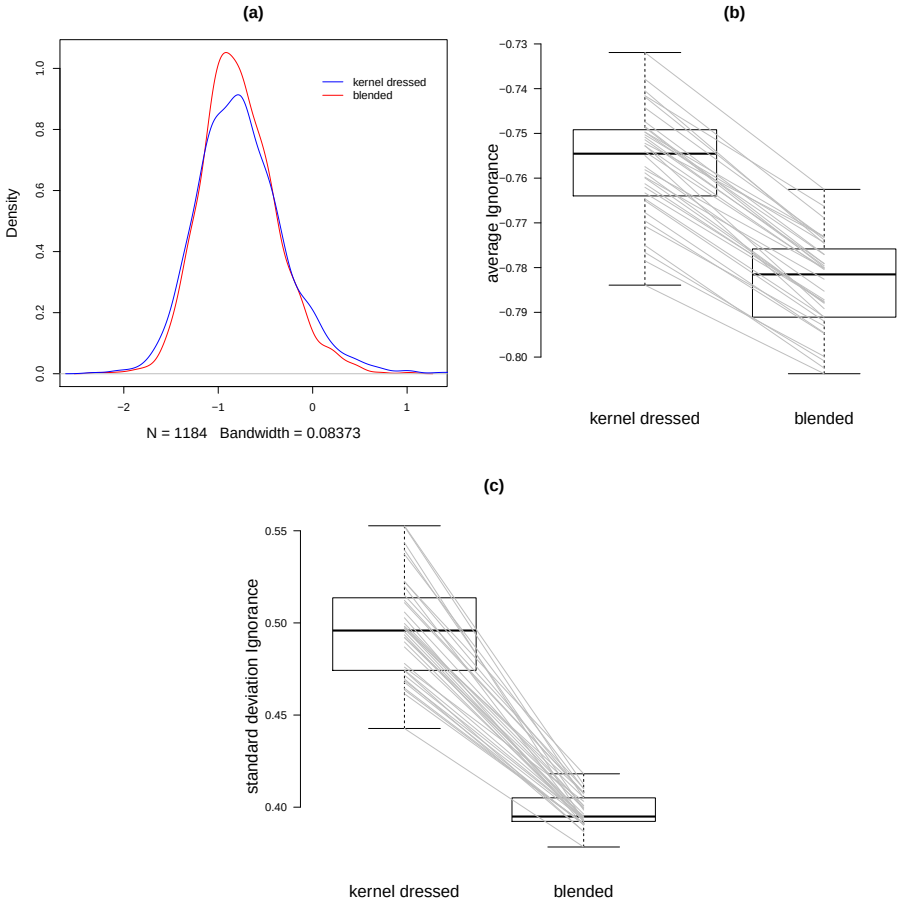


Figure 3.34: Average period forecasts, comparison of climatology blended forecast and kernel dressed (only) forecasts. Top left shows density of average period score over 1184 periods; top right shows box plots of the mean average period score over 36 X variables; bottom plot shows the range of standard deviation of the average period score for the same variables. In each of the box plots and for a given variable X_k a grey line is drawn between the value of the statistic in the kernel dressed case and in the blended case.

Model 10008 was compared with two sensitivity tests - the first 10008a uses the same value of σ but uses α values that are 10% lower (i.e. more weight put on climatology). The second 10008b keeps α the same but sets σ 10% higher. This has two effects: (1) the full range of scores is slightly narrowed (this is most noticeable in 10008a where the reduction in really bad scores is much reduced) and (2) the median score (and mean) is worse - but not much worse. So by giving more weight to climatology (or using a wider bandwidth) some skill is given up at the mean but

there are fewer forecast busts.

Examination of whether two forecasts can be blended to further improve the average score Given two models that provide forecasts over a period of observation. If there are times during the period when one model does reliably better than the other, but then other times when the situation is reversed, then there would be a benefit to blending the forecasts together. This has already been exploited above where the climatology forecast is blended with those from the models considered. The pairwise difference between the scores of each model were compared and in each case one model performed better than the other throughout the entire period. Hence there is nothing to be gained from blending two of the models together on this occasion.

3.11 Scoring forecasts - models 10009-12

This section uses the Ignorance score to compare all the models of system 80001 to see which (if any) outperforms the others, over 1184 periods. First the average scores at each observation time are compared to see if different models do better at different times through a period. Next, different score quantiles at each observation time are compared to see whether some models have worst forecast busts than others. Finally, the period average scores (i.e. the average score over all observation times within a period) are compared. Insurers typically offer cover over a period rather than at a point in time and it is therefore important to know whether a model does well over the whole period rather than just at certain points during the period, for this reason period average scores are also considered. Model 10011 emerges as the best model on most measures, consistent with the fact that its blending α parameter stays higher for longer during the period.

Figure 3.35 shows the mean Ignorance score at different times through the period for the 5 forecasts. As expected, forecast 10011 (where $F = IEF(X)$) scores considerably better than the others and indeed retains skill relative to climatology to the end of the period. Forecast 10010 ($F = 8$) scores worse than the others. 10009 ($F = E(IEF)$) does slightly better than 10008 for most of the period and 10012 ($F = AR(4)$) does worse at the start of the period than all the other models.

Figure 3.36 is similar but compares chosen quantiles of the score at different points during the period (over 1184 periods). Forecast 10011 does best for the low quantiles (i.e. when it does well it does really well). The story for high quantiles (i.e. when the forecasts are doing badly) is more interesting. At the start of the period all the methods are quite close. At the end of the period, however, forecasts from models 10008, 10009, 10010 and 10012 all see their worst skill scores diminishing - this is due to the fact that for each of these forecasts a great deal of weight has been put on climatology by this point. Conversely when forecast 10011 does badly it does really badly - because little weight has been put on the climatology.

Figure 3.37 compares the period average scores of the forecasts. The diagonal shows which row and column a given forecast falls within - and also shows the mean period average score over all 1184 periods. Despite its occasional very bad scores at the end of the period, forecasts from model 10011 have considerably better average scores than the other models. The right hand triangle shows the correlation between period scores; for example the correlation between the period scores of forecasts 10008 and 10010 is 0.6. The bottom triangle shows scatter plots of the period scores over all 1184 periods observed. In the scatter plots the x-axis always relates to the forecast named in the column and the y-axis relates the forecast named in the row. The line $y = x$ is shown for easy comparison. Forecast 10011 does better than the others in the majority of periods (since the scatter points are almost all one side of the line) and 10010 clearly tends to do worse more often than it does better.

The next chapter will explore the forecasts in an insurance setting and will discover whether the occasional busts of forecasts from model 10011 are serious impediment or something that can be tolerated given its general out-performance of the others.

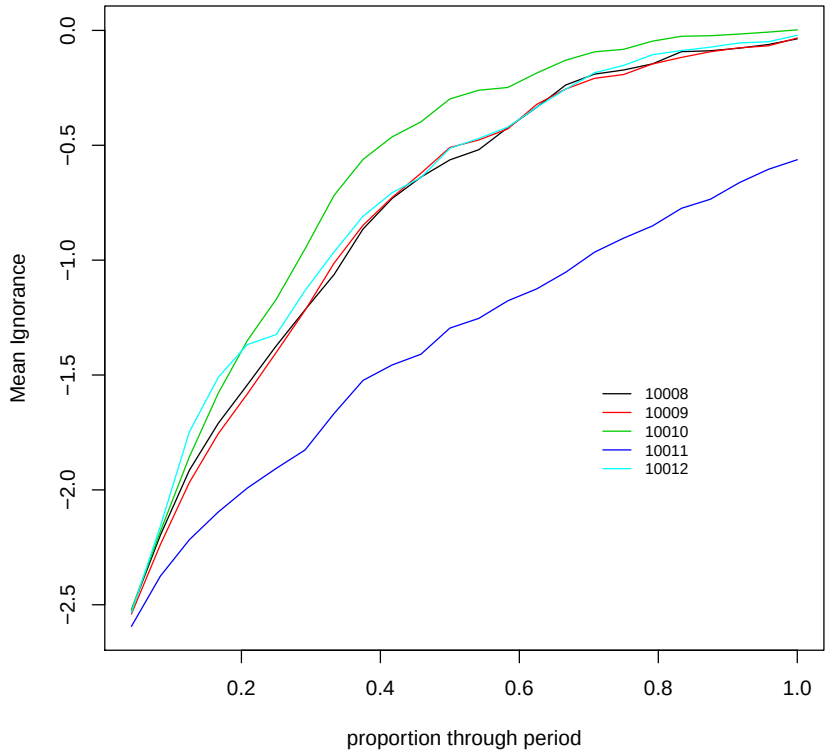


Figure 3.35: Comparison of scores for models 10008 to 10012. Mean score over 1184 periods - at different times during the period. Model 10011 outperforms the others and retains skill relative to climatology through the entire period.

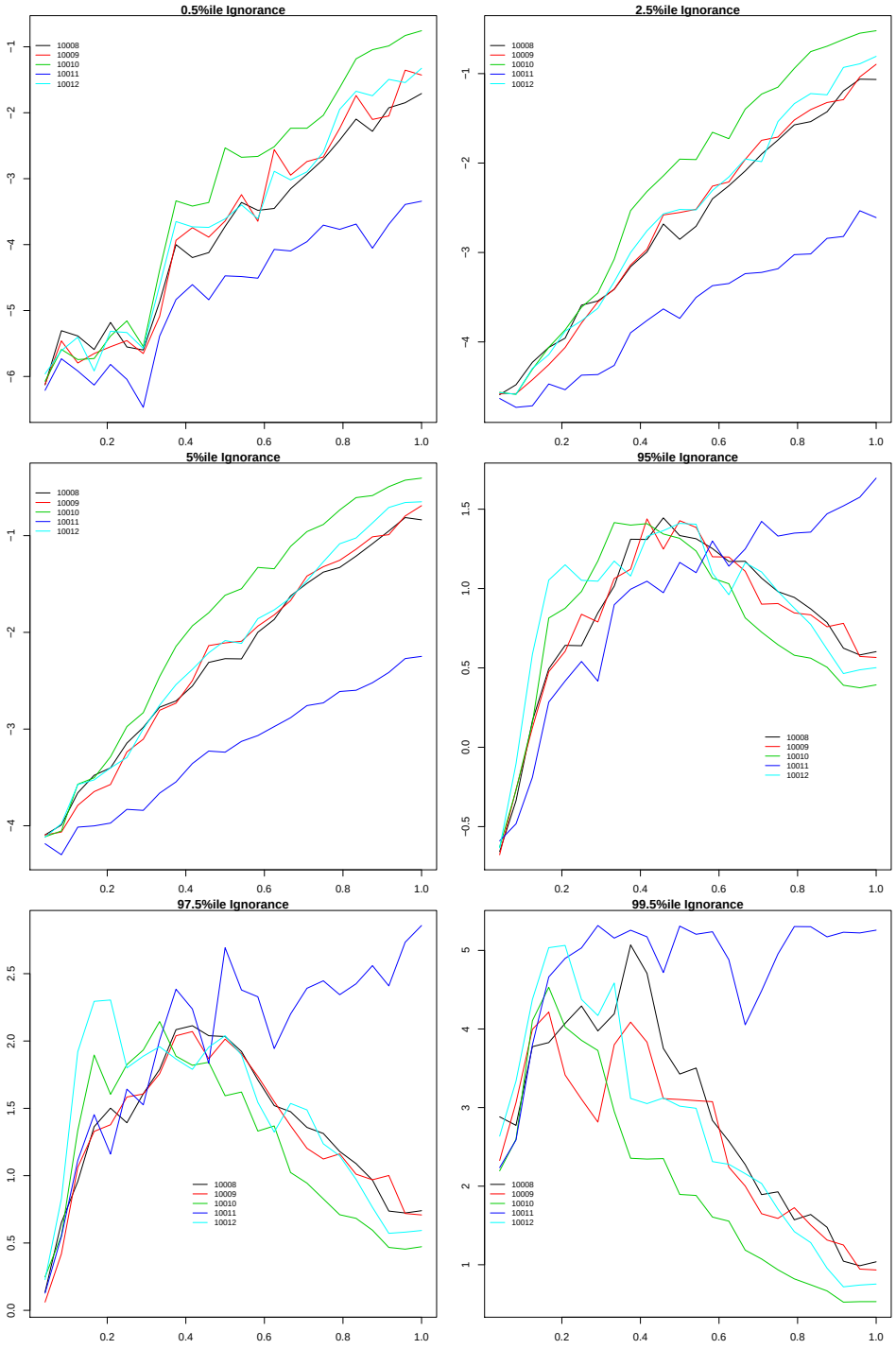


Figure 3.36: Comparison of skill score quantiles for forecasts from models 10008 to 10012. For the low quantiles (Top left, top right and mid left graphics) model 10011 performs better than the others throughout the period. For the high quantiles (mid right, bottom left and bottom right graphics) model 10011 performs best for the first half of the period but worst for the second half.

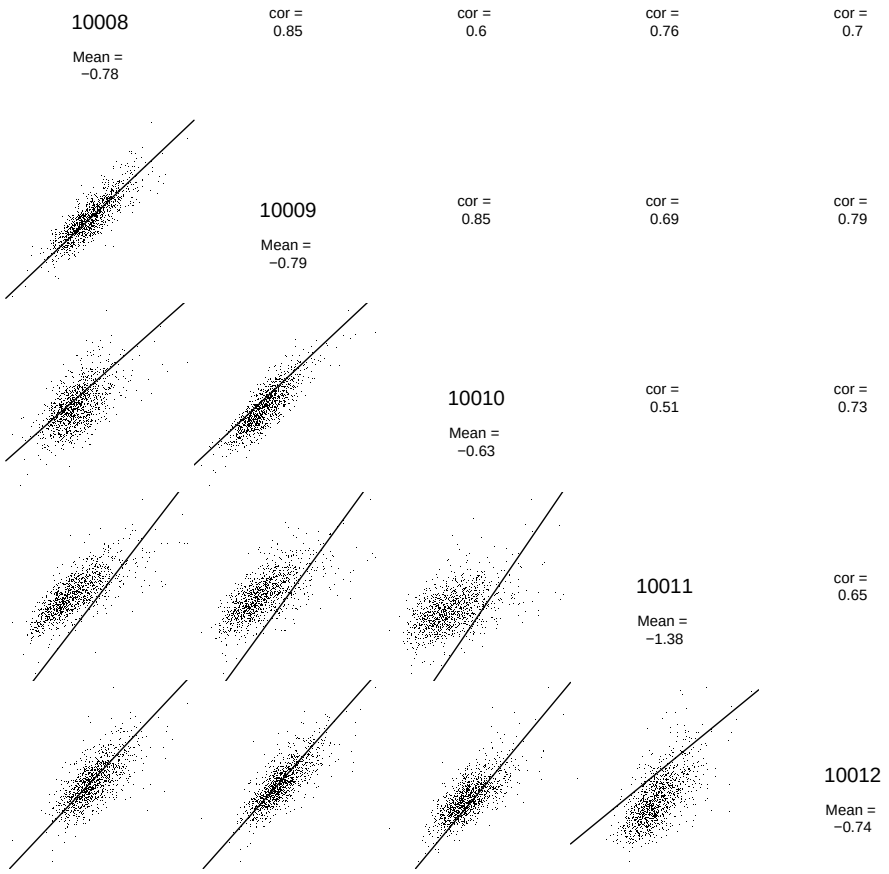


Figure 3.37: Comparison of forecasts 10008 to 10012. Diagonal shows forecast ID and also the mean of the period average scores over all periods. The lower triangle shows scatter plots comparing the period average score from each pair of models. Let Model A (x-axis) be defined by the model label in the column above and model B by the row label to the right, then each point in the scatter plot is the period average score from model A compared to that of model B; the line $y = x$ is shown for easy comparison. Model 10011 clearly outperforms the other models since the scatter points are almost all one side of the line. The top right triangle shows the correlation coefficient between model A and model B defined by the label in the column below and row to left.

3.12 Climatology Blended Forecasts: 80002-6

As described in section 3.9 the ensemble outputs from each model can be kernel dressed and climatology blended to produce best scoring forecasts. The Ignorance score continues to be used in all cases. The methodology described earlier is used except that the parameters for system 80001 are used as the central values of the grid of parameters tested. The grid size was enlarged, as necessary, to ensure that the minimum average Ignorance had been found in each case. The grid-best values of of α and σ are shown in figures 3.38 and 3.39 as empty circles along with the manually chosen monotonically increasing and decreasing final choices as solid lines; note that the circles have been jittered slightly in some cases so they do not overlap. The graphics are plotted on a 3 row and 2 column grid, the phrase 'position (i,j)' will be used to refer to the graphic in column j and row i. System 80001 has been described in detail above and this discussion is not repeated but values for easy comparison are repeated in plot (1,1). The following observations are made:

System 80002 $F = 20$**,** $h_x = 1$ **plot (1,2)** The α parameter falls faster away from the best value of 1 than system 80001. It also falls close to zero for forecasts at the end of the period (for all models but 10111). In tandem the σ parameter rises quickly to a value above 1.4 showing that a wide kernel dressing bandwidth is required to improve the skill score. Note that the grid-best values for σ can be variable at the end of the period due to sampling error. This is not a cause for concern, however, because this is where significant value is given to climatology and so the σ parameter plays little role in the forecast. The blending parameters of model 10111 (i.e. where $F = IEF(X)$) continue to give more weight to the forecast than the other models. In summary, system 80002 is less predictable as expected and the blending parameters cater for this.

System 80003 $F = 20$**,** $h_x = 0.1$ **plot (2,1)** The α parameter follows a similar pattern as system 80002. For each value observation during the period the value is less, however, indicating less weight is put on the forecast and more on climatology. This was not expected as the lower coupling parameter was expected to lead to better predictability because the closer to zero the coupling parameter is, the less

affect the Y variables have on the X variables; in the limit when h_x is zero the system and model (with same forcing) converge. Hence models in system 80003 were expected to be 'closer' to the system and therefore to give better predictability. The reason that this is not the case is suggested by the difference between systems 80002 and 80001. In system 80002 the forcing is much larger and this leads to lower predictability suggesting a general rule that larger effective forcing leads to lower predictability. The effective forcing of system 80003 is 19.9 compared with 18.9 for system 80002; this is due to the fact that the impact of the Y variables is to reduce the effective forcing on average (as mentioned above). Hence system 80003 is a higher forcing system and inherently less predictable than 80002. In summary the effective forcing exerts a much stronger influence on predictability than coupling which was not anticipated.

System 80004 $F = 10$**,** $h_x = 0.1$ **plot (2,2)** This system was chosen to be the pair of system 80001, it has the same forcing but lower coupling. Again this was expected to be more predictable and as with 80003 the opposite is true for similar reasons.

System 80005 $F = 9.1$**,** $h_x = 0.1$ **plot (3,1)** This system was chosen after it was discovered that system 80004 was not as predictable as expected. This system is chosen to have the same effective forcing as system 80001 but, due to the lower coupling parameter, was hoped to be more predictable. The results here are interesting. For the models *8, *9, *10 and *12 the weight put on the forecast remains higher for longer and also the kernel dressing bandwidth remains lower for longer, both of these indicate more predictability as hoped. The blending parameters of Model *11 give lower weight to the forecast than 80001. In system 80001 the α parameter stays close to 1 for the whole period, in system 80005 this model does retain a higher α for slightly longer than the other models but falls away from 1 a little under half way through the period. Model *11 makes use of the strong correspondence between a value of X and the effective forcing at that time, when the coupling is reduced the utility of this relationship is also reduced leading to less weight being placed on the forecast for this model.

System 80006 $F = 11$**,** $h_x = 1$ **plot (3,2)** The broad trend in the blending parameters in this system is quite similar to 80001. The stronger coupling than 80005 and greater weight placed on the forecasts from model *11 supports the hypothesis that a stronger coupling is actually beneficial to predictability in this case. Due to the higher effective forcing in this system ($E(IEF) = 9.98$) the models place less weight on the forecasts than for 80001, however, and the value of α falls further indicating more weight being put on climatology. This result is consistent with expectations.

3.13 Scoring forecasts in systems 80002-80006

Comparison of skill scores for each system and each observation time Figure 3.40 compares the average score at each observation time through the period for each of the systems. The same ordering is used as before and the graphics will be referred to using the same row/column format. System 80001, plot (1,1), has been described before. For system 80002, plot (1,2) the Ignorance score rises quickly and by half way through the period is zero indicating there is no skill above climatology. The same is true for system 80003, plot (2,1) except that the skill falls away quicker reaching the climatology average by time 0.4 - this is another illustration of the system that was expected to be more predictable being less so. Systems 80003, 80004 and 80005, each have low coupling of 0.1 and show little difference in average score at each observation time between the models. Models *11 stand out in each of the systems where the coupling is larger - for each of the observation times. In the case of systems 80001 and 80006 skill above climatology is retained to the end of the period for model *11.

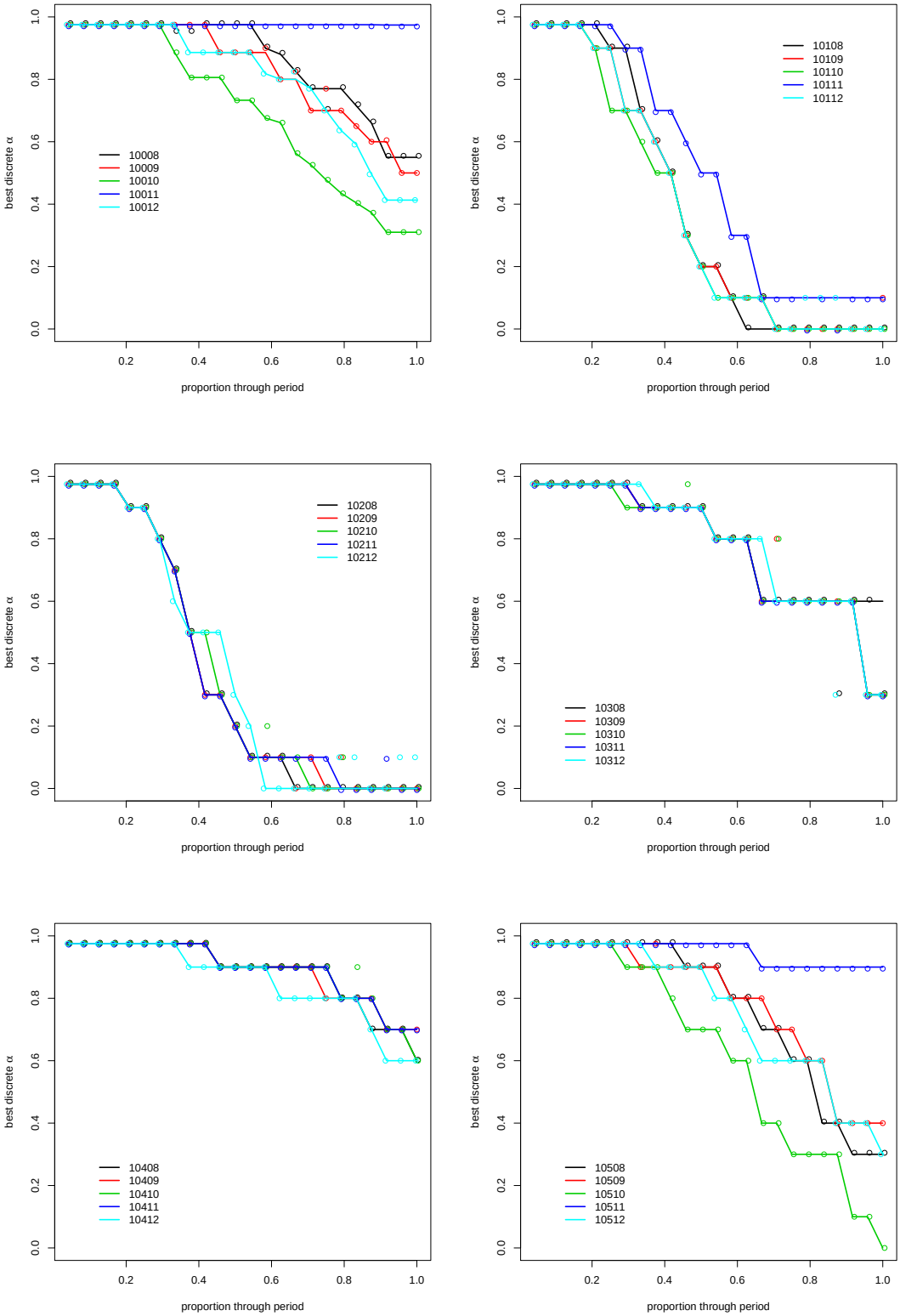


Figure 3.38: Grid-best α value over the period- lines represent manually smoothed values and dots represent the best points on the tested grid. The model numbers are shown in the legend of each plot. The corresponding system can be inferred from these.

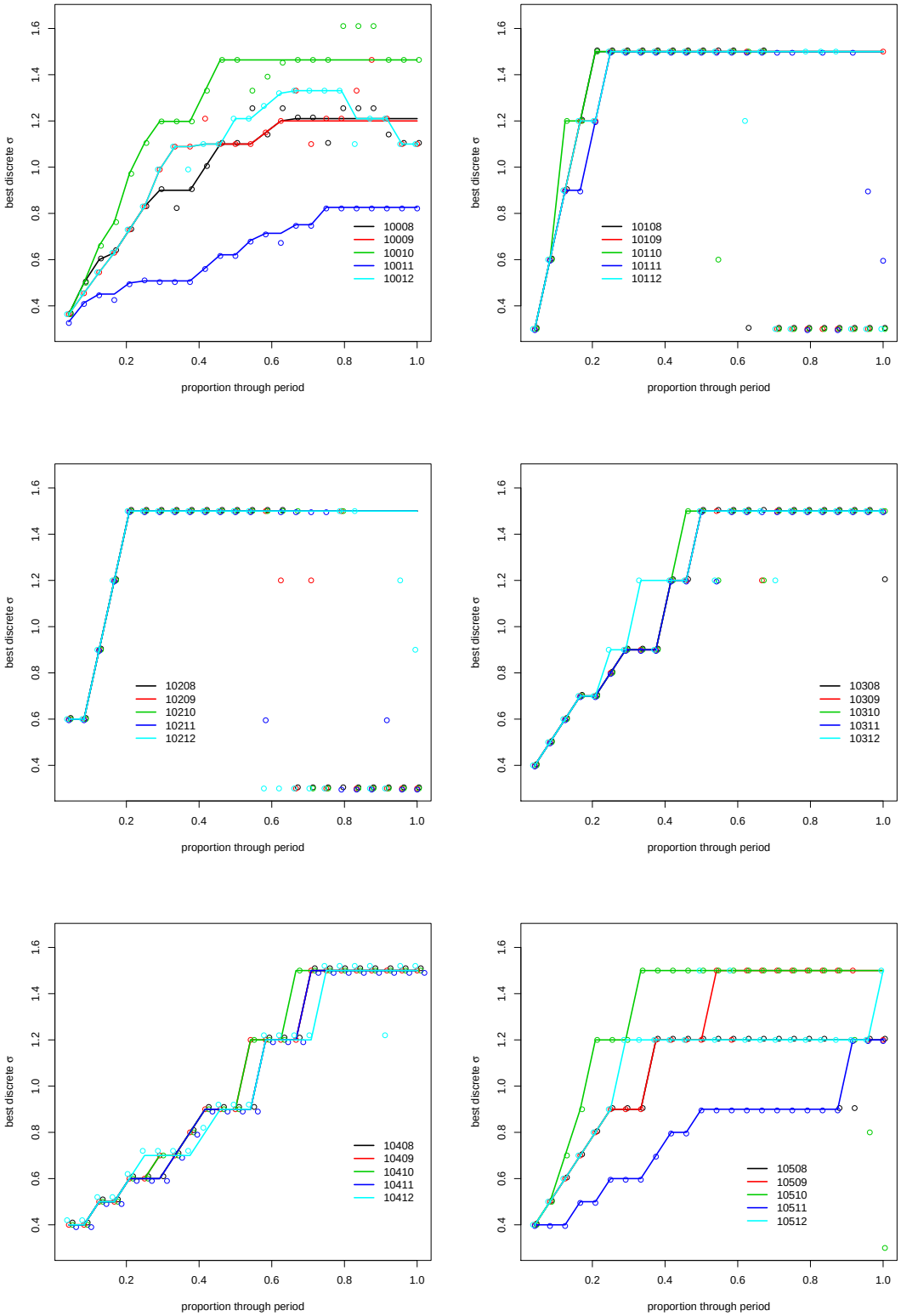


Figure 3.39: Grid-best σ - lines represent manually smoothed values and dots represent the best points on the tested grid The model numbers are shown in the legend of each plot. The corresponding system can be inferred from these.

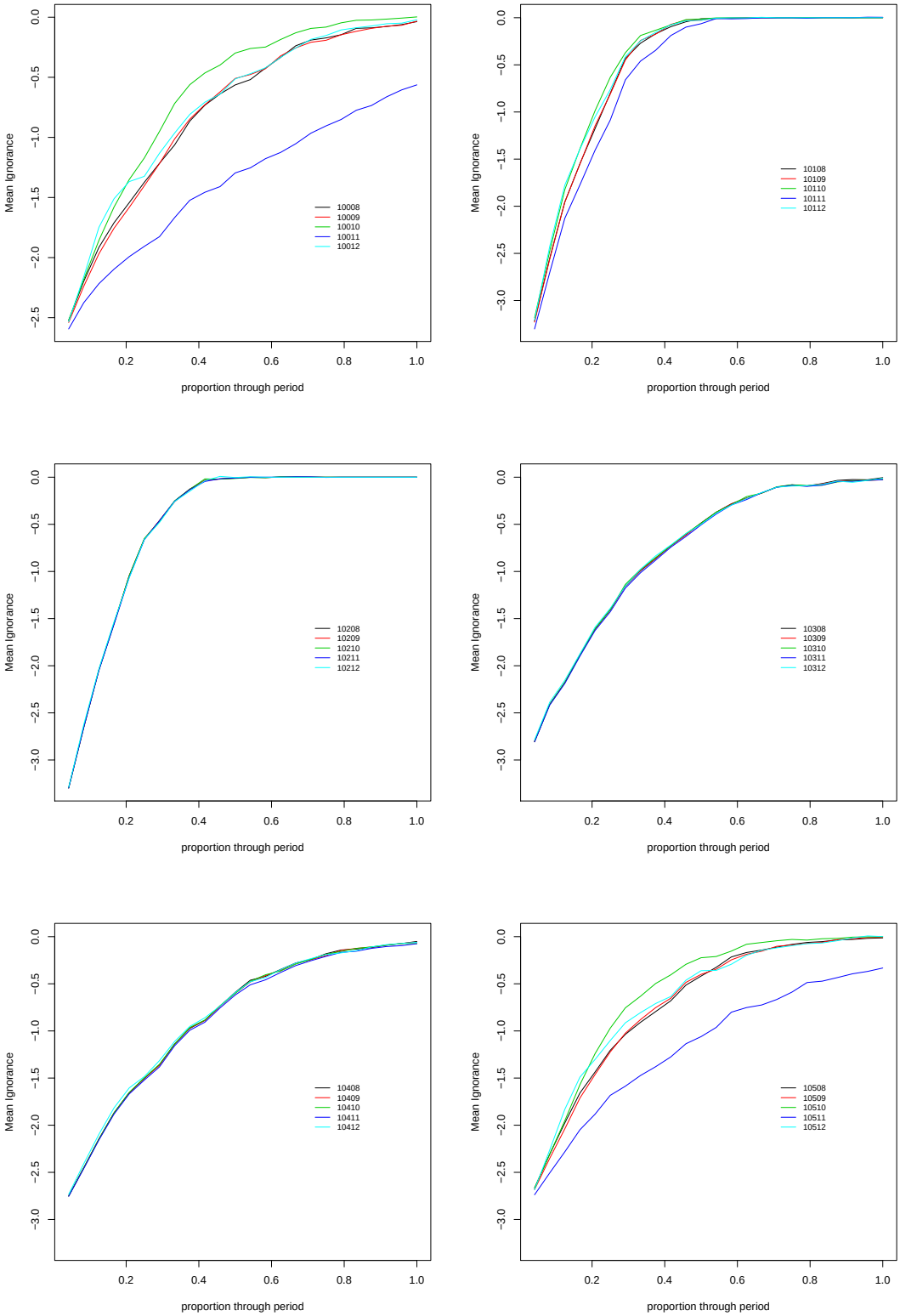


Figure 3.40: Mean Ignorance score at different proportions through the period $\frac{1}{24}, ...\frac{24}{24}$. Comparison of all forecasts from all model ensembles in all systems. Model ID is shown in the plot from which the system can be inferred.

Period average scores Figure 3.41 illustrates the period average scores for the various systems and various models described above. As noted above this is not a proper score but allows the models to be compared over a complete period. The average score for variable X_1 is highlighted by a red bar in each plot. In figure (b) a box plot of the scores for each of the 36 X variables is also shown in grey to indicate the variance in the estimate (given as usual that the X variables are identical in distribution). Figure (b) uses a bootstrap re-sampling approach to illustrates the uncertainty in the estimate. The different systems are demarcated by vertical grey lines in the plot. The key results are:

- As the forcing increases the predictability in the system (as measured by the skill score) decreases;
- Model *11 does best in all systems consistent with the results at each observation time. This is particularly the case where the coupling parameter is larger and the benefit diminishes materially when the coupling is small;
- Model *10 performs worst in all systems from this point of view;
- Models *8 and *9 perform similarly in all systems;
- Model *12 performs worse than the simpler model *8 despite its attempt to replicate the statistical behaviour of the IEF through an AR(4) process; this suggests that the lack of conformity with the dynamics of the system is a larger factor in model quality than the closer adherence of one of its parameters to the 'correct' statistical behaviour.

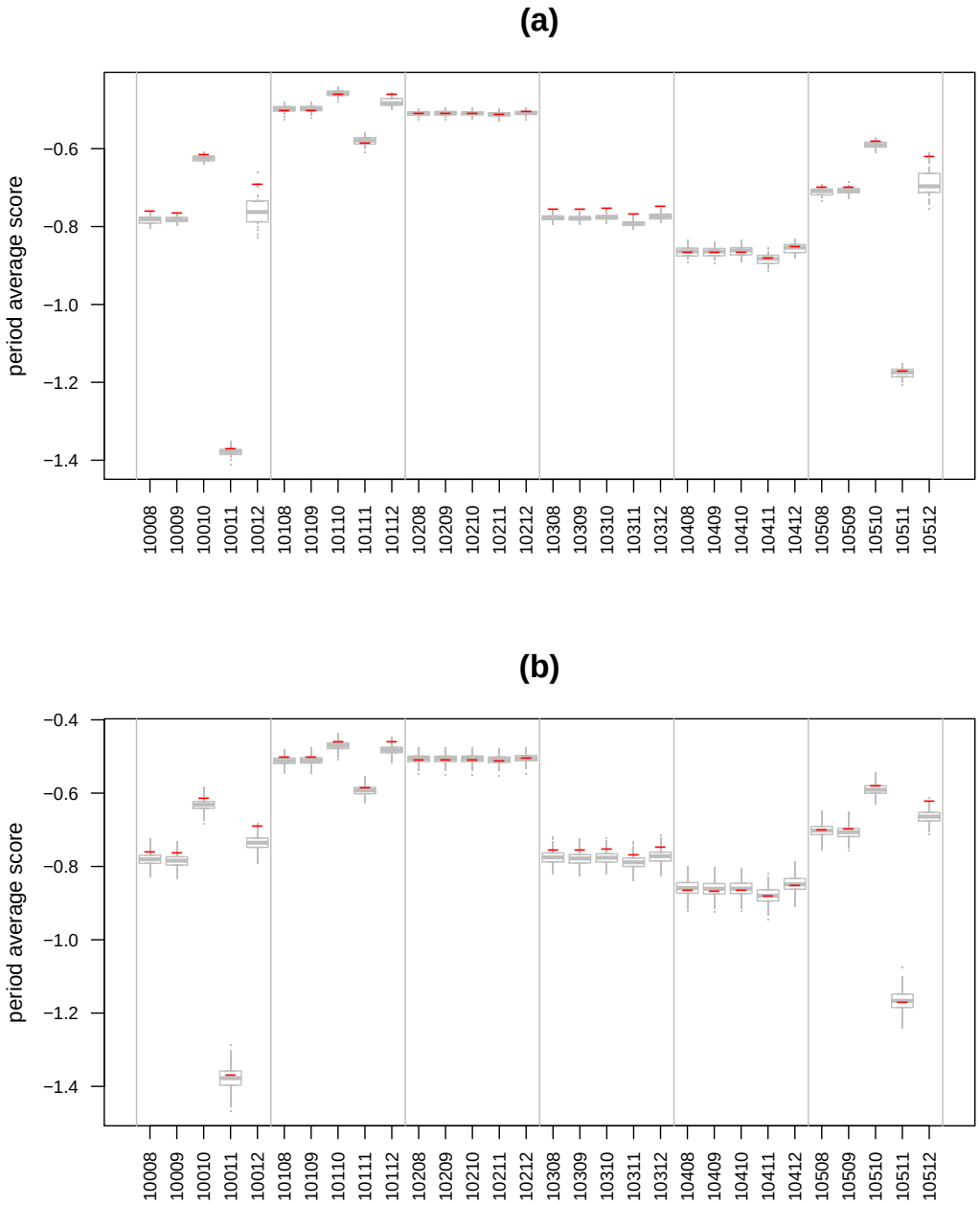


Figure 3.41: Comparison of period average scores for forecasts from each model. Figure (a) shows the average score for K=1 (red) compared to a box plot of the average score for the other 35 K variables. Figure (b) shows the average score for K1 (again red) but this time against a box plot of bootstrap resampled means. 2^9 samples each of size 2^9 are taken (with replacement) from the 1184 period scores available to indicate the uncertainty in the mean value. These resamples are chosen so that the same periods are chosen for each of the forecasts in each case.

3.14 Conclusions

This chapter has introduced six Lorenz 96 systems each with five models. Higher values of the forcing parameter F in the system lead to lower predictability. Reduced coupling h_x for a given level of forcing was expected to lead to increased predictability but this did not arise because the effective forcing in the system actually increases when coupling decreases and this is shown to be the dominant effect. Climatology blending is used at each forecast lead time over a chosen period of 24 observations. The blending parameters are observed and their behaviour illustrated and explained. As expected the weight put on the forecast diminishes as lead time increases and the kernel width increases; however the fact that these appear to operate in series rather than parallel was not anticipated. Models with a fixed forcing parameter do not score as well when compared to a model with a stochastic parameter based on the current state of the system. An AR(4) model for the instantaneous forcing parameter leads to observations that broadly match its statistics, but do not produce forecasts that score well. This shows that the common practice in statistics of matching process parameters is incomplete in the context of dynamical systems. The Lorenz 96 system will be used in the next chapter to explore the conditions in which forecasts can be useful in an insurance context.

Chapter 4

Predicting the Lorenz system - with applications to insurance

'***The key question concerning the usefulness of imperfect models, however, is not whether they will yield nearly optimal policies which could be obtained only by having perfect knowledge...but whether they can yield policies which are superior to an inactive policy allowing for no feedback.***'

Chow, 1976 [42]

This chapter considers whether forecasts can improve pricing in insurance. Section 4.1 describes a general method to create an insurable index from dynamical system output. This is analogous to various index products currently sold by the insurance markets today where, for example, the dynamical system can be the atmosphere or geological processes [116, 267]. The chapter ultimately seeks to test, by analogy, whether models of a dynamical system (such as WRF[1] or HadCM3[2]) can add value to insurance, using the Lorenz 96 system and models as a concrete example.

Before the key insurance question is tackled it is stressed that operational models only imperfectly describe the systems they are attempting to imitate [166]. For

[1]The **W**eather **R**esearch and **F**orecasting model (WRF) [276] is the numerical weather prediction model of the National Centre for Atmospheric Research (NCAR) [181], USA.

[2]The **Had**ley centre, **C**oupled **M**odel 3 (HadCM3) [169] is a coupled ocean and atmospheric model developed by the Hadley centre in the UK.

example, figure 3.19 of Chapter 3 shows that the range of target values taken by a model can be different to reality (system). Operational models, whilst not perfect, may still be useful [27,42] and section 4.2 introduces a new and general method, using skill scores, to recover a more useful 1-to-1 relationship between one dimensional projections of the models and system. Section 4.3 confirms that the method is successful in a special case when the sample space of the system and model are discretised in an ideal way; section 4.4 then demonstrates that the method remains successful when the relationship between the model and system is less idealised.

Section 4.5 gives a translation of the general setting into an insurance setting to provide a concrete example. The Lorenz system II and structurally imperfect models of it, described in Chapter 3, provide the basis for an example in section 4.6 showing that the use of the transformed model output in pricing leads both to improved profitability and to reduced risk of insolvency. Climatology blended pricing is also developed in this section; it is not as successful as the simpler expectation adjustment method, for this example. Section 4.7 concludes with a discussion for future work that could be undertaken using this new framework. The new work in this chapter is believed to be:

- Introduction of the ϕ-transformation method described in section 4.2 and the testing of this method in later sections 4.3 and 4.4;
- Application of the ϕ-transformation to the Lorenz 96 system in subsection 4.6.1;
- Development of a pseudo-insurance index, in general (4.1) and for the Lorenz 96 system (4.6);
- Use of ϕ-transformed model outputs in pricing of the Lorenz 96 index, via (1) updated expectation method and (2) climatology blending in subsection 4.6.2;
- Suggestions for extension of this work in future discussed in section 4.7.

4.1 Definition of an Insurable index

Weather risk transfer is currently a small market for insurers [185]. Traditionally, risk transfer is achieved by paying for the damage or loss of profitability that has actually occurred, determined after the fact by experts (loss adjusters), called in-

demnity insurance [47, 247]. Since the 1990s another form of risk transfer has been offered where the payouts are determined by agreed relationships between observed variables and the financial losses they relate to [5, 6, 39, 269]. For example a payout may be made each time the temperature exceeds 105F at the weekend (**'hot days'**); in this example, summarised in table 4.1, the system is the atmosphere and the particular variable is temperature. The insurance purchaser may be interested in the total impact of hot days at multiple locations in Florida over the summer. This generalises to an interest in multiple transformed variables measured over multiple times. In order to decide what payout they require for a given variable value the purchaser or broker would explore the likely impact that could arise; this could be achieved by a statistical analysis of past events or by using models [53]. The raw variables may need transforming into a more decision relevant index, for example the damage from hurricane winds. The following general description is developed with the above example in mind.

The following describes how an insurable index R may be created from observed variables of a dynamical system. Consider an arbitrary multidimensional system with multiple variables $X_1(t)$, $X_2(t)$, ...,$X_K(t)$. Given a vector of such system values at times $t_1, ...t_P$[3] let $f_k(X_k(t_i), \theta_k)$ represent the transformation from system variables to decision relevant variables at a particular time t_i, where θ_k refers to a set of exogenous parameters of the function f_k. Suppose the decision maker (e.g. a risk manager considering the purchase of insurance) is interested in multiple variables indexed by $H \subset \{1, 2, ...K\}$). Let a function of system variables be defined as follows:

$$R(X_k|k \in H) := \sum_{i=1}^{P} \sum_{k \in H} f_k(X_k(t_i), \theta_k) \tag{4.1}$$

Then, R is the decision relevant variable created from transforming various selected system variables measured at a number of stated times. In the example, R represents the revenue lost due to extreme temperatures at specified locations during the summer.

[3] As with Chapter 3 a sequence of times $1, ..P$ will be called a 'period'; which would typically be a year for an insurance contract but, in theory, be any chosen length of time.

Table 4.1: Terminology for insurance index showing general terms and a concrete example

General concept	Index insurance example
Problem	On hot days people may not attend outside venues leading to loss of ticket sales and other related revenue
Decision	Whether to purchase insurance against 'hot days' for a stadium operator on critical business days
System	Atmosphere
System variable	Temperature
Decision relevant variable	Reduction in ticket sales (and other sources of revenue) caused by non-attendance due to hot days
$f_k(X_k(t_i), \theta_k)$	Loss of revenue at a given venue and particular day
Period $t_1, ...t_P$	specified times (weekend dates) over the summer
$H = \{k_1, k_2, ..., k_n\}$	Locations of chosen venues in Florida
$R(X_k \mid k \in H)$	Total implied revenue impacts at specified locations over the summer
Model	WRF (NCAR)

4.2 Relating the system and its models

Model dynamics may not match the system precisely as shown in Chapter 3 where, for example, the model PDFs were not identical to those of the system. The model may, however, still have useful information about the system. This section considers a method whereby the model and system values can be related via a 1-1 relationship and demonstrates a method to estimate this using skill scores. To emphasise the difference between model and system the notation $\hat{X}_k$ will be used to denote the model variables, as distinct from system variables X_k.

Let Ω be the range of the function R (equation 4.1) applied to the system. Choose a partition[4] of intervals $A_j = (a_{j-1}, a_j)$. Define the indicator function $I(A_j, X, H)$ as:

$$I(A_j, X, H) = \begin{cases} 1 & R(X_k \mid k \in H) \in (a_{j-1}, a_j) \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

As a slight abuse of notation A_j will also be described as an 'event' which is deemed to have occurred if $I(A_j, X, H) = 1$. $I(A_j, \hat{X}, H)$ is defined similarly for model variables.

A simple approach to predicting whether the index defined in 4.1 will fall within a prescribed range (A_j) would be to see whether the model outputs fall within this range too, i.e. to look for cases when $I(A_i, \hat{X}, H) = 1$. This implies, however, that the thresholds in the system retain their meaning in the model. Chapter 3 showed

[4] A partition $\{A_j\}_{j=1}^{M}$ of a set Ω satisfies the following conditions: (1) $\Omega = \bigcup_{j=1}^{M} A_j$ and (2) $A_j \cap A_i = \phi \; \forall i \neq j$

that this is not the case in the model for Lorenz 96 where, for example, the 90th percentile from the model with forcing parameter $F = 10$ was 13% higher than that of the system. Note also that the 60th percentile was only 6.5% higher than the system so the relationship can vary.

A more general approach is to consider a possibly different interval ($B_j = (b_{j-1}, b_j)$) defined to have occurred when $I(B_j, \hat{X}, H) = 1$ and to test whether the occurrence of B_j in the model is predictive of the occurrence of A_j in the system. There is no a priori reason why $A_j = B_j$, other intervals may give better predictability and this can be tested. Specifically, choose another partition of intervals $\{B_i\}_{i=1}^{M}$ which covers the range of R applied to the model outputs (call this Θ , so that $\Theta = \bigcup_{i=1}^{M} B_i$). Note that M is the same in both partitions and implicitly defines a set function $\phi(A_i) := B_i$. The aim of this section is to find a mapping ϕ so that when B_i occurs in the model there is a good chance A_i will occur in the system. In creating these partitions the system and model spaces have been discretised such that there are M points in each space.

Given an ensemble of forecasts with N_{ens} members, labelled $\hat{X}_k^1, ...\hat{X}_k^{N_{ens}}$ (for a given k) define the relative frequency (λ) of hitting an interval B_i as follows:

$$\lambda(B_i, \hat{X}_k) := \frac{\sum_{g=1}^{N_{ens}} I(B_i, \hat{X}_k^g)}{N_{ens}}. \tag{4.3}$$

The model ensemble implies a relative frequency that event B_i occurs by counting the number of ensemble members that give values falling in that interval as described above. Since, by construction, ϕ is 1:1 and monotonic this also suggests a categorical probabilistic forecast p for the event A_i. Specifically,

$$p(A = A_i) := \lambda(B_i, \hat{X}_k.) \tag{4.4}$$

Observations determine which event A_i occurs; the indicator functions defined above will be zero in all cases apart from the one that occurs. For a given forecast and observation, the skill score can be calculated using any score appropriate for categorical forecasts. For example, if the Ignorance is used the score is calculated as follows.

$$S(p, A_i) = -\sum_i I(A_i, X_k) \log_2(\lambda(B_i, \hat{X}_k)). \tag{4.5}$$

Note that a partition $\{B_i\}_{i=1}^{M}$ is equivalent to an $M + 1$ dimensional vector subject to the constraint that $b_{j+1} > b_j$. Hence using a constrained multivariate

optimisation routine (such as the Nelder Mead algorithm) a partition $\{B_i\}_{i=1}^M$ can be found to optimise the average score S for a given partition $\{A_i\}_{i=1}^M$. This produces a piecewise linear non-decreasing function ϕ which relates the model and system, as sought. The above concept is illustrated in figure 4.1. Call this the **'Score Optimal Piecewise Linear Relationship (SOPLR)'** .

$$\phi = \text{argmin}(b_0, ...b_M)S(p, v); b_i < b_j \ \forall i, j. \tag{4.6}$$

Generalisations In the above definition the relative frequency is interpreted as a forecast probability. In Chapter 3, climatology blending was used to produce a continuous forecast with a minimal empirical score. Such a forecast defines a predicted probability density (p) for any value of the system. There are other methods[5] to produce a probabilistic forecast p and given any such a forecast the model probability of falling within the chosen interval can be calculated, in which case the relative frequency in equation 4.5 can be replaced with the value below:

$$\lambda'(b_1, b_2, p) := \int_{b_1}^{b_2} p(x)dx. \tag{4.7}$$

This is a generalisation of the relative frequency defined above. In fact let p be a sum of delta functions at the predicted values $\hat{X}_k^1, ...\hat{X}_k^{N_{ens}}$, each with integral $\frac{1}{N_{ens}}$, then the two definitions are equivalent.

Given any non-decreasing function ϕ, the forecast probability density $p_A(a)$ of the system value a being observed can be defined by the probability density p_B in the model as:

$$p_A(a) := p_B(\phi(a))|\frac{d}{da}\phi(a)| \tag{4.8}$$

As before a score can then be evaluated $S(p_A, a)$. With this definition one would then seek to optimise the score over all allowable functions ϕ. In general the data is not available to achieve this in practice. The piecewise linear method described above will be used for the rest of this chapter as it is sufficient for our purposes given the imperfections in the model.

[5]For example the Kernel Dressing or Climatology Blending of Chapter 2.

Summary of coming sections Sections 4.3 and 4.4 describe a series of experiments to test the ϕ-transformation method in idealised conditions. Section 4.3 describes a situation where the sample spaces of the model and system are naturally discretised into partitions. The derived SOPLR ϕ closely matches the true relationship ϕ' in two tests where the forecasts are of high and low quality respectively. The true relationship ϕ' is almost exactly recovered in a perfect partition example ($\{A_j\} = \{A'_j\}$). Section 4.4 tests a less idealised situation where the relationship ϕ' is defined by a monotonic continuous function. Again the ϕ-transformation method recovers a discrete approximation to the true relationship. The same partition is considered in two cases with different probability densities, this shows (as expected) that the ϕ-transformation method achieves a closer approximation where the data is plentiful. The overall conclusion is that the method is successful. The reader can skip to section 4.5 where the method is tested in an insurance context, unless they wish to see the full details.

4.3 Example C4.1: Naturally discretisable

Consider the situation where the system and model spaces are already partitioned. Observations are generated along with an ensemble of model outputs that have a built-in piecewise linear relationship ϕ'. Deliberate model failures are incorporated, the frequency and magnitude of which can be controlled to make the quality of the forecast as good or bad as required. The Score Optimal Piecewise Linear Relationship method will then be used to generate ϕ (equation 4.6) to test whether this is close to the true relationship ϕ'.

System and model partitions The following algorithm defines a process to create a series of observations and a corresponding ensemble of 'model' outputs from which a forecast can be created. The algorithm allows the quality of the forecast to be controlled. Let $\Omega' = (a'_0, a'_{M'})$ be an interval which defines the sample space of the System and $\{A'_j\}$ be a partition of intervals, $A'_j = (a'_{j-1}, a'_j)$, such that $\Omega' = \bigcup_{j=1}^{M'} A'_i$. A'_1 and $A'_{M'}$ are referred to below as **'boundary'** intervals and A'_j, $1 < j < M'$ are called **'Interior intervals'**. Note that this partition will be used to *define* the system, it is not the same as as the partition $\{A_j\}$ used to define

the mapping ϕ (equation 4.6). Let $\Theta' = (b'_0, b'_{M'})$ be an interval which defines the range of the model and $\{B'_j\}$ be a partition of intervals, $B'_j = (b'_{j-1}, b'_j)$, such that $\Theta' = \bigcup_{j=1}^{M'} B'_j$.

The following algorithm (see also figure 4.2) results in N_{obs} observations $v_1, ...v_{N_{obs}}$. For each observation v_i an ensemble of N_{ens} model outputs $w_{i,1}, ...w_{i,N_{ens}}$ are created. Let $p_1, p_2, p_3 \geq 0$ be such that $\sum p_i = 1$, these will be the proportions of 'Good', 'Near Miss' and 'Uninformative' ensemble members, respectively. Let $u \sim U(x, y)$ denote a sample from a uniform random variable with lower limit x and upper limit y, the character Q is used below to denote a random real number between zero and 1 (i.e. $Q \sim U(0, 1)$).

Algorithm for examples C4.1.x, for $x \in 1, 2$

- For $i = 1, 2, ...N_{obs}$
- Sample $j \in \{1, ...M'\}$;
- Define Observation: Sample $v_i \sim U(a'_{j-1}, a'_j)$.
- Create Ensemble: For each integer $g = 1, ...N_{ens}$ Choose from three options according to specified probabilities:
  - If $Q_1 < p_1$: 'High quality ensemble member' sample $w_{i,g} \sim U(b'_{j-1}, b_j)$;
  - If $p_1 \leq Q_1 < p_1 + p_2$: 'Near miss ensemble member'
    - If $1 < j < M'$ (observation is from an interior interval; then choose ensemble member from one of the intervals either side as follows)
      - if $Q_2 \leq 0.5$: $w_{i,g} \sim U(b'_{j-2}, b'_{j-1})$;
      - if $Q_2 > 0.5$: $w_{i,g} \sim U(b'_j, b'_{j+1})$;
    - Else if $j = 1$ sample $w_{i,g} \sim U(b'_1, b'_2)$ (observation is from left boundary, choose ensemble member from adjacent interval on right)
    - Else if $j = M'$ sample $w_{i,g} \sim U(b'_{M'-2}, b'_{M'-1})$ (observation is from right boundary, choose ensemble member from adjacent interval on left)
  - If $Q_1 > p_1 + p_2$: 'Uninformative ensemble member' sample an integer $k \in \{1, ...M'\}$, then sample $w_{i,g} \sim U(b'_{k-1}, b'_k)$;
  - Repeat to create N_{ens} ensemble outputs for given observation v_i

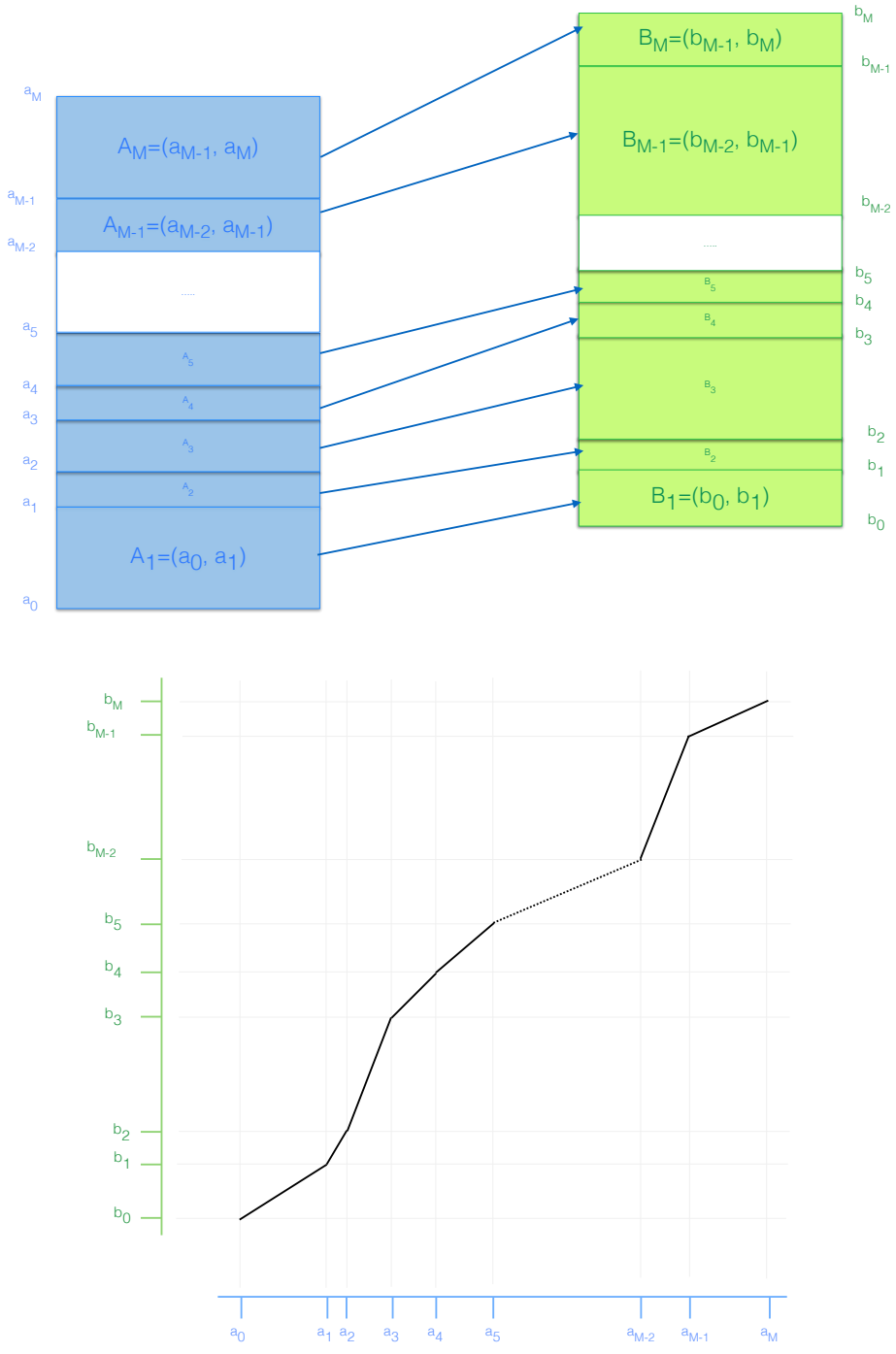


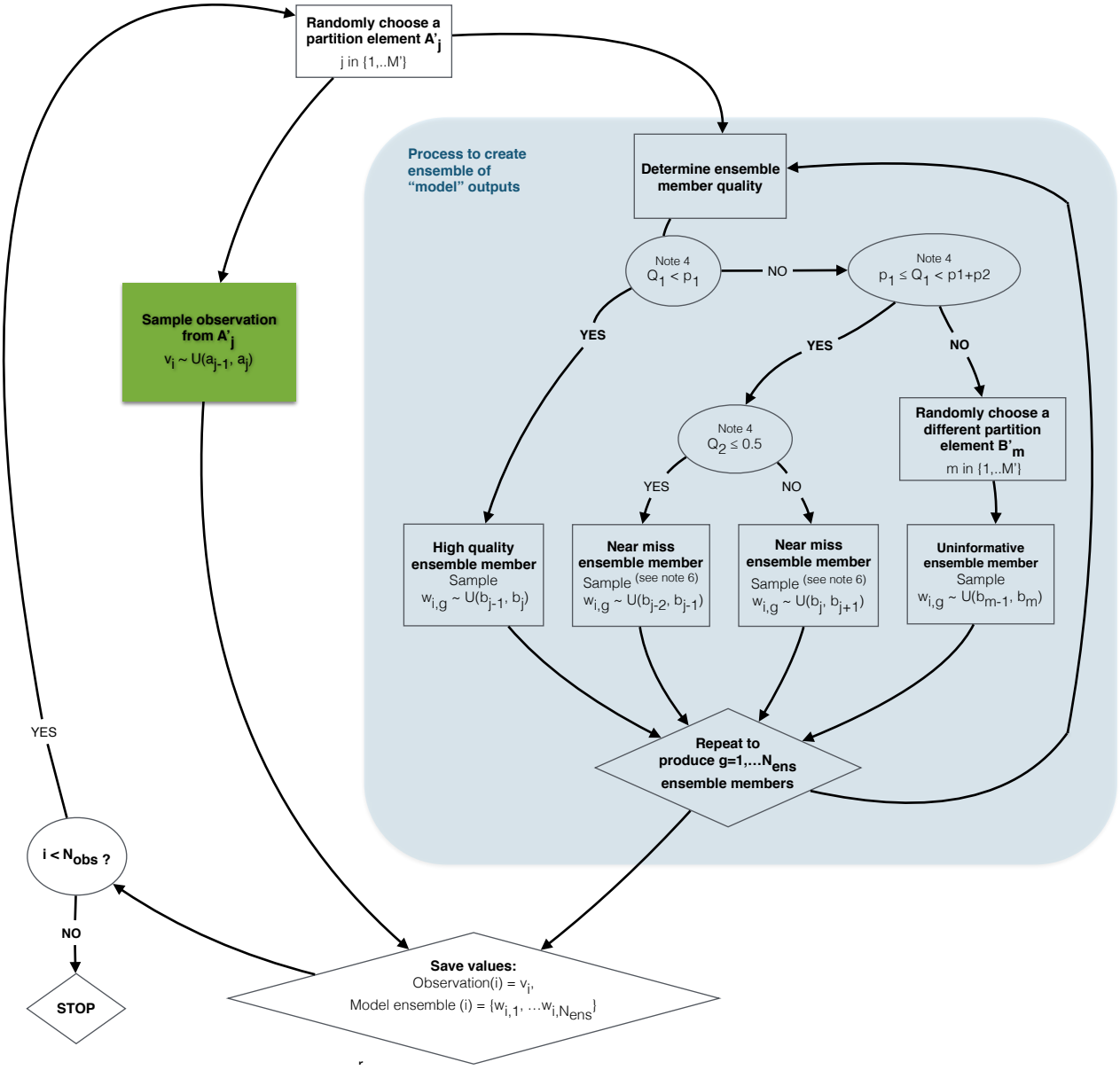
Figure 4.1: Discretisation of the system and model sample spaces, with resulting relationship ϕ. The top graphic shows the sample space of the system (the interval $\Sigma = (a_0, a_M)$ partitioned into sub intervals $A_1, A_2, ...A_M$, these are mapped in 1:1 correspondence to M intervals in the model sample space $\Theta = (b_0, b_M)$ partitioned by $B_1, ...B_M$. This can be represented by a piecewise linear, non-decreasing relationship $\phi : \bigcup_{i=1}^{M} A_i \rightarrow \bigcup_{i=1}^{M} B_i$ as shown in the bottom plot.

Using the above process, the quality of the forecast can be controlled. If the proportion of uninformative ensemble members (p_3) is high, the average quality of the resulting forecast will be low. If $v \in A'_j$ then define

$$\phi'(v) := mv + c \tag{4.9}$$

where $m = \dfrac{\max(B'_j) - \min(B'_j)}{\max(A'_j) - \min(A'_j)}$ and $c = \min(B'_j) - m.\min(A'_j)$. This is illustrated for a single ensemble member by the blue line in figure 4.4(a) (this figure is an output from experiment C4.1.1 defined on page 209). Note that, by construction, the model is not in 1-1 correspondence with the system. The diagonal line in figure 4.3(b) is the line $y = x$ and it is clear the ensemble values do not fall around this line, they are typically higher initially so that model values do not retain their meaning in the system. Nevertheless the model outputs and system are closely related and hence the forecast should have considerable predictive power. Note that since the intervals A'_j are sampled in an equally likely way the probability density of the observations can be controlled by the width of the intervals, figure 4.3(a) shows a histogram of observations for one example.

In order to calculate the Score Optimal Piecewise Linear Relationship (ϕ) it is necessary to specify a partition $\{A_j\}$. Recall, however, that the partitions A'_j generating the observations of the system or forecasts are assumed to be unknown. In the absence of any other information it is appropriate to choose a system partition where each interval has the same number of observations, referred to below as **'equal cardinality bins'**. Then the regions that are dense with observations will have a finer grid than those with sparse observations; this allows for a more granular relationship ϕ in those regions. The equal cardinality bin method is used below unless stated otherwise.



Notes:

1. System state space $\Omega' = U_{j=1,M'} A'_j$ where, $A'_j = (a'_{j-1}, a'_j)$
2. Model state space $\Theta' = U_{j=1,M'} B'_j$ where, $B'_j = (b'_{j-1}, b'_j)$
3. w ~ U(a,b) means sample from a uniform distribution with minimum value a and maximum b.
4. Q ~ U(0,1) is a random number between 0 and 1. In the circles, the notation $Q_1 < p_1$ (for example) means sample a value Q_1 and determine whether it is less than p_1
5. p_1 = P(good quality ensemble member), p_2 = P(near miss ensemble member), p_3 = P(uninformative ensemble member)
6. If j = 1, then $w_{i,g} \sim U(a'_1, a'_2)$; If j=M, then $w_{i,g} \sim U(a'_{M'-2}, a'_{M'-1})$

Figure 4.2: Example C4.1.1: Flowchart describing the observation and forecast ensemble creation process.

Experiment C4.1 specifications The following experiments test the ϕ-transformation method in various idealised examples: (1) C4.1.1(a) has forecasts of high average quality, (2) C4.1.1(b) forecasts of low quality; (3) C4.1.1(c) uses fewer equal cardinality bins; (4) C4.1.2 uses a perfect partition in the calculation of ϕ, to test whether the method recovers the true relationship ϕ'.

Example C4.1.1(a) - Forecasts of high average quality.

System and model parameters: $N_{obs} = 2^{12}$, $M' = 14$, $\Omega' = [0, 49]$, $\Theta' = [3, 49]$, $N_{ens} = 2^4$, $p_1 = 0.80$, $p_2 = 0.15$ and $p_3 = 0.05$. A'_j, B'_j are defined in table 4.2.
SOPLR parameters: M=7, the equal cardinality partition $\Omega = \bigcup_{j=1}^{M} A_j$ for ϕ is defined in table 4.3.

Table 4.2: Experiment C4.1.1(a) - definition of underlying system and model partitions

j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
a'_j	0	1	3	4	9	15	16	19	25	36	38	39	42	47	49
b'_j	3	6	9	12	17	19	25	30	32	33	34	37	42	45	49

Table 4.3: Experiment C4.1.1(a) - assumed, equal cardinality, partition for ϕ

j	0	1	2	3	4	5	6	7
a_j	0.00	2.84	8.34	15.86	24.05	37.69	41.87	48.99

Example C4.1.1(b) - Forecast with low average quality.

As for C4.1.1(a) but with $p_1 = 0.25$, $p_2 = 0.25$ and $p_3 = 0.5$.

Example C4.1.1(c) - smaller partition size for ϕ.

As for C4.1.1(a) but with the partition $\{A_j\}$ chosen to be comprised of four equal cardinality bins only, specified by the values in table 4.4

Table 4.4: Experiment C4.1.1(a) - assumed partition for ϕ

j	0	1	2	3	4
a_j	0	6.13	18.41	38.32	48.99

Example C4.1.2 Perfect system partition.

System and model parameters: $N_{obs} = 2^{12}$, $M' = 8$, $\Omega' = [0, 25]$, $\Theta' = [3, 32]$, $N_{ens} = 2^4$, $p_1 = 0.80$, $p_2 = 0.15$ and $p_3 = 0.05$ (forecasts with high average quality). A'_j, B'_j are defined in table 4.5.

SOPLR parameters: $M = 8$, the equal cardinality partition $\Omega = \bigcup_{j=1}^{M} A_j$ for ϕ is set equal to the true system partition (i.e. $a_j = a'_j$, $\forall j$).

Table 4.5: Experiment C4.1.2 - definition of underlying system and model partitions

j	0	1	2	3	4	5	6	7	8
a'_j	0	1	3	4	9	15	16	19	25
b'_j	3	6	9	12	17	19	25	30	32

Results for C4.1.1(a) This experiment shows that the SOPLR produces an appropriate discrete approximation to the true relationship; the method is successful. Figure 4.3(a) shows a histogram of observations with some areas of high and low density. $M = 7$ equal cardinality bins are illustrated by the blue tick marks, the width of the bin follows the density of observations as required. Figure 4.3(b) shows the observed values versus the model values (for one ensemble member only). The line $y = x$ is included to illustrate the high bias of model versus system values less than 30. Figure 4.4(a) shows the initial values (b'_j) chosen for the optimisation routine, these are just evenly spaced points within Θ illustrated by grey horizonal lines. The intersection of these with the chosen equal cardinality bins (blue tick marks and grey vertical lines) defines the initial trial relationship ϕ_{trial} used in the optimisation routine. Figure 4.4(b) shows the results after the optimisation routine. The blue line represents the true relationship ϕ', the black line represents the Score Optimal Piecewise Linear Relationship ϕ and the grey line ϕ_{trial}. Clearly ϕ is closer to ϕ' than ϕ_{trial} suggesting success of the method.

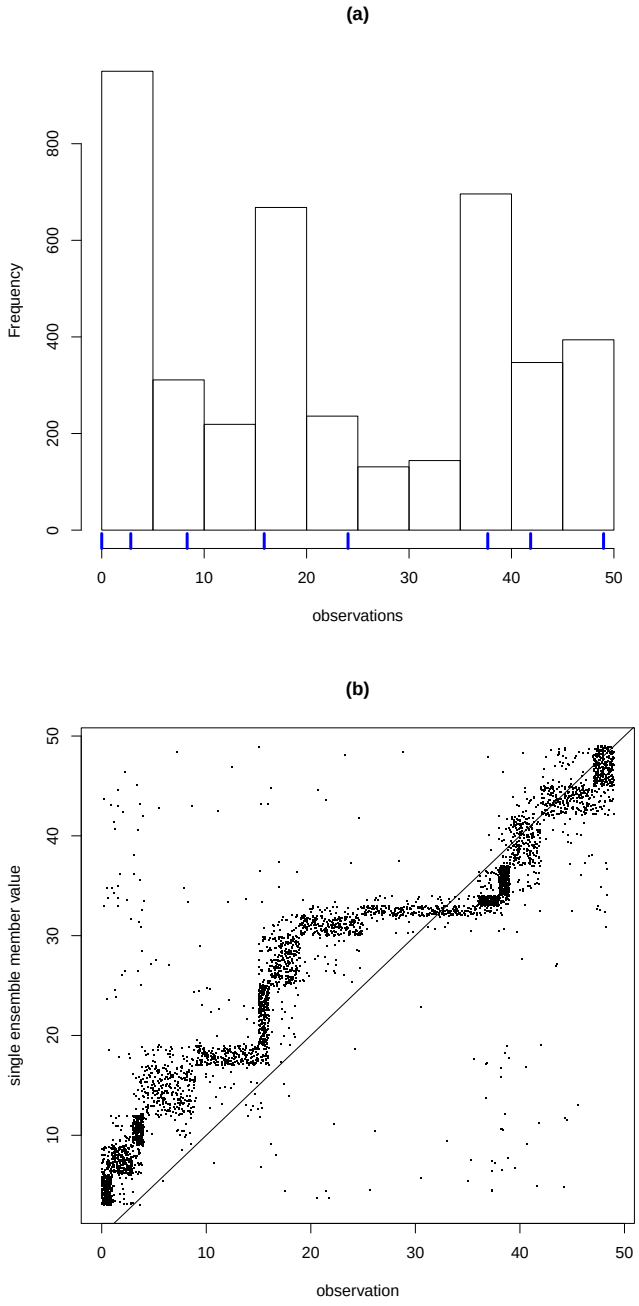


Figure 4.3: Example C4.1.1: Figure (a) Histogram of pseudo observations of the system. End points of equal cardinality bins indicated by the blue tick marks), Figure (b) Scatter plot of pseudo observations of the system (x-axis) with corresponding forecast values (y-axis) for one ensemble member. The line $y = x$ reveals that the model tends to predict values above the system in some regions of the distribution.

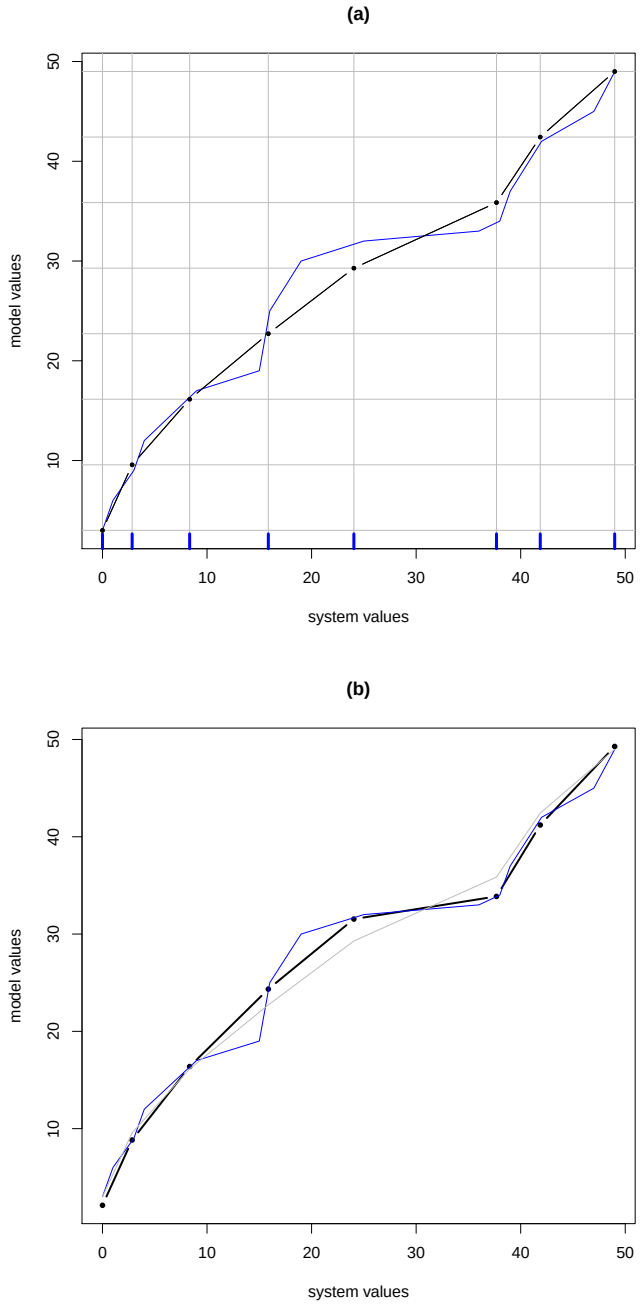


Figure 4.4: Example C4.1.1: Figure (a) Blue line shows the true relationship ϕ between the system and the model. The blue tick marks show an equal cardinality partition of the system sample space. The model space is subdivided into equal length intervals and the black line shows a line drawn between the points the form the intersection between the interval end points and the equal cardinality partition - this is the relationship that forms the initialisation of the optimisation routine. Figure (b) the blue line shows the true relationship ϕ' and the black line shows the result of the optimisation routine (i.e. the estimator ϕ) the SOPLR is closer to the true relationship in figure (b) than in the initial partition (a).

Results for C4.1.1(b) This experiment shows that the method remains successful when the forecast quality is lower. Figure 4.5(a) illustrates the results of an ensemble member of lower average quality compared to the corresponding observations. The scatter plot has points filling much of the space $\Omega \times \Theta$ due to the chosen lower ensemble member quality ($p_3 = 0.5$). As before, figure 4.5(b) shows ϕ' in blue, ϕ_{trial} in grey. The SOPLR ϕ in black remains close to ϕ' suggesting that the method continues to work when forecast quality is lower.

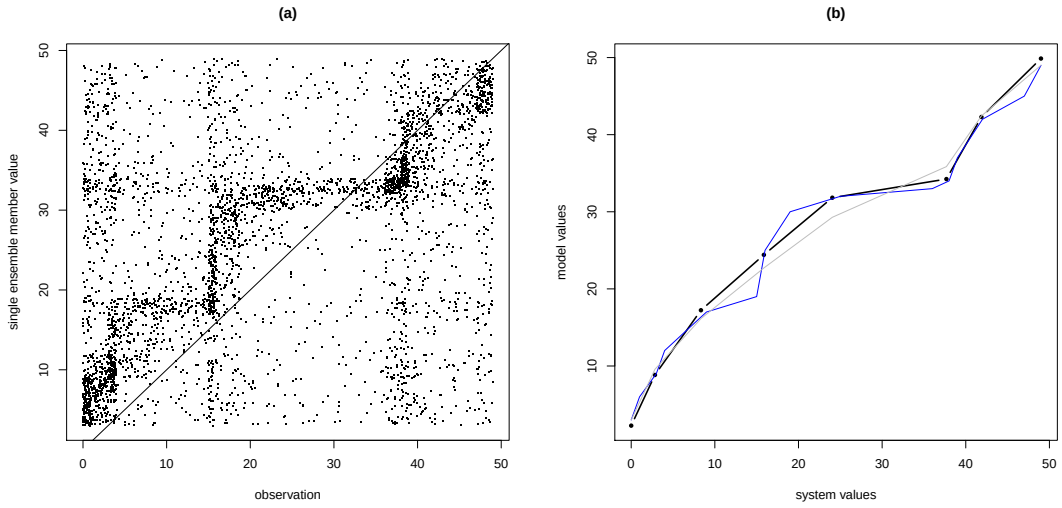


Figure 4.5: Example C4.1.1: Left plot show a scatter plot of observations of the system versus one forecast ensemble member. This is a lower quality forecast than in graphic 4.3 as evidenced by the greater scatter of points. The right plot shows the true relationship ϕ (blue line) and the estimator in black. Despite the poor quality of the forecast the estimator closely aligns with the true relationship.

Results for example C4.1.1(c) Figure 4.6 shows that the method has again been successful in approximating the true relationship ϕ' although, arguably the fit from ϕ at the end points is worse than ϕ_{trial}.

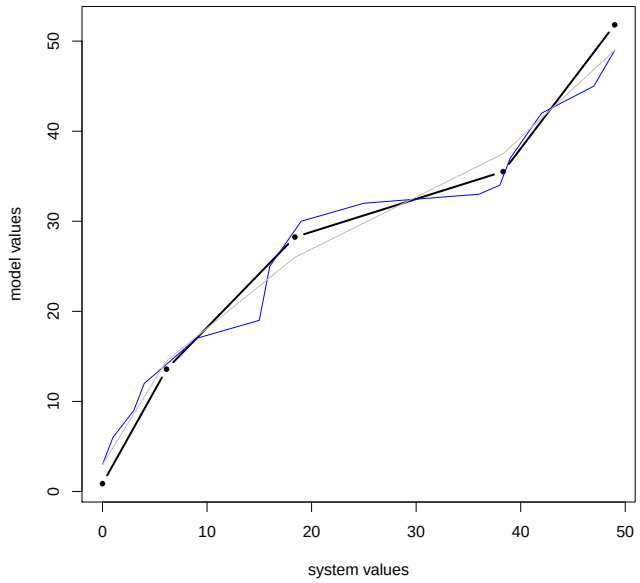


Figure 4.6: Example C4.1.2: Coarse partition of the system space with just 4 equal cardinality bins. The true relationship (blue line) and estimator (black line) are close together indicating success of the method despite the coarse partition of the model space. The initial trial relationship ϕ_{trial} is shown in grey.

Results for example C4.1.2: Perfect system partition This experiment shows that the true relationship is (almost) recovered when the system partition is perfect. The system partition used in the optimisation is set to be exactly the same as the partition used to generate the observations and ensemble values, even though this would not normally be known in practice. This is to test whether the SOPLR approach finds the true ϕ'. Figure 4.7(d) shows that ϕ is not exactly the same as ϕ' for the last interval but agrees everywhere else. The error occurs in a region of lower observation density (see histogram 4.7(a)).

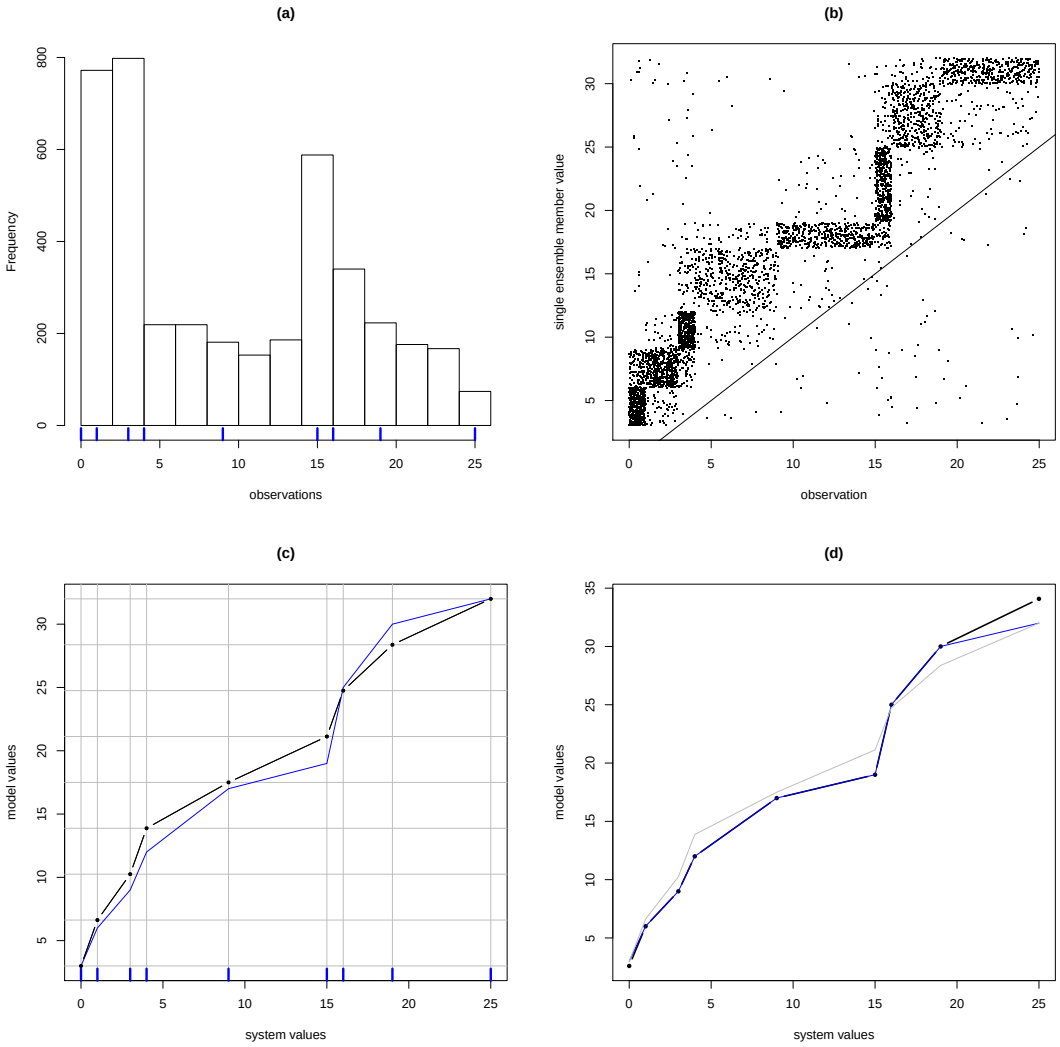


Figure 4.7: Example C4.1.3: Perfect system partition. The optimisation routine doesn't quite find the true ϕ'. The figure types follow those already described in this section. Figure (a) a histogram of pseudo observations of the system. Figure (b) a scatter plot of system versus observed values. Figure (c) the initial estimate for the optimisation routine. Figure (d) the result of the optimisation (ϕ, black line) versus the true relationship (ϕ' blue line) versus the initial trial (ϕ_{trial} grey line).

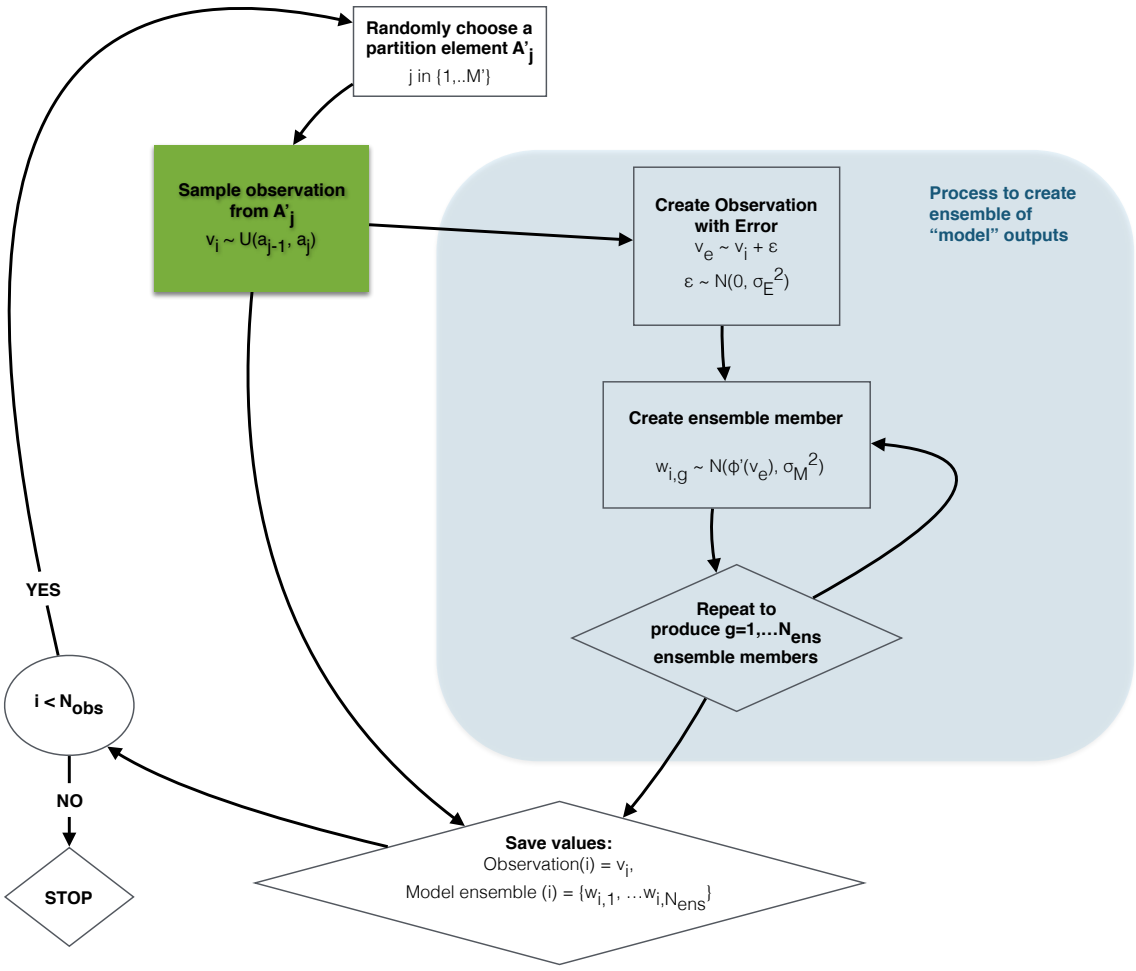
4.4 Example C4.2 - continuous relationship

In this example the system values are sampled as before - but ensemble values are then created directly from the observations in a way that mimics initial condition uncertainty as follows:

System and model partitions Choose system and model partitions as in example C4.2.1. Define the true relationship ϕ' as in equation 4.9. The following algorithm (see also figure 4.8) results in N_{obs} observations $v_1, ...v_{N_{obs}}$. For each observation v_i an ensemble of N_{ens} model outputs $w_{i,1}, ...w_{i,N_{ens}}$ are created. Let $x \sim N(\mu, \sigma^2)$ denote a sample from a Gaussian distribution with mean μ and variance σ^2. Let σ_E be the standard deviation of observational error. Let σ_M be the standard deviation of ensemble member outputs.

Algorithm for examples C4.2.x, for $x \in 1, 2$

- For $i = 1, 2, ...N_{obs}$;
- Sample $j \in \{1, ...M'\}$;
- Define Observation: Sample $v_i \sim U(a'_{j-1}, a'_j)$;
- Define an observation-with-error: $v_e = v_i + \epsilon$ where $\epsilon \sim N(0, \sigma_E)$;
- For $g = 1, ...N_{ens}$ define an ensemble of model values as $w_{i,g} \sim N(\phi'(v_e), \sigma_M)$;
- Repeat to produce N_{obs} observations, each with associated model outputs.



Notes:

1. System state space $\Omega' = U_{j=1,M'} A'_j$ where, $A'_j = (a'_{j-1}, a'_j)$
2. Model state space $\Theta' = U_{j=1,M'} B'_j$ where, $B'_j = (b'_{j-1}, b'_j)$
3. If $a_{j-1} < v < a'_j$, then $\phi'(v) = mv + c$ where $m = (b'_j - b'_{j-1}) / (a'_j - a'_{j-1})$ and $c = b'_{j-1} - m\, a'_{j-1}$
4. $w \sim U(a,b)$ means sample from a uniform distribution with minimum value a and maximum b.
5. $x \sim N(\mu,\sigma^2)$ means sample from a Gaussian distribution with mean μ and variance σ^2

Figure 4.8: Example C4.2.1: Flowchart describing the observation and forecast ensemble creation process.

Example C4.2.1(a) ϕ Continuous - equal cardinality bins

System and model parameters: $N_{obs} = 2^8$, $M' = 7$, $\Omega' = [0, 100]$, A'_j is defined in table 4.6. ϕ' is a continuous relationship illustrated by the blue line in figure 4.9(b) and defined by choosing 50 values of x and $y = \phi'(x)$ which are listed in full in appendix B.

SOPLR parameters: M=7, the equal cardinality partition $\Omega = \bigcup_{j=1}^{M} A_j$ for ϕ is defined in table 4.7.

Table 4.6: Experiment C4.2.1 - definition of underlying system partition

j	0	1	2	3	4	5	6	7
a'_j	0	10	20	40	80	90	95	100

Table 4.7: Experiment C4.2.1 - assumed, initial equal cardinality, partition for ϕ

j	0	1	2	3	4	5	6	7
a_j	0.15	6.92	20.1	41.54	81.85	90.73	94.52	99.97

Example C4.2.1(b) ϕ Continuous - with additional end points

System and model parameters: As for 4.2.1(a)

SOPLR parameters: M=9, the partition $\Omega = \bigcup_{j=1}^{M} A_j$ for ϕ is defined in table 4.8 and is created by adjoining two additional points at the ends of the partition.

Table 4.8: Experiment C4.2.1 - assumed, initial equal cardinality, partition for ϕ plus end points

j	0	1	2	3	4	5	6	7	8	9
a_j	0.15	1	6.92	20.1	41.54	81.85	90.73	94.52	98.97	99.98

Example C4.2.2 - Different observation density.

System and model parameters: $N_{obs} = 2^8$, $M' = 7$, $\Omega' = [0, 100]$, A'_j is different to 4.2.1 and is defined in table 4.9. ϕ' is as for example 4.2.1.

SOPLR parameters: M=7, the equal cardinality partition $\Omega = \bigcup_{j=1}^{M} A_j$ for ϕ is defined in table 4.10. Note due to the different system definition the equal cardinality bins are also different, end points have also been added as in 4.2.1(b)).

Table 4.9: Experiment C4.2.2 - definition of underlying system partition

j	0	1	2	3	4	5	6	7
a'_j	0	30	40	45	50	60	80	100

Table 4.10: Experiment C4.2.2 - assumed, initial equal cardinality, partition for ϕ plus end points

j	0	1	2	3	4	5	6	7	8	9
a_j	0.5	1	32.61	41	44.18	49.26	59.81	79.97	98.81	99.81

Results for example C4.2.1(a) and (b) This experiment shows that the ϕ-transformation method continues to work in a less idealised setting and that end-point corrections produce a better fit at the extremes of the data. Figure 4.9(a) shows a histogram of observations which illustrates the chosen, equal cardinality partition with blue tick marks. Figure 4.9(b) plot shows the observations (x-axis) against model ensemble values on the y-axis (black dots) and also shows the true relationship ϕ' as a blue line. Figure 4.10(a) shows the resulting SOPLR for example 4.2.1(a). The optimisation routine chooses end points that are considerably above the true relationship (blue line) at the upper end point. By adding additional points at either end the fit is better as shown in figure 4.10(b) which contains the results of example 4.2.2(b). This end-point correction method is used in the Lorenz 96 example later in the chapter and also in example 4.2.2.

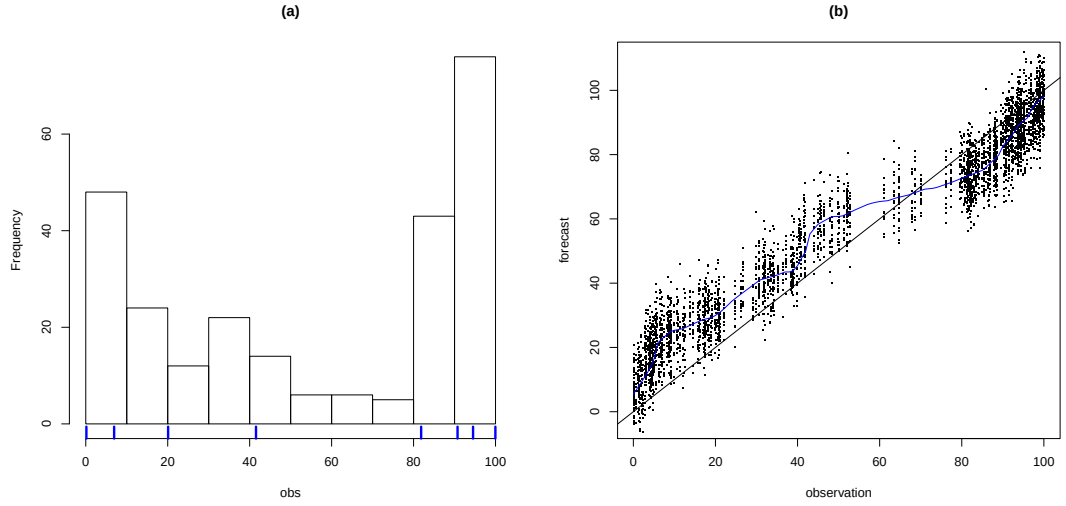


Figure 4.9: Example C4.2.1: Left plot, a histogram of observations of the system with equal cardinality bins shown by the blue tick marks. Right hand plot shows observations plotted against forecast ensemble values (all ensemble members)

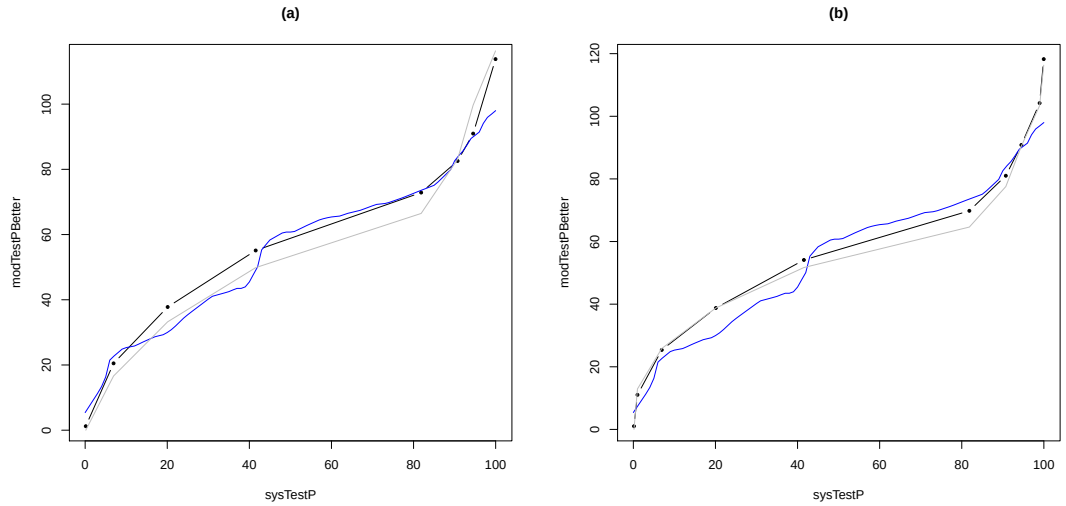


Figure 4.10: Example C4.2.1: Figure (a) shows the results of the optimisation routine for equal cardinality bins. Figure (b) shows the result using the same partition but with two additional points added close to the end points of the first and last intervals - this helps to suppress the overshoot at the end points.

Results for example C4.2.2 The experiment shows that the SOPLR will tend to have a better fit where the density of observations is higher. Example C4.2.2 is a variation where the relationship ϕ' is the same, but where the probability density of different observation values within the support Ω is different (as shown by the histogram of observation values in figure Figure 4.11(a)). 4.11(b) shows ensemble outputs against the observations. Figure 4.11(c) shows the the SOPLR where the fit is better in the centre of the curve because that is where the highest observation density occurs. As noted above this may be undesirable if extreme values are important to the user. Note the use of the Ignorance score will mitigate this partially because it will score extremes very badly and will tend to reward ϕ relationships that mitigate this as far as possible. As noted in Chapter 2, this feature of the Ignorance score is sometimes criticised (e.g. Selten [224]) but in this setting it is a benefit.

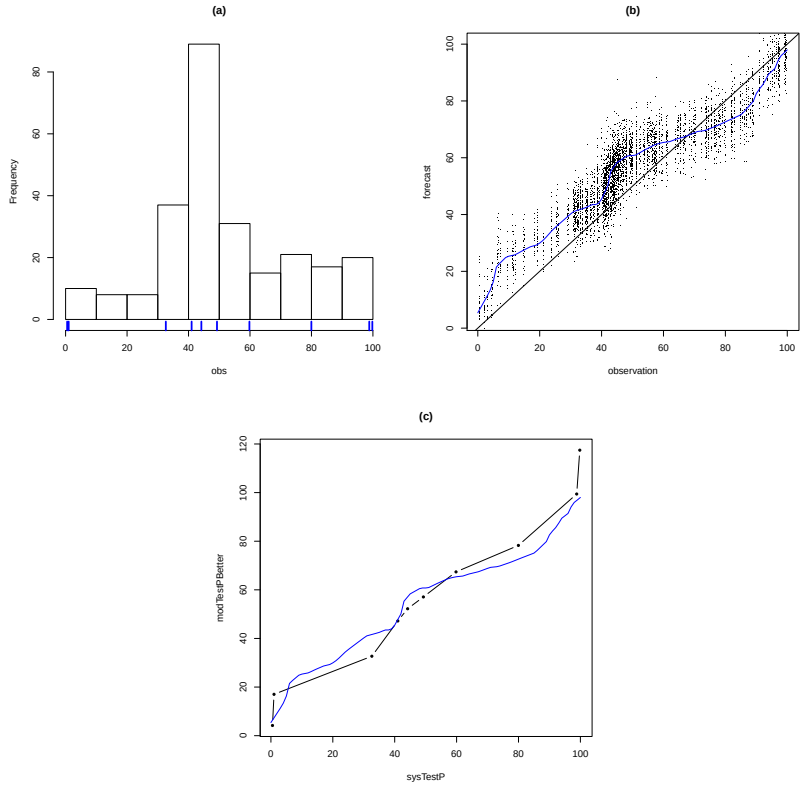


Figure 4.11: Example C4.2.2: Figure (a) shows histogram of observations against equal cardinality bins (blue tick marks), note that the density is now highest in the centre of the distribution. Figure (b) shows the true relationship ϕ' as a blue line with a scatter plot of ensemble values versus system values. Figure (c) shows the true relationship in blue and shows the SOPLR ϕ in black. The fit is now closest in the centre where the density of observations is highest.

4.5 Translations to the challenge of insurance

Section 4.1 described the decision relevant function R which transforms system variables into more relevant outputs. The values of R over time can be thought of as an 'index'. In theory an insurance product could be designed to give coverage that is triggered by an index from any system, provided the purchaser can demonstrate they would suffer loss [53], otherwise it counts as gambling [258, 259] or investing [215]. Such products are now offered in practice, such as the El-Nino index insurance offered by Kalista global [35] which pays out if the index passes a threshold value, or the weather index products of Nephila Capital [185]. In the following example the index is defined by equation 4.1; other index definitions have been used in practice [91, 241, 247].

Climatology pricing The traditional approach to pricing insurance uses the past claims history and uses the statistics of this data to derive the price [52]. For example a standard distribution might be fitted to the data and the premium calculated using the expected index value ($E(R)$) and standard deviation, or an explicit return on capital. In the language of forecasting such pricing methods use the 'climatology' of the insurance losses. In later sections this will therefore be referred to as **'climatology pricing'** . A simple pricing formula for a premium rate of W_{clim} which targets a return (γ) on capital (κ) is as follows. This is broadly based on the method defined by Kreps [128] which is discussed in more detail in Chapter 5:

$$W_{clim} = E(R) + \gamma\kappa(R) \tag{4.10}$$

In words W_{clim} is the premium rate that will cover the expected level of claims $E(R)$ with a profit margin $\gamma\kappa(R)$ expressed as a target return on capital held, expenses are ignored.

Define the 'quantile' of a random variable Y which has a return period of τ as follows:

$$Q_\tau(Y) = \{y | P(Y > y) = \frac{1}{\tau}\} \tag{4.11}$$

The capital requirement κ is defined using the current UK insurance regulatory requirement which requires that a company be able to survive a large insurance

claim that could arise with 1 in 200 probability. The premium charged can be taken into account since this will be 'reserved' until the period of risk has expired. Therefore, in the current example the capital is defined as:

$$\kappa = Q_{200}(R) - W_{clim} \tag{4.12}$$

This may appear circular since the premium is used in the capital definition and then the capital is used to define the premium. This is not the case and the premium can be calculated directly as follows:

$$W_{clim} = \frac{E(R) + \gamma Q_{200}(R)}{1 + \gamma} \tag{4.13}$$

Note that R is the index defined earlier and has been transformed into financial terms, so although the inputs into R are physical variables (such as windspeed or temperature) its value is in appropriate currency units.

ϕ-transformed pricing Earlier sections in this chapter described a method for transforming model output to the system sample space by deriving a non-decreasing SOPLR ϕ. Assume an ensemble $\{\hat{X}_k^j\}$ is given where k denotes variables in the system as usual and j denotes an ensemble member. A 'ϕ-transformed' forecasted index value $\hat{R}^j$ can be calculated as follows:

$$\hat{R}^j = R(\hat{X}_k^j | k \in H) := \phi(\sum_{i=1}^{P} \sum_{k \in H} f_k(\hat{X}_k^j(t_i))) \tag{4.14}$$

The ϕ-transformed price is then calculated as a function of the entire ensemble (let this be denoted $\hat{R} = \{\hat{R}^1, ...\hat{R}^{N_{ens}}\}$):

$$W_\phi = E(\hat{R}) + \gamma\kappa(\hat{R}) \tag{4.15}$$

Where for example the component $E(\hat{R})$ could be taken as the average of the ensemble members, or methods such as climatology blending could be used to deriver a forecast PDF from which this component is calculated. Similarly the quantile $Q_{200}(\hat{R})$ underlying the capital requirement $\kappa(\hat{R})$ can also be calculated directly from an ensemble with sufficiently many members - or calculated from a forecast probability density function using using analytic or numerical methods. In the the following examples two uses of the ϕ-transformed price are tested: (1) Updated Expectation which uses the forecast to update the expected index value in the pricing

formula, but does not reassess the capital requirement and (2) Full Blending which uses climatology blending (Chapter 2) to create a probability forecast for the index such that both the expected claims and capital are recalculated. Defined as follows:

Updated expectation method This method assumes that the capital will not be recalculated using the forecast. Hence the only part of the price to be updated relates to the expected claims. Specifically, the expectation in the pricing formula 4.10 is replaced with the average the ensemble of forecasted index values $\hat{R}$, so that.

$$\mathrm{W}_{UE} = E(\hat{R}) + \gamma\kappa(R) \tag{4.16}$$

Where $E(\hat{R}) = \frac{\sum_{j=1}^{N_{ens}} \hat{R}^j}{N_{ens}}$ denotes the average value of the forecasted index and $\kappa(R)$ is the capital requirement for the climatology price.

Full blending method This method (illustrated in subsection 4.6.3) uses climatology blending, defined by equation 2.62 Chapter 2. Using the notation of that equation, p is the PDF created by kernel dressing the ensemble of N_{ens} ϕ-transformed index values $\hat{R}^j$ and f_u is the climatology of R. These are trained using the same past observations that were used to choose the relationship ϕ. This process produces a PDF r of index values in the coming year. The expectation in the pricing formula is replaced with the mean of the forecast distribution $E(r)$. Likewise the capital, $\kappa(r)$, is calculated by determining the 1 in 200 value of the forecast distribution allowing for premiums charged. Using this method the actual capital held is assumed to be increased or decreased according to the forecast.

$$\mathrm{W}_{Blend} = E(r) + \gamma\kappa(r) \tag{4.17}$$

In either of the above pricing methods the average premium will be of interest to policyholders - if it is higher then the insurance is less affordable in the long run. Hence this metric can be used to compare strategies.

4.6 Example C4.3 Lorenz 96

Recall Lorenz System II defined in Chapter 3, equation 3.2 and also models of this system defined by equation 3.1. The following section will use the Lorenz 96 X

variables as a concrete example of a dynamical system from which an index can be constructed. The key features of the six systems of Chapter 3 are summarised in table 4.11 below:

Table 4.11: Summary of important features of systems 80001-80006

System	Predictability	Best Model	Worst model	Median X_1	Range of X_1	Median IEF	Range IEF
80001	Baseline	10011	10010	Baseline	Baseline	Baseline	Baseline
80002	Low	10011	10010	> 80001	Twice 80001	>> 80001	> 80001
80003	Low	Little difference, 10011 just wins	Little difference	> 80001	Twice 80001	>> 80001, > 80002	Very low
80004	Medium	As 80003	As 80003	As 80001	As 80001	> 80001	Very low
80005	Medium	As 80003	As 80003	As 80001	As 80001	As 80001	Very low
80006	High	10011	10010	As 80001	As 80001	As 80004	As 80001

Notes for table 4.11

1. 'Predictability' is high if at least one model has a period average score less than -1.0, is medium if at least one model gives a period average score less than -0.7 and low otherwise;

2. 'Best' and 'Worst' models are defined by their skill scores (measured in bits of information) at each time during the period and also for the period average - since the models behave consistently under each measure there is no ambiguity here; and

3. The descriptions in this figure are intended to give an approximate relationship between systems; phrases like 'Twice 80001' are to indicate that the range of values is approximately double that of system 80001 - not to mean that they are exactly double. >> indicates 'much greater than'.

It will be demonstrated that forecasts of this system can be used to create robust insurance pricing. First, the relationship between the Lorenz system values and the Index R is defined. Parameters are chosen to 'tune' the index to be similar in each of the systems 80001, ... 80006. The SOPLR ϕ is then determined for the index. Competition effects are then discussed along with definitions of profitability and insolvency. The two forecast based pricing methods above are used and insurance results are shown in detail for one of system/model pair and then selected results from the other systems and models are discussed.

The following uses terminology common to Catastrophe models of the insurance industry. The Ground Up Loss (denoted D below) is the total damage (typically to a building) before the application of any insurance terms and conditions; the Gross Loss (denoted L below) is the damage after such conditions have been applied (see below) and is the amount payable by the insurer (often called the **'claim'** or the **'loss'**).

Definition of Ground up Loss (D) The relationship between a given hazard index and the damages caused will depend on the hazard. Often the relationship

follows some sort of power rule - for example, Emmanuel [74] notes that, hurricane winds tend to cause damage related to the cube of the windspeed. It may be that damages only start to be caused above a threshold level. With this motivation in mind, in this chapter the Ground Up Loss (D) is defined as follows:

$$D(X_{t_j,k}) := \lambda \max(X_{t_j,k} - \tau, 0)^{\rho} \tag{4.18}$$

Where: λ is a scalar that allows the losses to be tuned between systems to achieve a similar index (explained further below); τ is a threshold to ensure that events do not occur all the time; and ρ is an exponent to enable losses to created with a desired level of heavy tail.

Definition of Gross Loss (L) Various insurance terms and conditions exist in practice. It is typical, however, that the payouts only start above a certain level (the **'attachment point'**) and are often subject to an overall **'limit'** . Sometimes, there is no limit, or equivalently the limit is infinite. Motivated by this the Gross Loss (L) is defined as follows, where Δ is the attachment point and Λ is the limit.:

$$L(D) = \min(\max(D - \Delta, 0), \Lambda) \tag{4.19}$$

Definition of $f_k(X_k(t), \theta_k)$ The transformation function from section 4.1 is then defined as the combination of the D and L as follows:

$$f_k(X_k(t_j), \theta_k) = L(D(X_{t_j,k})) \tag{4.20}$$

where $\theta_k = (\lambda, \tau, \rho, \Delta, \Lambda)$ are the set of exogenous parameters which define the Ground up Loss and Gross Loss .

Definition of $R(X)$ The index function R in this concrete example is then defined using equation 4.1, where $H = \{1\}$, so this example considers an index created from just one Lorenz variable X_1, this could be extended to consider multiple Lorenz variables. In this example $P = 24$ so that observations are made at times $t_1, ...t_{24}$ and the index is the aggregate index value over these observations.

The following examples are split into two groups. In each example prices are determined using climatology pricing, updated expectation and full blending:

Examples 4.3.1.x Various models of system 80001.

These examples consider system 80001 only, and explores the ϕ-transformed pricing methods for each of the models explored in Chapter 3 (i.e. 10008, ...10012). Experiments are referred to by the last significant digits of the *model* ID because the system is fixed. **Each experiment ID in this group has the form 4.3.1.x, where** $x \in \{8, 9, 10, 11, 12\}$;

Examples 4.3.2.y All systems - with single model type.

These examples consider systems 80001, 80002, 80003, 80004, 80005 and 80006. Models 10011, 10111, 10211, 10311, 10411 and 10511, respectively are used to produce ensemble outputs for these systems. Collectively these models are referred to below as model class *11, where the * is a wildcard. **Each experiment ID in this group has the form 4.3.2.y where** $y \in \{1, 2, 3, 4, 5, 6\}$ **refers to the last significant digit of the system ID.**

In examples 4.3.2.y the other models were not considered because of time constraints; but these could be explored in future work. Model class *11 was chosen because it had the highest α values in the blending work described in Chapter 3 and the best skill scores, as such it is the forecast with the best predictability.

Tuning the index for different systems Recall from Chapter 3 that six different systems were considered each with different levels of forcing and coupling. These were chosen so that predictability varies. The range of values X_k produced by each system, however, is different. For example, when $F = 20$ the extreme X_k values are larger than when $F = 10$. As such, if the same parameters were used in each case, the index described above would be very different in the various systems. In order to compare the insurance contracts for different levels of predictability, *only*, the parameters $\theta_k = (\lambda, \tau, \rho, \Delta, \Lambda)$ need to be tuned so that a similar range of index values arise.

Define the **'quantile ratio'** (QR) as:

$$QR_{RP}(Y) = \frac{Q_{RP(Y)}}{E(Y)} \tag{4.21}$$

where $E(Y)$ denotes the expectation of Y and Q is defined by equation 4.11. This variable is of interest to insurers since it shows, for high return periods, how extreme values relate to the average outcome.

In the following examples, first ρ is, arbitrarily, set to 3 in analogy with wind speed. Then a quantile ratio of 16:1 is sought in each system, by adjusting τ (this ratio is chosen based on a study of some of the Catastrophe distributions common to insurers[6]). Next the indices are scaled using λ so they each have approximately the same mean. Finally the deductible is set to have a fixed relationship to $Q_{50}(D)$. The details of this tuning process are as follows, these are assumed to occur in the order presented. Although $H = \{1\}$ so that X_1 only is involved in the definition of R, commentary is retained for other values of k to illustrate the sampling error in the parameter choices.

Choosing ρ This parameter is set equal to 3 in all systems.

Choosing τ Table 4.12 shows the value of τ chosen by trial and error to approximately achieve the target quantile ratio of 16:1 in each system. Note that this parameter can be chosen independently from λ which will scale the mean and Q_{200} equally (and this scaling therefore cancels out in the ratio). Therefore this step can be carried out first without affecting the value of λ. Define

[6]Based on my own experience of Catastrophe Modelling.

$QR_{200,k} = QR_{200}(X_k)$ as the quantile ratio for the kth Lorenz variable. The table shows the 'mean quantile ratio' defined as:

$$\text{mean quantile ratio} = \frac{\sum_{k=1}^{36} QR_{200,k}}{36} \tag{4.22}$$

and also the range $(\min(QR_{200,k}), \max(QR_{200,k}))$, where the minimum and maximum are taken over all k variables. In each system the range of observed quantile ratio is within $+/-10\%$ of the target.

Choosing λ Initially set $\lambda = 1$ in the definition of D for each of the systems. Let $\mu_{k,sys} = \frac{\sum_{t_j=1}^{N_{sim}} D(X_{t_j,k})}{N_{sim}}$, be the average Ground Up Loss for variable k in the given system sys, where N_{sim} is the number of observations of, Ground Up Loss, D in the time series and $sys \in \{80001, ...80006\}$. Let $\mu_{80001} = \frac{\sum_{k=1}^{36} \mu_{k,80001}}{36}$ be the average D overall all 36 Lorenz variables for system 80001. Then define $\lambda_{k,sys} = \frac{\mu_{k,sys}}{\mu_{80001}}$ and define $\lambda_{sys} = \frac{\sum_{k=1}^{36} \lambda_{k,sys}}{36}$. Given this definition, $\lambda_{80001} = 1$ and the value for other systems ensures that the average D will be the same in each system. Table 4.12 shows the chosen values of $\lambda = \lambda_{sys}$ for each system. The range $(\min(\lambda_{k,sys}), \max(\lambda_{k,sys}))$ is also shown and in each system the range is within $+/-8\%$.

Choosing Δ As defined in equation 4.19, the Gross Loss to the insurer is the Ground Up Loss less any deductible Δ up to a limit Λ. In the following example the limit is assumed to be infinite. The target deductible is defined to be $\Delta_k = \frac{1}{20} Q_{50}(D(X_k))$ so that it bears a consistent relationship to a major loss in each of the systems; the value of 20 is arbitrary and chosen from personal experience[7]. Table 4.12 shows the chosen deductible values for each of the systems considered. It is interesting to note that after normalisation of τ and λ the resulting deductibles are all quite close together. The value chosen was defined as: $\Delta = \frac{\sum_{k=1}^{36} \Delta_k}{36}$, the average over all 36 Lorenz variables; the range $(\min(\Delta_k), \max(\Delta_k))$ is also shown which is within $+/-12\%$ of the chosen value.

Results of tuning process The distribution of Ground Up Loss, D, in all the systems is very similar, as intended. Predictability can now be explored in these

[7]The author has worked in the financial services industry for over 20 years and at Lloyd's of London, concentrating on catastrophe modelling, for over 10 years.

Table 4.12: Insurance index example, chosen parameter values for each Lorenz 96 system

system	chosen values of τ	mean quantile ratio	range of quantile ratio	Chosen values of λ_{sys}	Range of $\lambda_{k,sys}$ values	Chosen values of deductible Δ	Range of Δ_k
80001	2.50	15.93	14.61 - 17.03	1.00	0.946 - 1.069	18.0	16.6 - 19.2
80002	2.10	16.09	15.10 - 16.76	0.102	0.095 - 0.111	17.7	16.2 - 18.8
80003	1.25	16.00	14.50 - 17.40	0.0655	0.0612 - 0.0720	17.1	15.3 - 18.8
80004	2.00	16.06	15.18 - 17.00	0.441	0.421 - 0.464	17.8	16.6 - 19.0
80005	2.05	16.05	14.63 - 17.42	0.5985	0.5602 - 0.6249	17.9	16.5 - 19.6
80006	2.60	16.03	15.22 - 16.67	0.7273	0.6725 - 0.7600	18.0	16.5 - 19.5

systems and results compared across them (the distributions for Gross Loss, L, are also similar to one another). As expected (due to symmetry), the behaviour for each X_i are similar.

4.6.1 Determine ϕ transformation for Lorenz 96 index

The following derives the SOPLR ϕ for the index R required by examples 4.3.1.x and 4.3.2.y defined above. $N_{ens} = 2^4$ predictions of the insurance index are produced by each of the models and $N_{obs} = 2^9$ observations of the index are used in the optimisation routine to determine the SOPLR ϕ. Figure 4.12 shows a histogram of observed index values from system 80001 for illustration. The sensitivity of the SOPLR to the training data is illustrated for example 4.3.1.11, then comments on the different forms of ϕ are made separately for example groups 1 and 2.

Illustrating sensitivity to training data Figure 4.13 shows, for example 4.3.1.11, the SOPLR for five different (non-overlapping) sets of observations with accompanying model outputs (each of size N_{obs}). These illustrate the sensitivity of the SOPLR to different training data sets. In the figure the x-axis has been truncated to focus on the more probable (lower) index values for illustration. The black line relates to the training data set that is used to define ϕ for later use, the other illustrative examples are shown in blue but are not used hereafter. Key points are:

- The partition values (evident from different x-axis positions of the dots in figure 4.13) are different in each case because the method of equal bin sizes is used with different observations and this results in different interval endpoints in each data set, this is more evident for the extreme index values due to their scarcity;

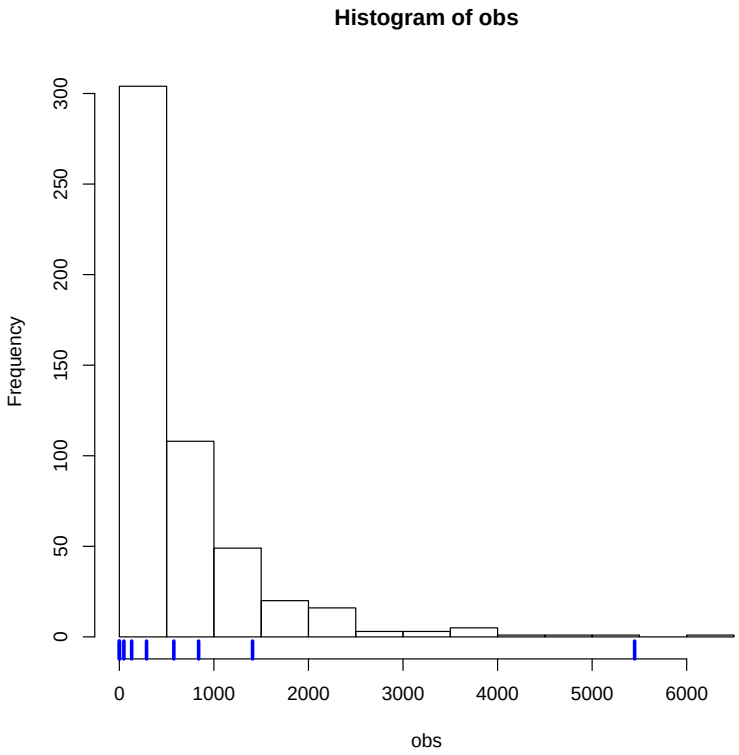


Figure 4.12: Example C4.3: Histogram of observed index values from System 80001 - y-axis is count of observations per bin out of 512 observations. Blue tick marks show the equal cardinality bins.

- All cases are above the line $y = x$, reminding us that the model is not the system and a naive approach to using model outputs for prediction would be flawed;
- The lines show a broadly similar pattern, but are far from identical, especially for observations for index values greater than 2000 (not shown) where the density is low. This indicates sensitivity to the training data set.

Results for Examples 4.3.1.x Figure 4.14 shows the estimated SOPLR ϕ for experiments 4.3.1.x. The results are broadly intuitive based on previous findings. Recall from Chapter 3 that model 10008 (black line) is chosen so that the forcing parameter is equal to the value in the system (i.e. $F = 10$), but that this means that the instantaneous effective forcing in the system is less on average. This causes model 10008 to produce larger extremes. In other words it tends to overshoot the system. Hence the relationship ϕ would be expected to be above the line $y = x$. Model 10009 uses the average forcing from the system, but still produces larger

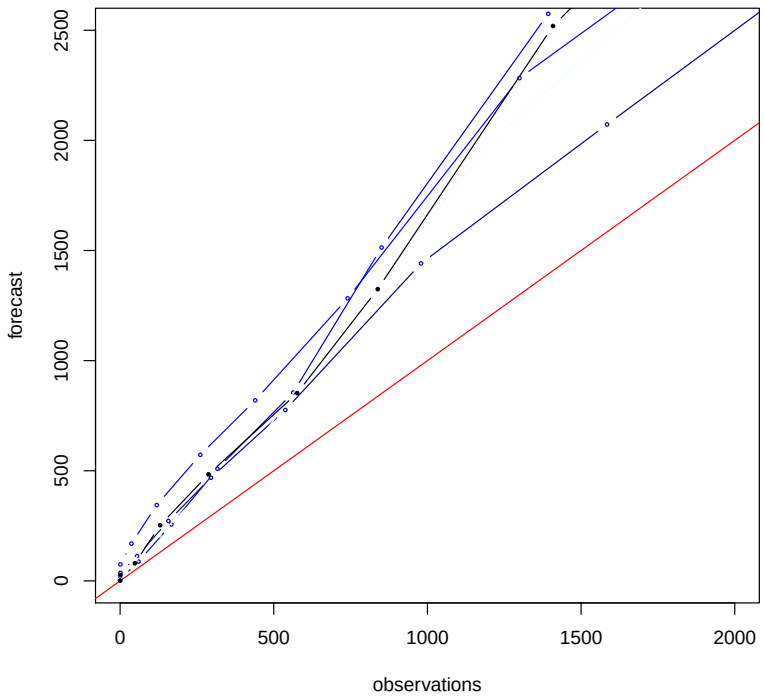


Figure 4.13: Example 4.3: Various examples of the SOPLR ϕ for different sets of observations. System 80001 and model 10011 are illustrated. Black line shows the result for the training data set. Blue lines show four other non-overlapping observation data sets to illustrate the stability of the estimator. The relationship is clearly above the red line $y = x$. The lines are not all identical indicating the estimator is quite sensitive to the observations, but do show the same pattern.

extremes than the system as shown in Chapter 3. As such, the relationship between the model and system would be expected to be closer to $y = x$ for 10009 than for 10008. Model 10010 uses a value chosen so that the 1 in 200 value is approximately equal to the system. As the Ignorance score tends to punish missed extreme values, model 10010 might be expected to be closer still to the line $y = x$. This is what occurs: model 10009 (red line) is closer to the line $y = x$ and 10010 is closer again. Model 10011 was chosen such that the forcing varies for each variable is dependent on the value of X_k (recall that it scored best out of all the methods). As such, the relationship ϕ might be expected to be closer to the line $y = x$ and this is the case for observed index values above 600. The least intuitive outcome is model 10012 which did not score so well but appears to have a relationship closer to $y = x$ than the others, for a portion of the curve it is actually below the line which may be consistent with the fact that at the 50^{th} percentile it actually undershoots the

system as shown in Chapter 3.

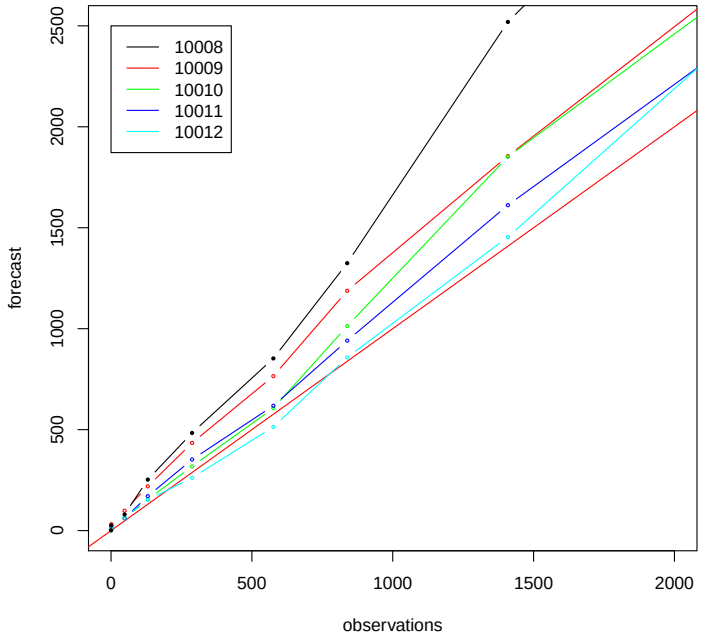


Figure 4.14: Example C4.3.1.x: SOPLR ϕ. The line $y = x$ is shown in solid red for comparison.

Results for examples 4.3.2.y Figure 4.15 compares the SOPLR ϕ for the system/model pairs defined by experiments 4.3.2.y. Given the sensitivity to training data any conclusions about the relative levels of the the SOPLR is tentative at best. For systems 80002 and 80003 (where F=20) ϕ is closer to the line $y = x$ than the other systems which have lower forcing parameters.

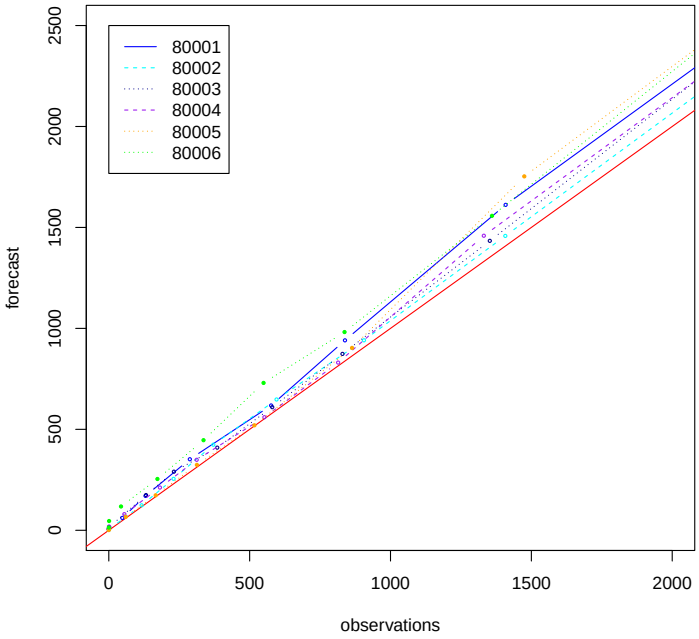


Figure 4.15: Example C4.3.2.y: SOPLRs ϕ. The line y=x is shown in solid red for comparison. Note that ϕ is above the $y = x$ line in each case.

4.6.2 Competition, profitability and insolvency

The previous section derived the SOPLR ϕ for each of the Lorenz systems of Chapter 3, and for model class *11. Each of these models was run with $N_{ens} = 2^4$ different initialisations (defined in Chapter 3) to produce an ensemble of forecasts for each time period. Four different sets of $N_{obs} = 2^9$ observations of the index were produced (each based on different system data than the training set used to determine ϕ) to enable comments to be made on the stability of the results. Each index observation R is accompanied by N_{ens} forecasts $\hat{R}^j$ where $j \in \{1, ...N_{ens}\}$. This section uses the SOPLR to derive a ϕ-transformed index price. As described above the climatology of the index R is used to determine the climatology price W_{clim}. Recall, this is the constant price that, on average, will be sufficient to pay for expected claims and provide the required return to shareholders based on required capital.

Competition effects The price arising from pricing models (the amount underwriters would *like* to charge) is typically referred to as the **'technical price'**

whereas the price actually charged is the **'market price'**. It is rarely possible to charge more than the market price for insurance, an issue that will be explored in Chapter 6. Although there is a degree of loyalty which allows some difference in price between insurers. In practice, as noted in a discussion with an experienced Lloyd's underwriter [179], underwriters tend to operate a dual strategy. If they believe that their pricing is more accurate (and lower) then they will charge a lower price and acquire market share from their competitors. Conversely if they believe the market price is too low they will tend to write less business[8] but keep the premium at the market price. When their technical price is lower than the market price underwriters may not lower their price fully to the technical price because they can charge more and still be the cheapest on the market. This pricing approach is defined in equation 4.23 where the market price is defined to be the climatology price W_{clim}. In the case where the technical price is above the market price underwriters often still write some business to ensure that they retain a presence in the marketplace. The impact of this approach will be explored using a simple rule which inflates or deflates the amount of business written depending on the technical price (defined in equation 4.24). W_{Blend}, W_{UE} and W_{clim} defined above are all technical prices. Define $\tilde{W}_m$, the market price for a given method m, as follows:

$$\tilde{W}_m = \begin{cases} W_m(1+\delta) & \text{if } W_m < \frac{W_{clim}}{(1+\delta)} \\ W_{clim} & \text{otherwise} \end{cases} \tag{4.23}$$

where W_m is the technical price for a given method m. δ is a price inflation factor which allows the insurer to charge more than the technical price when they know their competitors are charging even more.

Now assume that the insurer has a rule, $\beta(W_m)$ (defined in equation 4.24), which determines the change in business volume they will accept for a given technical price (assuming 100% as a starting point if $W_m = W_{clim}$). This implicitly assumes that the insurer receives sufficient requests for quotations that they can control the volume of business; in practice this may not always be the case. The business volume rule

[8]When insurers speak of 'business written' or 'business volumes' the are referring to the number of policies sold. Therefore 'write less business' means sell fewer policies or take a smaller share on any co-insured risks.

could take various forms but in this example is defined below:

$$\beta(W_m) = \begin{cases} \beta_{max} & \text{if } W_m < W_{min} \\ \beta_{max}(\frac{W_{clim}-W_m}{W_{clim}-W_{min}}) & \text{if } W_{min} \leq W_m < W_{clim} \\ \beta_{min}(\frac{W_m-W_{clim}}{W_{max}-W_{clim}}) & \text{if } W_{clim} \leq W_m < W_{max} \\ \beta_{min} & \text{if } W_m \geq W_{max} \end{cases} \tag{4.24}$$

Where W_{min} and W_{max} are threshold prices beyond which the change in business volume is constant. β_{max} and β_{min} are the maximum and minimum volume changes for each threshold price respectively.

In all the examples below: $W_{min} = 500$, $W_{max} = 3000$, $\beta_{min} = -0.9$ and $\beta_{max} = 0.75$. This means that for technical premiums equal to or lower than 500 the underwriter will write 75% more business than if they charged the market price. For technical premiums greater than 3000 they will choose to write just 10% of the business they would have done. δ is set at 20% in all examples unless otherwise indicated. These parameters are purely illustrative.

Profitability The current example uses a simplistic approach which will be extended in Chapter 6. Profit is simply calculated as the difference between the price charged and the payment actually made. Positive profits will be made some years and not others and by counting the years in which this occurs an approximate probability of positive profitability can be estimated. This is denoted 'P(+profit)' in the results tables below. It is also interesting to look at the average amount of profit conditional on the fact that it is either positive or negative since some strategies may protect against large negative profits at the cost of lower positive ones. These are denoted '+ve profit' and '-ve profit' in the results tables. The level of profitability indicates whether a strategy is better or worse for the shareholders. Policyholders are arguably indifferent to this, although profits that are too high might indicate they are getting poor value for money; conversely profits that are too low might indicate an insurer that is weakly financed which may therefore fail if large claims arise. The level of profit will depend on the volume of business sold, controlled by the rule β defined above. Specifically **'profit (π)'**, for a given period, is defined as:

$$\pi = (\tilde{W}_m - R) * (1 + \beta(W_m)) \tag{4.25}$$

Insolvency As explained in Chapter 1 the strict definition of insolvency is complex. This chapter defines insolvency when the payout in the year exceeds the premium charged and capital held. Equivalently (in this simple example) if the sum of profits and capital is less than zero then the company is insolvent. Again, by counting the number of instances an approximate probability of insolvency 'P(insolvent)' can be estimated and this can be used to compare strategies. This measure is of interest to both shareholders and policyholders; the latter particularly because it indicates how likely the company will be to pay their claims; the former because it defines the probability that the value of their investment will be lost.

4.6.3 Full Blending pricing - parameter calculation

This section shows (1) how the blending parameters α and σ_m were chosen in the Full Blending method for example 4.3.1.11 and (2) the results of using this method for the remaining examples.

Method to choose blending parameters Figure 4.16 illustrates how the blending parameters were chosen for example 4.3.1.11. Different values of α and the kernel dressing bandwidth are trialled and the average Ignorance score over $N_{obs} = 2^9$ observations of the index versus the forecast is calculated. The combination of parameters which lead to the lowest score on this grid is chosen. The process is carried out in two stages - first a wider grid (α in steps of 0.1 and the bandwidth in steps of 10) - then once a likely region is found the process is repeated around the candidate value in steps of 0.02 for α and 2 for the bandwidth.

Figure 4.17 gives one example of a forecast probability density function arising from the blending process. The blue tick marks show the forecast ensemble values. The blue density line shows the results of kernel dressing these points (note this uses the same bandwidth as the blended forecast and so has not been chosen to be score-optimal). The green line shows the PDF of the climatology of the index. Finally the black line shows the PDF of the blended forecast. In this example the blending parameter $\alpha = 0.8$. The two vertical lines show the 1 in 200 (99.5%) quantile values of the kernel dressed and blended forecasts. This highlights very clearly the role of blending in retaining extreme values within the forecast; without this the regulatory

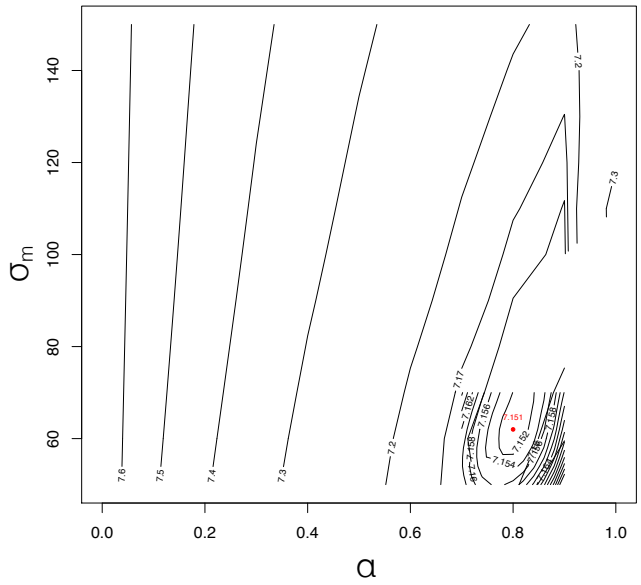


Figure 4.16: Illustration of selection of low scoring parameters for the blending process. Red dot shows the chosen value and the two sets of contours (of the average score) illustrate the two step process for choosing it. System 80001 and model 10011.

quantile $Q_{200}(\hat{R})$ would only be 1179 compared to 3170 with blending; in other words the assessed capital requirement for kernel dressing only would be one third of the blended value.

Blending parameter results for all examples Figure 4.18 shows parameters that result from the blending method. These are chosen to give a low score on a grid of bandwidths that are separated by a step of size 2 (i.e. 80,82,84 etc) and α values that are separated by steps of 0.02. As such, they are not necessarily truly 'optimal' but are expected to be close to the value an optimisation routine would chose - the latter option was not used to due run time constraints.

- For system 80001 and the five forecast methods 10008,...10012. Forecast 10011 has the highest α value i.e. high weight to the dressed ensemble; this result is consistent with those of Chapter 3. Forecast 10010 has a similar α value to 10012 but a thinner kernel width indicating that more emphasis is placed on the ensemble values in the former case.
- For forecast system *11 the higher predictability systems (80001 and 80006) have α values that are closer to 1 than the other systems. Systems with

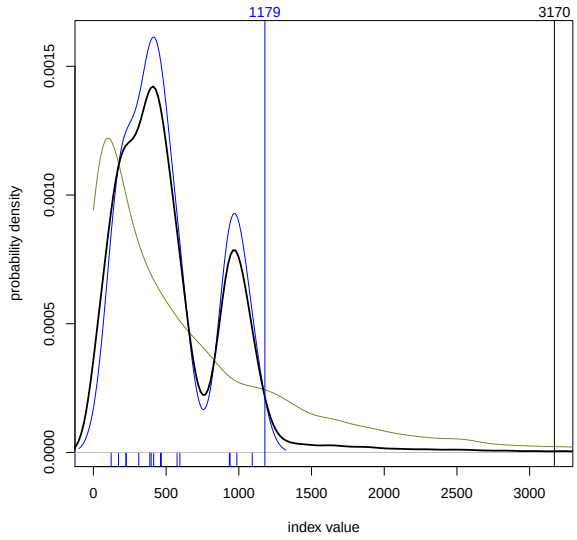


Figure 4.17: PDF of index distributions: Green = climatology, Blue=Kernel dressed forecast, Black = Blended forecast

medium predictability (80004 and 80005) have low α but also a reasonably thin kernel dressing bandwidth which will give considerable weight to the forecast ensemble but preserve the tail of the climatology. Finally, low predictability systems (80002 and 80003) have low α and wider bandwidths.

4.6.4 Updated Expectation Results - Example 4.3.1.11

Figure 4.19 shows a histogram of the technical price using the Updated Expectation method which can vary considerably depending on the forecast. The climatology price (1088) is shown by the vertical red line.

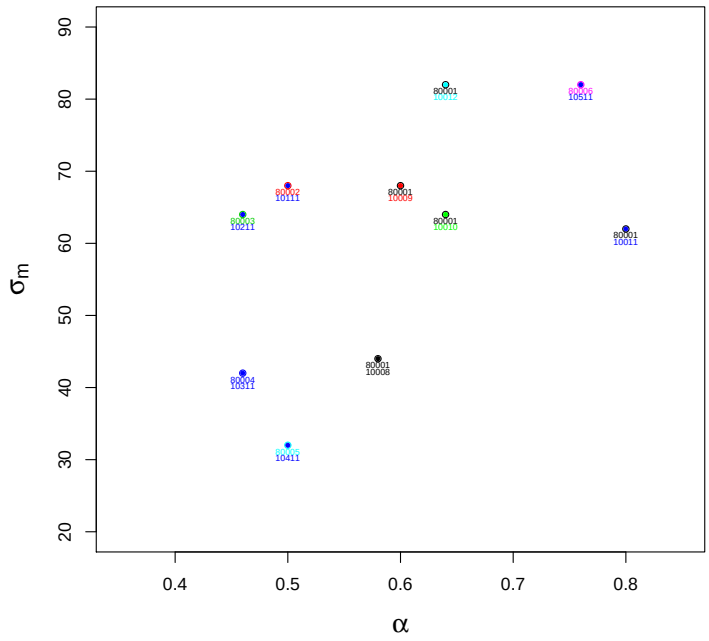


Figure 4.18: Grid-best blending parameters for various systems and models. Plot labels show the system ID on the top and the model ID on the bottom, the inner dot colour represents the forecast, outer ring represents the system - labels are in the same colours.

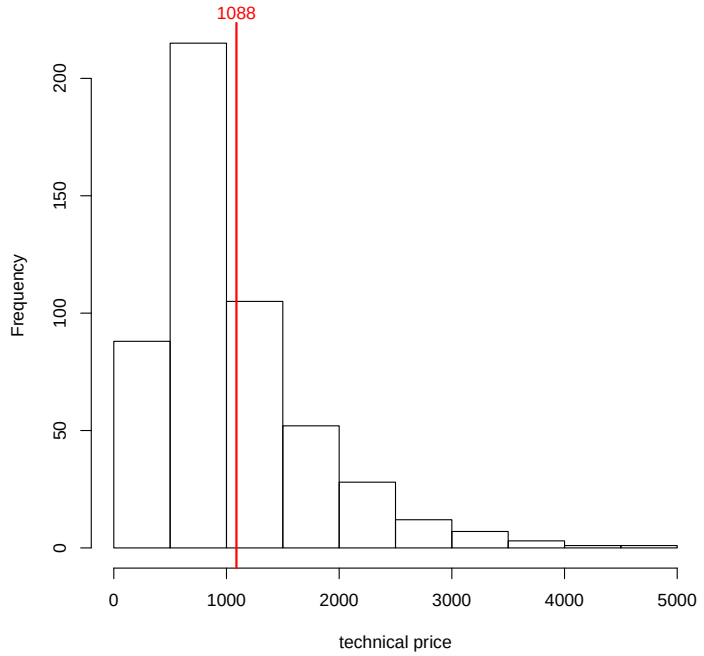


Figure 4.19: Lorenz 96 Index example: Histogram of technical price based on the Updated Expectation method. Vertical red line indicates the climatology price.

Table 4.14 shows a comparison between the climatology price and three variants of the updated expectation method defined in table 4.13:

Table 4.13: Definition of columns in results tables

Column name	Description
W_{clim}	Results when the climatology price is used in all cases
Variant A	Assumes the technical price can be charged at all times - business volumes are 100% for each price (i.e. $\beta_{max} = \beta_{min} = 0$ in equation 4.24 and $\tilde{W}_{UE} = W_{UE}$))
Variant B	β is defined by 4.24 as usual, but $\delta = 0$ in equation 4.23
W_{UE}	Prices are W_{UE} defined above (i.e. β and δ as defined without adjustment)

As already mentioned it is rarely possible to charge the technical price at all times. As such, Variant A is largely theoretical. It could arise if all companies in the market switched to ϕ-transformed pricing. Key results are:

- **Variant A** Despite some very large premiums on occasion the average technical price (1020) is actually lower than the climatology price (1088). The ranges beneath the figures denote the high and low values from the four distinct observation data sets. For example, in the case of Variant A the lowest average price was 978 and the highest 1080. With Variant A no competition effect is modelled (and when climatology pricing is used there is no competition at all) therefore the difference in average profit between climatology pricing and technical pricing is simply the difference between the average premium rate (since the claims are the same in each case). Variant A is less profitable on average, but gives a 6% higher chance of making a profit in the year. When profits are made they tend to be smaller when positive, but also losses are less extreme. The probability of being insolvent under Variant A is lower than using the climatology price indicating that the forecast is able to give a warning against very large index levels allowing adequate prices to be charged in those years, thereby avoiding insolvency.
- **Variant B** introduces competition effects but assumes that no uplift is applied to the technical price (i.e. $\delta = 0$) in this case the average premium is much lower because the premium is capped at W_{clim} and the probability of making a profit is actually lower than for climatology pricing but the probability of insolvency is much lower.
- **Updated expectation** W_{UE} This pricing method includes a 20% uplift whenever the technical price is more than 20% lower than the market (climatology)

price. Despite this, the average price (891) remains lower than the climatology price. Average profits (444) are higher than those using climatology price although the probability of a profit in a given year is slightly lower (78.4% compared to 79.4%). The probability of insolvency remains much lower than climatology pricing. As such this method is arguably better for all stakeholders: policyholders, shareholders and regulators. In the case of regulators the supervisor of the individual company will arguably prefer that the company is less likely to go insolvent; the insurance regulator in total may be less happy that some other companies will be taking up excess business and are consequently more likely to become insolvent. Such a regulator may take a more strategic view, however, and believe that in the long term all companies may adopt the more accurate pricing method; they may even encourage such a change.

Table 4.14: Example 4.3.1.11 Results for climatology pricing and three variants of updated expectation

	W_{clim}	Variant A	Variant B	W_{UE}
avg premium	1088	1020	812	891
	1088-1088	978-1080	794-823	877 - 897
avg profit	408	339	326	444
	330-475	317-384	275 - 391	390-511
avg +ve profit	752	512	587	709
	724-769	504-518	560-605	676-727
avg -ve profit	-923	-654	-510	-519
	-955 - -892	-697 - -633	-570 - - 487	-577 - -491
P(+profit)	0.794	0.853	0.762	0.784
	0.750 - 0.832	0.832 - 0.895	0.719 - 0.811	0.740 - 0.824
P(insolvent)	0.00684	0.00244	0.000977	0.000977
	0.0039 - 0.0098	0 - 0.0039	0 - 0.00391	0 - 0.00391

4.6.5 Selected results for all other examples

The following are a series of observations and general statements about the selected results of the example 4.3. Examples 4.3.1.x and 4.3.2.y are discussed separately. The example names are shortened to *x and *y below. Recall that *x relates to model 1000x in system 80001 and *y relates to system 8000y. The text below therefore sometimes refers to 'model *x' and 'system *y'.

Example 4.3.1.x Tables 4.15, 4.16 and 4.17 show the results for this example for climatology, Updated Expectation and Full Blending respectively. Climatology results are the same for each column of table 4.15 because the same system (80001) is used in each case. The columns are retained for easier comparison with other tables.

- **Updated Expectation** When comparing the Updated Expectation method for models *11 and *10 table 4.16 shows that the average premium charged is the same. Yet *11 is much more profitable (444 vs 393) due to lower negative profits in bad years despite the fact it makes slightly less profit in good years. Overall *11 has a higher probability of making a profit than *10. It also has a much lower probability of insolvency. Model *10 actually does worse than climatology pricing for the majority of metrics - but does have a lower probability of insolvency, so on this critical metric model *10 is still adding value.
- **Updated Expectation** All of the models in this example lead to a lower chance of insolvency than climatology pricing despite all charging lower premiums on average.
- **Full Blending** The Full Blending pricing method results for model *8 versus *11 are shown in table 4.17. Model *8 leads to a considerably higher average premium of 953 vs 872, an increase of 81, yet only manages to create increased profits of 28 with only a very slight increase in probability of making a profit. In this case the probability of insolvency is the same between the two models.
- **Full Blending** In general, Full Blended pricing in system 80001 leads to lower probability of insolvency than climatology pricing; but not as low as using the simpler Updated Expectation method except for model *12. The fact that Full

Blending has higher insolvency probability than Updated Expectation may be due to the fact that the capital held also adjusts according to the forecast so that funds are not available to survive a major claim when there is a forecast bust. This issue is discussed further in Chapter 5.

Example 4.3.2.y Tables 4.18, 4.19 and 4.20 for climatology, Updated Expectation and Full Blending respectively.

- **Climatology** Using system *2 as an example, table 4.18 shows, that despite creating an index with a similar mean and overall spread, the insurance statistics between systems can be different - making them difficult to directly compare. For example the average premium is 1050 compared to 1090 in system *1 some 40 lower, yet profits are 96 lower. Perhaps this is to be expected since system *2 produces more frequent extremes than *1. The probability of making a profit is lower in this system than *1 for example. However, it is interesting that the probability of insolvency is actually lower in *2, which is a surprise.
- **Updated Expectation** Continuing with system *2 as an example table 4.19 shows that in this case the Updated-expectation method does not lead to increased profits, or even a decreased probability of insolvency. Similar comments can be made for system *3. This is not a surprise because Chapter 3 showed that the α values in the blending exercise were low from early on in the period showing that the models quickly loose their predictive power in these systems. The Updated expectation method produces lower average premiums, higher profitabilty and lower risk of insolvency for all of the systems *1, *4, *5 and *6.
- **Full Blending** The Full Blending pricing method again under-performs the Updated Expectation method (as shown in table 4.20) in all systems except *6 where it is more profitable - though still more likely to go insolvent. As noted, the additional insolvencies may be caused when the blending method leads to a reduction in capital in a year where the actual index is high (i.e a forecast bust). It would be possible to restrict the change in capital (e.g. apply bounds to $\kappa(\hat{R})$) to test whether the Blending method is then more robust.

Table 4.15: Example 4.3.1.x - Prices use Climatology method (W_{clim})

	4.3.1.8	4.3.1.9	4.3.1.10	4.3.1.11	4.3.1.12
avg premium	1088	1088	1088	1088	1088
	1088 - 1088	1088 - 1088	1088 - 1088	1088 - 1088	1088 - 1088
avg profit	407.6	407.6	407.6	407.6	407.6
	330.3 - 474.8	330.3 - 474.8	330.3 - 474.8	330.3 - 474.8	330.3 - 474.8
avg +ve profit	752.2	752.2	752.2	752.2	752.2
	724.3 - 769.5	724.3 - 769.5	724.3 - 769.5	724.3 - 769.5	724.3 - 769.5
avg -ve profit	-922.7	-922.7	-922.7	-922.7	-922.7
	-955.1 - -892	-955.1 - -892	-955.1 - -892	-955.1 - -892	-955.1 - -892
P(+profit)	0.7944	0.7944	0.7944	0.7944	0.7944
	0.75 - 0.832	0.75 - 0.832	0.75 - 0.832	0.75 - 0.832	0.75 - 0.832
P(insolvent)	0.006836	0.006836	0.006836	0.006836	0.006836
	0.003906 - 0.009766	0.003906 - 0.009766	0.003906 - 0.009766	0.003906 - 0.009766	0.003906 - 0.009766

Table 4.16: Example 4.3.1.x - Prices use Updated-expectation method (W_{UE})

	4.3.1.8	4.3.1.9	4.3.1.10	4.3.1.11	4.3.1.12
avg premium	985.2	942.2	890.6	891.3	978.1
	976.6 - 989.8	939.8 - 947.8	885.3 - 899.2	877.3 - 897.9	968.8 - 982.6
avg profit	474.1	445.4	392.9	444.1	457.3
	426.6 - 526.2	394.1 - 497.3	332.2 - 449	389.8 - 511.3	410.1 - 507.8
avg +ve profit	749.4	732.2	734.4	709.9	737.3
	726.2 - 766.2	705 - 744.4	707.2 - 765.4	676 - 727	713.1 - 755.9
avg -ve profit	-526.9	-568.1	-718.6	-519.1	-552.2
	-563 - -476.8	-621.6 - -508.8	-771.9 - -654.4	-576.6 - -491.3	-591.1 - -524.8
P(+profit)	0.7837	0.7788	0.7642	0.7837	0.7827
	0.7422 - 0.8262	0.7383 - 0.8184	0.7227 - 0.8145	0.7402 - 0.8242	0.7363 - 0.8242
P(insolvent)	0.001465	0.001953	0.004395	0.0009766	0.002441
	0 - 0.001953	0 - 0.003906	0 - 0.007812	0 - 0.003906	0 - 0.003906

Table 4.17: Example 4.3.1.x - Prices use Full Blending method (W_{Blend})

	4.3.1.8	4.3.1.9	4.3.1.10	4.3.1.11	4.3.1.12
avg premium	952.7	904.1	837.4	872.3	940.5
	942.7 - 959.5	899.7 - 910.9	831.4 - 847.7	858 - 880.5	930.4 - 945.3
avg profit	448.7	406.3	326	421.4	425.4
	399.8 - 500.8	349.7 - 457.6	263.2 - 383.9	365.3 - 494.8	375.7 - 477.5
avg +ve profit	730.5	704	690.9	722.2	717.8
	708.4 - 750.4	677.8 - 723.5	666.4 - 722.4	692.6 - 737.2	691.3 - 737.3
avg -ve profit	-557.8	-596.5	-763.6	-637.3	-592.6
	-608 - -507.8	-666 - -532.2	-808.2 - -701.1	-692.1 - -610.7	-658 - -563.2
P(+profit)	0.7808	0.7705	0.7485	0.7788	0.7769
	0.7402 - 0.8242	0.7305 - 0.8164	0.707 - 0.7988	0.7344 - 0.8223	0.7285 - 0.8164
P(insolvent)	0.001465	0.002441	0.004883	0.001465	0.001953
	0 - 0.001953	0 - 0.003906	0 - 0.009766	0 - 0.003906	0 - 0.003906

Table 4.18: Example 4.3.2.y - Prices use Climatology method (W_{clim})

	4.3.2.1	4.3.2.2	4.3.2.3	4.3.2.4	4.3.2.5	4.3.2.6
avg premium	1088	1054	969.6	1134	1132	1065
	1088 - 1088	1054 - 1054	969.6 - 969.6	1134 - 1134	1132 - 1132	1065 - 1065
avg profit	407.6	312.2	268.7	378.6	396.7	353.7
	330.3 - 474.8	264.5 - 339.9	241.4 - 299.9	320.9 - 402.6	366.9 - 411.1	325.9 - 392.8
avg +ve profit	752.2	656.9	597	763.6	775	710.9
	724.3 - 769.5	641.8 - 685.2	576.4 - 610.1	756 - 775.3	760.3 - 791.8	685.2 - 723.4
avg -ve profit	-922.7	-769.4	-692.6	-971.6	-912.7	-965.5
	-955.1 - -892	-784.3 - -747.1	-737.2 - -641.6	-1085 - -852.7	-956.6 - -854.9	-1008 - -930.7
P(+profit)	0.7944	0.7583	0.7456	0.7783	0.7759	0.7866
	0.75 - 0.832	0.7344 - 0.7832	0.7305 - 0.7598	0.7637 - 0.793	0.7676 - 0.7812	0.7617 - 0.8125
P(insolvent)	0.006836	0.003906	0.00293	0.005371	0.003906	0.003418
	0.003906 - 0.009766	0 - 0.007812	0 - 0.005859	0.003906 - 0.007812	0.001953 - 0.005859	0.001953 - 0.005859

Table 4.19: Example 4.3.2.y - Prices use Updated-expectation method (W_{UE})

	4.3.2.1	4.3.2.2	4.3.2.3	4.3.2.4	4.3.2.5	4.3.2.6
avg premium	891.3	969.5	899.1	996	985.3	869.7
	877.3 - 897.9	963.5 - 975.7	895.5 - 905.9	989.4 - 1004	979.4 - 989.7	864.1 - 873.6
avg profit	444.1	324.5	277.6	421.9	421.9	373.2
	389.8 - 511.3	279.1 - 361.1	246.4 - 320.6	384 - 457.2	397.8 - 436.1	339 - 413
avg +ve profit	709.9	703	661.1	790.8	787.3	677.6
	676 - 727	686.8 - 725.9	643.2 - 674.8	783.8 - 804.8	776.5 - 794.3	656 - 692.7
avg -ve profit	-519.1	-734.1	-718.3	-686.4	-719.7	-601.2
	-576.6 - -491.3	-772.9 - -689.3	-797.2 - -646.6	-771.8 - -622.5	-866 - -657.6	-645.7 - -555.7
P(+profit)	0.7837	0.7368	0.7222	0.7505	0.7573	0.7622
	0.7402 - 0.8242	0.7207 - 0.752	0.707 - 0.7383	0.7324 - 0.7695	0.752 - 0.7617	0.7402 - 0.7852
P(insolvent)	0.0009766	0.005859	0.008301	0.00293	0.002441	0.001953
	0 - 0.003906	0.003906 - 0.007812	0.003906 - 0.01172	0 - 0.005859	0 - 0.007812	0 - 0.003906

Table 4.20: Example 4.3.2.y - Prices use Full Blending method (W_{Blend})

	4.3.2.1	4.3.2.2	4.3.2.3	4.3.2.4	4.3.2.5	4.3.2.6
avg premium	872.3	907.6	885.1	950.8	942.3	888.4
	858 - 880.5	900.4 - 913.9	882.3 - 892.3	943.1 - 961.2	935.3 - 946.9	883.4 - 891.5
avg profit	421.4	260.2	265.6	376	376.8	391
	365.3 - 494.8	212.8 - 303.2	233.1 - 310	339.4 - 412.9	352.7 - 390.5	352.7 - 433.6
avg +ve profit	722.2	686.3	662.4	771.9	765.3	732.1
	692.6 - 737.2	669.9 - 701.9	645.8 - 671.8	761.6 - 780.4	755.8 - 778.3	706 - 752.2
avg -ve profit	-637.3	-814.4	-742.7	-728.4	-764.4	-707.4
	-692.1 - -610.7	-858.5 - -788.2	-835.9 - -661.6	-824.2 - -666.7	-933.8 - -684.6	-746.1 - -640.7
P(+profit)	0.7788	0.7163	0.7178	0.7363	0.7456	0.7632
	0.7344 - 0.8223	0.6973 - 0.7344	0.7031 - 0.7324	0.7129 - 0.7559	0.7363 - 0.7559	0.7422 - 0.7891
P(insolvent)	0.001465	0.007812	0.009766	0.00293	0.003906	0.00293
	0 - 0.003906	0.005859 - 0.009766	0.003906 - 0.01367	0 - 0.005859	0 - 0.007812	0 - 0.005859

4.7 Further work

Testing indices comprised of multiple variables The Lorenz 96 example described in this chapter was based on an index set H which only contained one variable. Equation 4.1 allows for an index to be created over multiple variables. To make easy comparisons between systems it may be helpful to consider a portfolio with identical expected payouts regardless of the number of variables included. To achieve this the Ground up losses must be scaled as follows. If there are $|H|$ regions in the portfolio (say) then, in the notation of section 1:

$$f_k(t_j, \Theta) = \frac{1}{|H|} L_{t_j,k} \tag{4.26}$$

Where $\Theta = \{\tau, p, \Delta, \Lambda\}$

Extend the results set to other systems Results for the remaining systems and models considered in Chapter 3 could be created to further test the robustness of the conclusions. Also, different Lorenz systems could be explored for example with different forcing values or coupling, or with a trend in the forcing to illustrate the further difficulties of pricing in a non-stationary system. Finally different ODEs (such as the Duffing model considered in Chapter 2) could be considered to allow exploration for systems with different dynamics. Outputs from climate models could be taken to create indices from those - this would be a step closer to using the methods in the chapter to create viable insurance indices in practice.

Use insurance industry model The insurance industry model to be developed in Chapter 6 could be further developed to take claims information from the Lorenz 96 system - then issues such as competition effects could be further explored in this system as could the use of reinsurance or other risk mitigation tools. The Lorenz 96 system also serves as a useful **'claims generator'** which other insurance industry practitioners could use to test their models. Being a dynamical system it arguably produces more realistic behaviour than standard statistical distributions (such as runs of high values, waves and physical bounds on the largest values).

Use blended forecasts through the time period The Blended-pricing method described in this chapter uses the annual index - and then compares the forecast

ensemble values to the observed index. The work in Chapter 3 created blended forecasts for each time step during the year (i.e. 24 forecasts). This showed that in the first part of the time period the forecasts had considerable skill in some cases - but the latter time periods less so (as evidenced by the reduction in α value and widening of the kernel dressing bandwidth). It would be possible to develop a pricing method that made use of the forecast for part of the year only - and then used the climatology price for the remainder of the year. This would be a good proxy for the current situation in hurricane pricing where predictability degrades over the hurricane season.

Putting a value on the forecast If the forecast can be used to improve profitability, increase financial strength or lower average premium rates then it is successful. The question then arises: how much would the company pay for the forecast? McCarthy [165] notes that it is 'always a good idea to look at the outcome of an experiment if it is free', but what if it isn't? If the company absorbed the costs itself then any amount less than the difference in average profitability, between using the models and not, would leave the sophisticated company better off. In practice, however, companies would wish to retain some of the additional profits to make it worth their while to change processes. It would be important to assess the probability of insolvency to check that this additional cost does not increase it. Alternatively the company could pass the forecast costs to policyholders whenever the technical price is sufficiently below the climatology price (in the case where competitors are not making use of the forecasts). A hybrid strategy, that charged policyholders for the forecast when possible, but absorbed the costs otherwise, would likely be used in practice.

4.8 Conclusions

This chapter has used the Lorenz 96 system to explore whether forecasts can be useful for insurers. First the technique of ϕ-transformation was introduced which finds the piecewise linear relationship between the system and a model that has the highest average skill score. The technique is shown to work in two idealised situations.

The Lorenz 96 system is used as a proxy for an atmospheric (or other dynamical) hazard from which insurance protection may be sought. The system variables are converted into more decision relevant values (in this case by applying a threshold and power rule so that higher values are given more weight and then applying insurance deductibles and limits). A general insurance index is then described which integrates multiple decision relevant variables over multiple time periods. The index is priced using traditional climatology techniques and then using two techniques which use ϕ-transformed forecasts. The first technique simply adjusts the expected index values (arising from an ensemble of modelled outputs) in the pricing formula; the second uses climatology blending. The first technique is shown to be more successful. In general it is concluded that the forecasts were useful. Specifically they improved profitability, led to lower premiums on average and also to less chance of insolvency. As such, all major insurance stakeholders would have increased utility through the use of forecasts: shareholders, policyholders and regulators.

The models in the Lorenz 96 example are closely related to the system. Although they regularly diverge from the system they can match it closely for at least part of the period of insurance. The insurance payout is defined precisely by the system values and as such where the model is close to the system it will provide a very good estimate of the payout. The next chapter considers hurricane risk where, it is argued, there is less correspondence between forecasts and the eventual insurance payout. The value of forecasting will be assessed in this more challenging setting.

Chapter 5

The value of forecasting in hurricane insurance

'***There is an inherent curiosity amongst the general public for any quantitative information about how active the upcoming [hurricane] season is likely to be, and we show in this manuscript that this is possible. These models will likely continue to be modified and improved in the years ahead as additional data and improved physical insights become available.***'

Philip J. Klotzbach and William M. Gray 2009 [126]

'***Under the highly idealized conditions of this experiment, there is a clear advantage to the hypothetical property insurance firm of using seasonal hurricane forecasts to adjust the amount of reinsurance it purchases each year ... But when a more realistic seasonal forecast skill is assumed, the potential value of forecasts becomes significant only after more than a decade.***'

Kerry Emmanuel et al 2012 [75]

The insurance industry exposes itself annually to losses from hurricanes. To date the most costly year was 2005 when hurricane Katrina cost the industry USD80bn [248][1]. Seasonal weather forecasting methods are becoming more sophisticated [125]. It is likely that the skill and capabilities of these forecasts will increase over the

[1]This was material when compared to an estimated, total, insurance premium for catastrophe risks of USD73bn in 2009 [254].

coming decades [126]. This chapter seeks to investigate the extent to which models that can make skilful forecasts of some elements of risk, but not others, are useful[2] for insurers[3]. For example, currently the forecasting group 'Tropical Storm Risk' based at University College London can forecast [221] the number of Atlantic storms that make landfall with some skill, but not whether one will hit Miami.

A simple model of insured hurricane losses is created. Basin hurricane counts are sampled from a Poisson distribution, then successively sub-sampled to create landfall numbers and major city hits. The severity of landfalling storms is also simulated as are insurance losses which are modelled by assuming a simplified 1-1 relationship between hurricane strength and loss. The parameters to simulate basin number, landfall proportion and severity are based on the HURDAT[4] data set from 1955 to 2010[5]. Various pricing methodologies are then considered which make use of successively better forecasts. It is shown that simple business volume scaling methods that react to forecast information can improve expected profitability. More sophisticated forecasts, if used in pricing, can lead to reduced capital requirements in quiet years, but would lead to lower profitability unless steps were taken to ensure premium levels are on average no lower than traditional pricing methods. Where pricing is changed it is assumed that the whole market have adopted the pricing method; otherwise competition effects would severely restrict their adoption. Finally it is noted that, given the natural variability of this system, it is very difficult to distinguish between an underwriter who is good or just lucky. For natural catastrophes it is easy to be 'fooled by randomness' [249]. The original elements of this chapter are believed to be:

- Development and presentation of a simple statistical hurricane landfall process;

[2]'Useful' forecasts, from an insurance perspective, would be, for example, those that help to improve profitability or postpone insolvency for longer.

[3]Summary conclusions from this chapter were presented by the author [163] at an R Met Soc meeting in London in February 2011 and then in detail at Cass Business School in March 2011 [162]. The work was carried out to support the publication of a Lloyd's of London report [139] 'Forecasting Risk': The value of long-range forecasting for the insurance industry".

[4]Hurricane data from the US National Hurricane Centre is updated annually. The data is referred to as HURDAT (**Hur**ricane **Dat**abases.)

[5]This work was carried out in 2011. The results of this chapter are discursive and do not depend on the latest available data set being used.

- Illustration of efficacy of different quality forecasts using adjusted pricing based on a standard reinsurance pricing method first proposed by Kreps [128];
- Comparison of a return-on-capital pricing method with the Kreps method;
- Comments on perception of underwriter performance in the context of the hurricane process.

5.1 Description of simple hurricane landfall model

For a hurricane to cause a major loss the following has to occur: (1) a hurricane forms; (2) it makes landfall; (3) at landfall it is intense, and finally; (4) the landfall location occurs where exposure density is high (i.e. it impacts a major urban or commercial centre). The basic simulation examined in this chapter is as follows (see also figure 5.3):

- Let $N_B \sim \text{Poisson}(\lambda)$ be the number of hurricanes that form in the North Atlantic Basin, where λ is the expected number of hurricanes in the year;
- Let the the number of these that make landfall $N_L|N_B \sim \text{Binomial}(N_B, q)$, where q is the probability that a basin hurricane will make landfall;
- Let N_C be the number of these which hit a major city or commercial centre so that $N_C|N_L \sim \text{Binomial}(N_L, c)$, where c is the proportion of landfalling hurricanes that are city hits;
- Simulate the Saffir Simpson[6] strength of each landfalling storm, according to their observed frequencies (call these $sa_1, ...sa_{N_L}$) assume this is independent to landfall location, uniformly sample N_C of these, which are deemed to be the city hits, assume a 1-1 correspondence between strength *of a city hit* and financial loss ($S_i = S(sa_i)$) distribution;
- Calculate the Premium charged P_j where j denotes the pricing method used (defined in section 5.2);
- Calculate the profit as $\pi = P - \sum_{i=1}^{N_C} S_i$.

[6]The Saffir Simpson hurricane wind scale [183] is a series of ranges of wind speed split into named categories: Tropical Storm, plus 1-5 Hurricane strength. These were designed to broadly correlate with the damage caused, for example: category 1 states that 'dangerous winds will produce some damage' and category 5 that 'catastrophic damage will occur'.

The following discussion is not particularly complex; but its implications are new and give insights into the difficulties of making use of forecasts in an insurance context.

Number of hurricanes generated N_B: $N_B \sim Poisson(\lambda)$. It is standard to use a Poisson distribution for catastrophe hazard frequencies [117, 120, 124, 171]. The variance in average number of hurricanes generated per year since 1955 is 6.9 which is larger than the mean of 6.1. This is known as 'over-dispersion' since with a Poisson distribution the variance should equal the mean. Allowance for over-dispersion can be made in various ways such as using a Negative Binomial distribution [66] but here the Poisson expectation parameter is rounded up to $\lambda = 7$. See figure 5.1 which compares histograms of actual hurricane numbers and a Poisson(7) distribution sample.

Number of landfalls N_L: $N_L \sim binomial(N_B, q)$. On the assumption that each hurricane generated in the Atlantic has an equal probability of making landfall the total landfall count is the sum of N_B Bernoilli trials [101] resulting in a Binomial distribution. The long term average proportion of landfalling hurricanes from those generated (based on HURDAT data up to 2010) is $q = 24\%$ see figure 5.2. The assumption of constant landfall proportions is contentious in the literature [46, 48, 55, 120, 161, 242]. This is not a constraint for this chapter, however, because the hurricane model is deliberately simple to enable clear investigation of the efficacy of forecasts in insurance pricing.

Number of city hits N_C: A simple model of whether a major city is hit is defined as follows:

- The US east coast is around 3600 miles long [111];
- Assume that the destructive winds from each hurricane falls into a 'slot' exactly 90 miles wide - so there are 40 such slots on the US coast;
- Assume there are 10 major population centres on the coast;
- Say that each city is sufficiently far away from the others, so there is a zero probability of a hurricane hitting two, also assume that each city is in the middle of a coastline 'slot' (defined above);

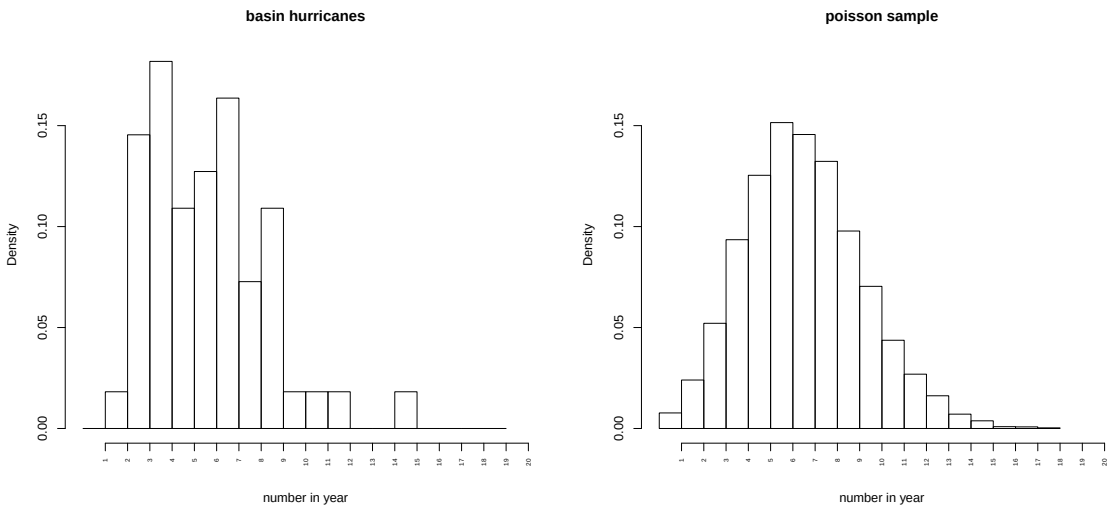


Figure 5.1: Histograms of actual hurricane counts per year (left) and simulated from Poisson distribution (right)

- Assume that a hit on each coastal slot is equally likely, and therefore there is a probability $c = \frac{10}{40}$ that a landfalling storm will hit a major city.

Using the above model the number of city hits is $N_C|N_L \sim binomial(N_L, c)$. Equivalently this is $N_C \sim \text{Poisson}(\lambda qc)$, where λqc is a product of the three parameters described above.

Severity of events The landfall intensity distribution is calculated from the following table (based on HURDAT data from 1955 to 2010). Insurance losses $S(sa)$ are assumed to be directly related to Saffir Simpson categories as shown in table 5.1. The loss column is not based on an analysis of past costs; it is intended to reflect the approximate relationship between category and strength[7]. Clearly one would expect the cost to increase as the strength of storm increases. In practice there is a large variation of insurance loss with storm category. 'Superstorm' Sandy was not even intense enough to be categorised as a hurricane at landfall yet led to an industry loss of USD36bn [248].

[7]This is based on my own personal experience of catastrophe models and Lloyd's disaster scenarios [146].

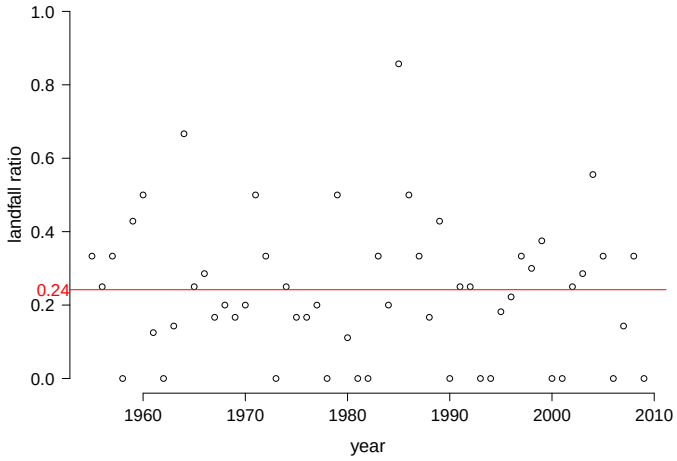


Figure 5.2: Hurricane landfall proportion (ratio) by year since the 1950s based on HURDAT data to 2010. The long term average of 24% is illustrated by the horizontal red line.

Table 5.1: Mapping of Hurricane categories to assumed insurance losses

saffir simpson (sa)	**landfall count since 1955**	**proportion**	**loss S(sa) USDbn**
1	31	38.2 %	1
2	20	24.7 %	3
3	23	28.4 %	15
4	5	6.2 %	70
5	2	2.5 %	130

5.2 Kreps' Pricing using forecasts

The following subsections describe various pricing methods to be investigated. These are all based on the work of Rodney Kreps [128]. Actual pricing methods used by individual insurers and reinsurers are likely more sophisticated but the methods illustrated capture the essence of pricing: to cover expected losses and provide a return on capital to investors that is consistent with the size of the risk taken on. Here the standard deviation term in Kreps' approach is a proxy for riskiness. The price for a claims process with mean μ and standard deviation σ is defined (see equation 2.1 in Kreps' paper) as follows:

$$P = \mu + R\sigma + E \tag{5.1}$$

Where R is the **'Reluctance'** a measure of the insurer's appetite for the given risk (a high reluctance means a low appetite and consequently a high technical price will be set) and E are the expenses of the contract, which are ignored in this chapter.

Naive company Let a **'Naive company'** be defined as one that uses past claims to estimate future premiums and which does not take account of forecasts that use a physical or statistical model.

Naive pricing The calculation of the Naive premium P_0 is derived as follows. First recall that the total claims in the year T_C arise from a compound Poisson process:

$$T_C = \sum_{i=1}^{N_C} S_i \tag{5.2}$$

where $N_C \sim Poisson(\lambda qc)$ and each S_i is IID according to table 5.1. Using the linearity of the expectation operator and the law of total expectation[8]:

$$\mu = E(T_C) = E(N_C).E(S_i) = \lambda qc E(S_i) \tag{5.3}$$

The variance of the total claims can be found by using the Law of Total Variance[9]:

$$\begin{aligned} Var(T_C) &= E(S_i)^2 Var(N_C) + E(N_C)Var(S_i) \\ &= \lambda qc E(S_i^2) \end{aligned} \tag{5.4}$$

The Kreps' formula for the Naive premium P_0 is therefore as follows:

$$\begin{aligned} P_0 &= E(N_C)E(S_i) + R\left(E(S_i)^2 Var(N_C) + E(N_C)Var(S_i)\right)^{\frac{1}{2}} \\ &= \lambda qc E(S_i) + R\left(\lambda qc E(S_i^2)\right)^{\frac{1}{2}} \end{aligned} \tag{5.5}$$

Control Experiment A sample of N_{smpl} claims is produced using the basic simulation above and using the Naive Price (P_0). This will be taken as the 'control' experiment against which other pricing methods will be compared.

[8] $E(Y) = E_X(E(Y|X))$

[9] $Var(Y) = E_X(Var(Y|X)) + Var_X(E(Y|X))$

Variant1: Basin frequency known approximately: change business volume This first variant therefore assumes that the control experiment represents 100% of the company's planned business and that this will be reduced (by a factor α_2) in a year where basin frequency is forecast to be high and increased (by α_1) when the forecast suggests low frequency. This method could be used in a market where the majority of companies still use Naive pricing [139]. It is assumed that the forecaster can say whether the number of Atlantic basin storms are 'high', 'medium' or 'low' compared to an average year (climatology). Define a function f as follows:

$$f(N_B) = \begin{cases} high & N_B > E(N_B) + k.\sigma(N_B) \\ medium & N_B \in [E(N_B) - k.\sigma(N_B), E(N_B) + k.\sigma(N_B)] \\ low & N_B < E(N_B) - k.\sigma(N_B) \end{cases} \tag{5.6}$$

Where $\sigma(X)$ is the standard deviation of the random variable X. f is therefore a three valued random variable. Then the profit π_1 is as follows:

$$\pi_1 = \begin{cases} \frac{1}{(1+\alpha_2)}.(P_0 - \sum_{i=1}^{N_C} S_i) & f(N_B) = high \\ (P_0 - \sum_{i=1}^{N_C} S_i) & f(N_B) = medium \\ (1+\alpha_1).(P_0 - \sum_{i=1}^{N_C} S_i) & f(N_B) = low \end{cases} \tag{5.7}$$

Variant2: Basin frequency known approximately - adjust premium rate For the rest of the variants in this section (unless otherwise indicated) it is assumed that the whole market has now adopted annually revised pricing, taking seasonal forecasts into account. In this case the premium adjustments below would not need to result in a reduction in business volumes. In this variation the price, P_2, is calculated as:

$$P_2 = \begin{cases} P_0(1+\beta_1) & f(_B) = high \\ P_0 & f(N_B) = medium \\ \frac{P_0}{(1+\beta_2)} & f(N_B) = low \end{cases} \tag{5.8}$$

Thus if a high frequency season is forecast the technical premium is increased (by a factor β_1); but is reduced (by β_2) if the frequency is forecasted low.

Variant 3: Frequency in basin known perfectly Here it is assumed that the N_B is known perfectly. In this case $N_C \sim binomial(N_B, q.c)$. Therefore the premium is calculated as:

$$P_3 = q.c.N_B.E(S) + 30\% \left(E(S)^2.q.c.(1-q.c).N_B + q.c.N_B.VAR(S)\right)^{\frac{1}{2}} \tag{5.9}$$

Note in this case that $P_3|N_B$ is a random variable (i.e. varying each year), and that $E(P_3|N_B) \neq P_0$. This is illustrated in figure 5.6.

Variant 4: Number of *landfalling* storms known perfectly In this case the number of landfalls (N_L) is known and therefore $N_C \sim binomial(N_L, c)$. The premium is therefore calculated as:

$$P_4 = c.N_L.E(S) + 30\% \left(E(S)^2.c.(1-c).N_L + c.N_L.VAR(S)\right)^{\frac{1}{2}} \tag{5.10}$$

Variant 5: Severity known approximately In this variant it is assumed (as in variant 4) that the number of landfalling hurricanes is known perfectly. For the landfalling storms an approximate strength category for each landfalling storm is produced enabling the forecaster to calculate a potential loss (L_p) (an upper bound on possible losses assuming all storms are direct hits). From this, three severity grades are published ('high', 'medium', 'low') as follows:

$$g(L_p) = \begin{cases} high & L_p > Q(L_p, k_3) \\ medium & otherwise \\ low & L_p < Q(L_p, k_4) \end{cases} \tag{5.11}$$

Where $Q(L_p, k_i)$ is the k_ith quantile[10] of the potential loss. Given this additional information the company is assumed to adjust pricing as follows:

$$P_5 = \begin{cases} P_4(1+\beta_3) & g(L_p) = high \\ P_4 & g(L_p) = medium \\ \dfrac{P_4}{(1+\beta_4)} & g(L_p) = low \end{cases} \tag{5.12}$$

Note the use of P_4 in the above formula, which already allows for the number of landfalling storms precisely. The additional adjustment allows for the aggregate

[10] $Q(Y,k) = \{y|P(Y \leq y = k\}$

landfall strength in addition. A variant of this approach (call this 5b) could be to use the additional information by scaling line-size, as in variant 1, as follows:

$$\pi_{5b} = \begin{cases} \frac{1}{(1+\alpha_4)}.(P_0 - \sum_{i=1}^{N_C} S_i) & g(L_p) = high \\ (P_0 - \sum_{i=1}^{N_C} S_i) & g(L_p) = medium \\ (1+\alpha_3).(P_0 - \sum_{i=1}^{N_C} S_i) & g(L_p) = low \end{cases} \tag{5.13}$$

Note the role played by P_0 in this case.

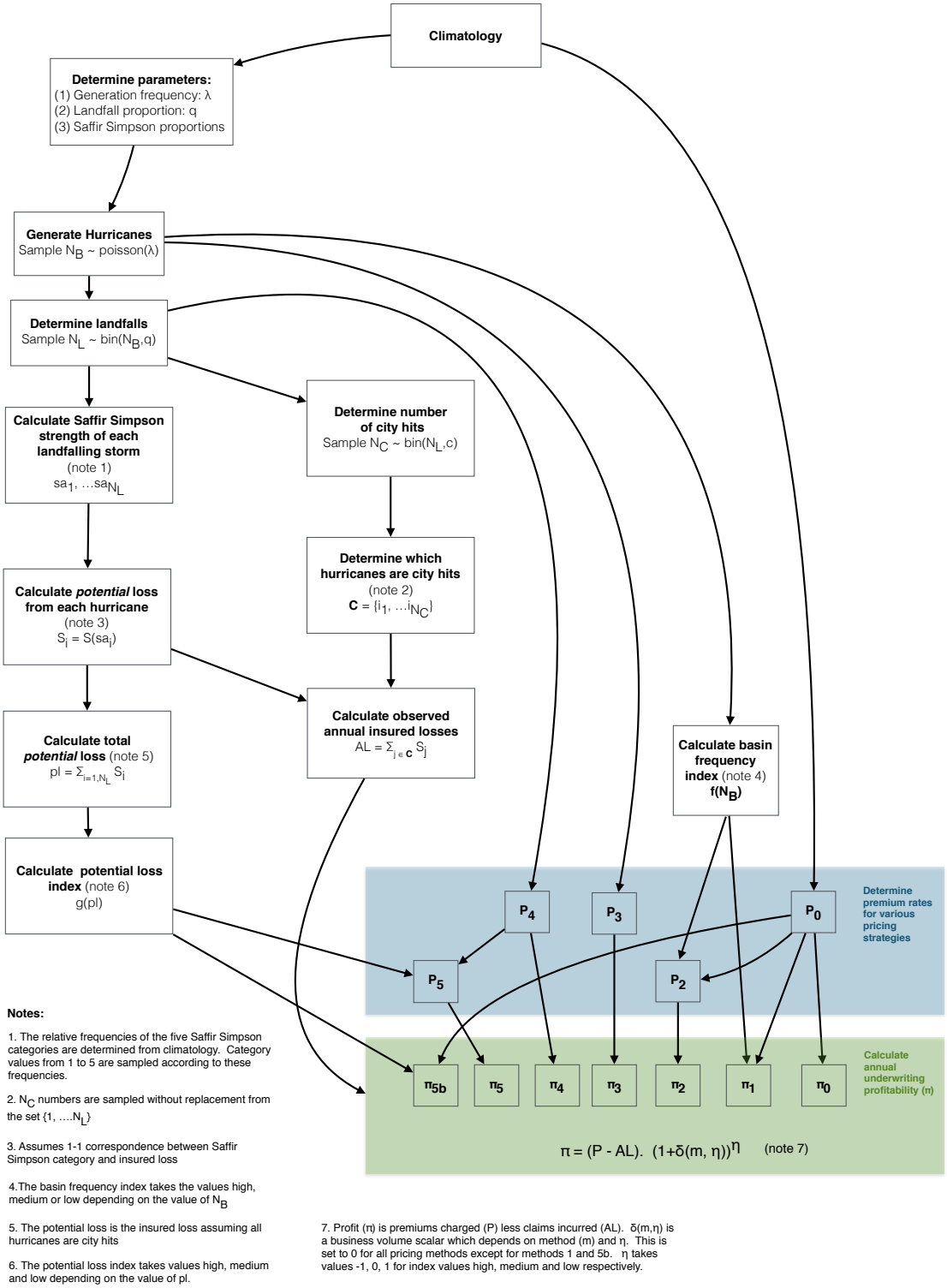


Figure 5.3: Flow chart for stationary climate experiments

Experiment C5.1.x Stationary climate

This is a series of experiments where **x** in 5.1.**x** denotes the pricing variant: 0,1,2,3,4,5,5b.

Sample size: $N_{smpl} = 2^{15}$

Hurricane generation process: $\lambda = 7, q = 0.24, c = \frac{10}{40}$. The assumed frequency of each category of hurricane and the relationship between category and insurance loss is defined in table 5.1.

Reluctance parameter: Kreps' [128] suggested that a figure of around 30% was consistent with reinsurance pricing at the time he published his paper (1990) and this is used throughout this chapter.

Basin frequency index f: : $k = 0.4$

Potential severity index g: For this simulation the following assumptions are made: $k_3 = 0.66$ and $k_4 = 0.33$. These were chosen so there is an equal chance of each category.

Method 1: : $\alpha_1 = \alpha_2 = 0.1$.

Method 2: : $\beta_1 = \beta_2 = 0.1$.

Method 5: : $\beta_3 = \beta_4 = 0.1$. There is an argument that β_3 should be set higher than this to reflect the skewed nature of the severity distribution. But this theoretical consideration has to be offset by the level of premium volatility the market would bear in practice.

Method 5b: : $\alpha_3 = \alpha_4 = 0.1$

5.3 Results: Kreps Pricing

Basin frequency index The choice of $k = 0.4$ in the definition of the basin frequency index f (equation 5.6), means that years with less than 6 hurricanes[11] are referred to as 'low' frequency years, and years with more than 8 are referred to as 'high' frequency. The approach described leads to the following probability table:

[11] $6 \approx 7 - 0.4\sqrt{7}$

basin frequency index (x)	**P(f=x)**
low	30 %
medium	43%
high	27 %

This is roughly symmetric around 'medium' and gives similar chance of a high, medium or low year (it isn't possible to set k for these to be exactly equal because the storm count is an integer). This business volume scaling method is similar to that used by Emmanuel et al [75] who vary the amount of reinsurance purchased depending on a forecast. In both cases the retained business will vary according to the risk. Emmanuel et al conclude 'When the seasonal forecasts are perfect, there is a clear advantage to using them to adjust the amount of risk retained by the primary insurer'.

Exceedance Probability (EP) curve A (very) simple industry catastrophe loss model arises from this process. The simulation of hurricanes through to landfall and insurance loss generates equally likely years of loss. These can be sorted to produce Monte-Carlo estimates of the probability of exceeding losses of a given size. These Exceedance Probability curves (glossary, Chapter 1) are used by the insurance industry to illustrate the levels of risk being assumed. Figure 5.4 shows the annual aggregate losses and their associated probabilities from the simple model described above. Two thirds of years have zero loss (not illustrated). The work presented allows the relative strength of the various pricing approaches discussed below to be compared and general conclusions drawn. Future work could extend this simple model by using a more sophisticated catastrophe model and using a more complex pricing formula. What is novel in this thesis is the framework itself and this initial exploration of it.

Sampling method A Where boxplots or confidence intervals are shown in the following sections they have been created as follows:

1. Given an experiment with a sample of $N_{smpl} = 2^{15}$ outputs;

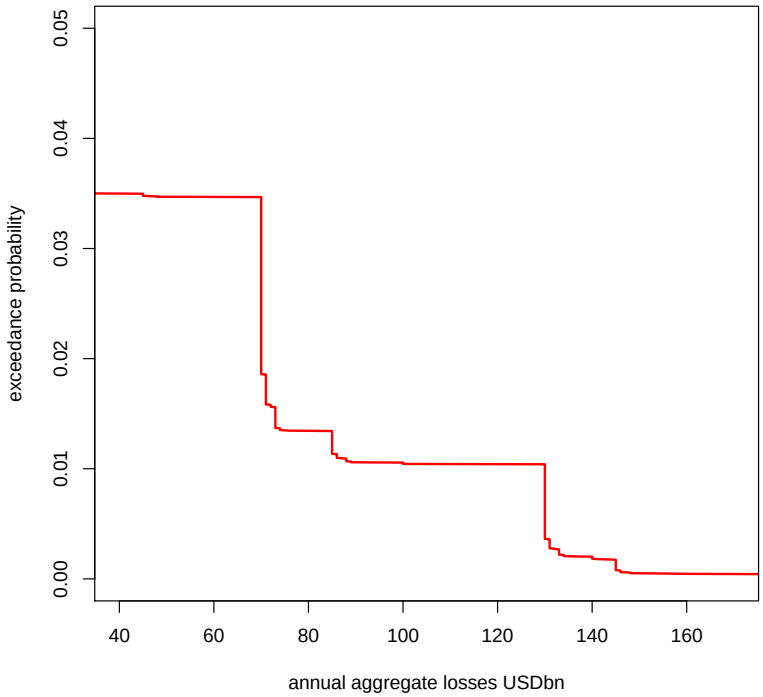


Figure 5.4: Experiment C5.1.x: Exceedance Probability (EP) curve of hurricane losses generated from simple landfall model.

2. Create a sub-sample by selecting $N_{sub} = 2^{14}$ values without replacement (ensuring that the same simulation numbers are chosen for each pricing variant to enable comparison);

3. Produce statistics for the sub-sample(s);

4. Repeat from step 2 $N_{box} = 2^{10}$ times to produce a set of bootstrap results.

Pricing method 0 (Control) The premium $P_0 = 10.88$ is charged in all years. The average profit ($E(\pi_0) = 5.55$) with a 95% confidence interval of (5.36, 5.74) using sampling method A. Profits, (illustrated in figure 5.5) are made in the majority of years with a few years with small losses (i.e. negative profits) and a tiny fraction with losses less than USD50bn arising when major hurricanes are city hits.

Premiums Figure 5.6 shows the various premium levels that arise under the pricing variants (grey). The average premium level is also shown (black). Note that the average premium for variants 3,4 and 5 are all *lower* than the control.

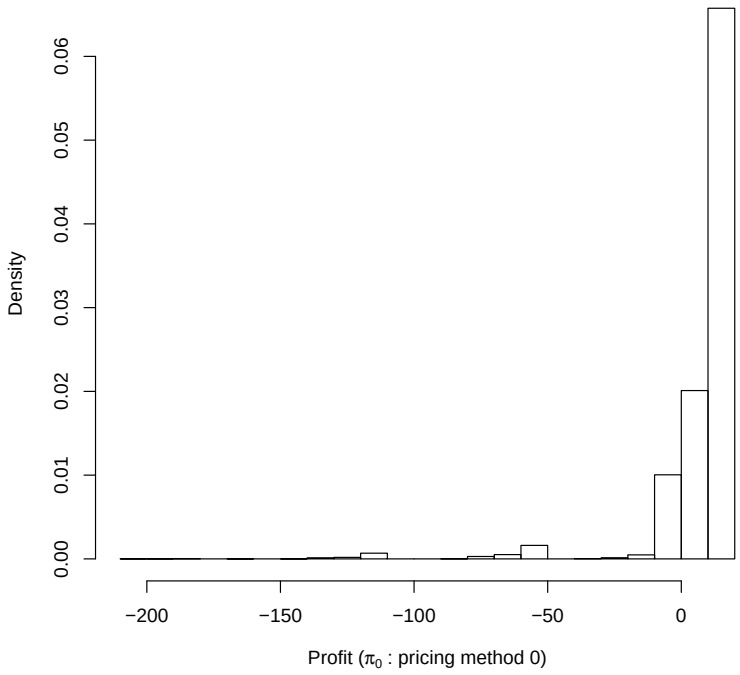


Figure 5.5: Experiment C5.1.0: Histogram of profits π_0 from the control experiment; x-axis shows profit π for pricing method 0 in USDbn.

Explanation why average premiums are lower in some cases The average premium for variants 3,4 and 5 is lower than the control. Consider for example variant 3 (variants 4 and 5 are affected similarly) and recall the premium formula for the control is:

$$P_0 = qc\lambda E(S_i) + R\left(qc\lambda E(S_i)^2 + qc\lambda Var(S_i)\right)^{\frac{1}{2}} \tag{5.14}$$

So,

$$E(P_0) = P_0 = qc\lambda E(S_i) + R\left(qcE(S_i)^2 + qcVar(S_i)\right)^{\frac{1}{2}}.\lambda^{\frac{1}{2}} \tag{5.15}$$

Compare this to,

$$P_3|N_B = qcN_BE(S) + R\left(qc(1-qc)N_BE(S_i)^2 + qcN_BVar(S_i)\right)^{\frac{1}{2}} \tag{5.16}$$

So, since $E(N_B) = \lambda$,

$$E(P_3) = qc\lambda.E(S_i) + R\left(qc(1-qc)E(S_i)^2 + qcVar(S_i)\right)^{\frac{1}{2}}.E({N_B}^{\frac{1}{2}}) \tag{5.17}$$

For the chosen parameters the term $qc\lambda.E(S_i) = 5.42$. This term is the same for *both* expectations and the term involving $E(S_i)^2$ is clearly lower for (P_3) (due to the

$(1 - qc)$ term). In the specific simulation:

$$R\left(qcE(S_i)^2 + qcVAR(S_i)\right)^{\frac{1}{2}} = 2.061 \tag{5.18}$$

compared to

$$R\left(qc(1-qc)E(S_i)^2 + qcVar(S_i)\right)^{\frac{1}{2}} = 2.047 \tag{5.19}$$

In the particular simulation $E(N_B^{\frac{1}{2}}) = 2.59$ compared to $E(N_B)^{\frac{1}{2}} = 2.64$ but it is generally true that $E(N_B^{\frac{1}{2}}) < E(N_B)^{\frac{1}{2}}$ due to Jensen's inequality and the fact that the square root function is concave. In the particular simulation, therefore:

$$E(P_0) = 5.42 + 2.061 * 2.65 = 10.88 \tag{5.20}$$

compared to

$$E(P_3) = 5.42 + 2.047 * 2.59 = 10.72 \tag{5.21}$$

Therefore it is clear that under variant 3 a lower premium is calculated because the Kreps formula gives credit for the lower standard deviation. The reduced risk is passed straight to the policyholder. In practice the insurer may wish to retain some profit to avoid a price that is less than the Naive price.

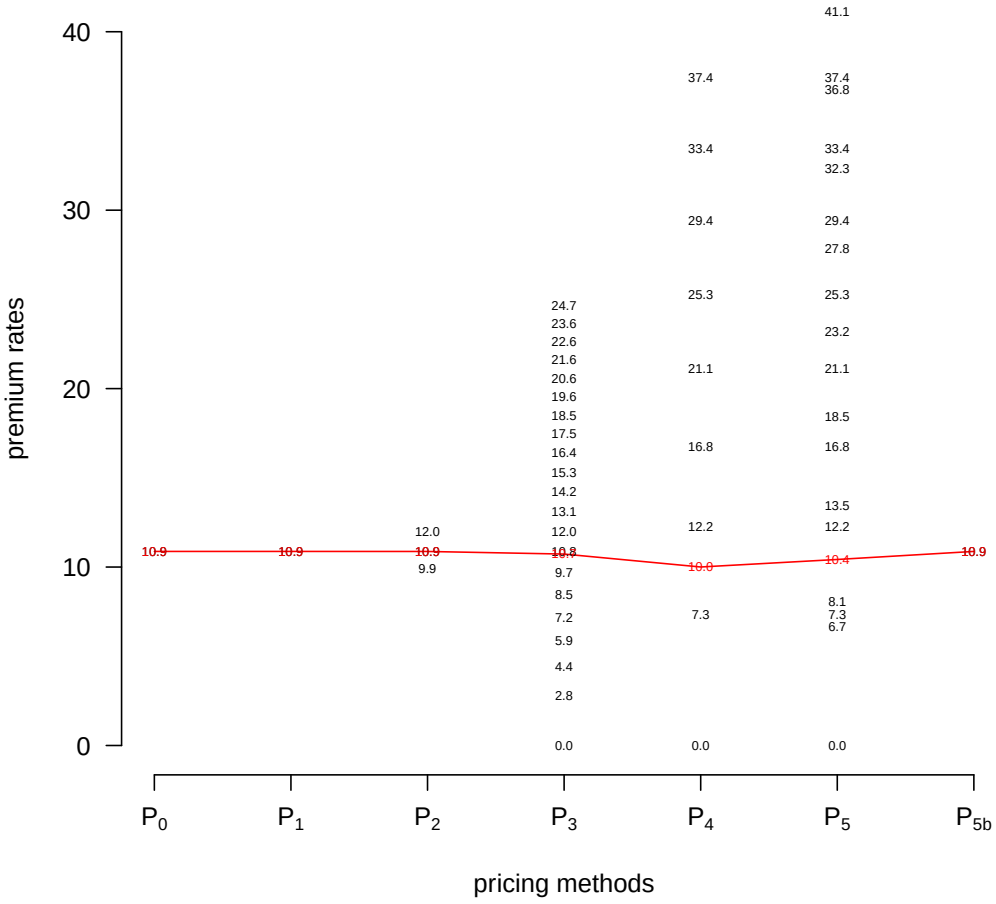


Figure 5.6: Experiment C5.1.x: Spread of premium rates arising from the different pricing methods indicated by the label on the x-axis, values are shown in black with the premium rate in text plotted at the appropriate level). The mean premium under each method is highlighted in red with a line joining them (this may not be a premium level that is ever charged). Note that the average premium rates for variants 4 and 5 are lower than the control.

Premium relative to the number of basin hurricanes Under some of the pricing variants the premium will be different depending on the number of basin hurricanes that are forecasted. Figure 5.7 illustrates how the premium varies. The key points from this graphic are as follows:

- by construction the control (0) and variants 1 and 5b are all equal and show a constant premium in all cases;
- variant 2 shows a premium which is higher in a 'high' activity year and lower in a 'low' activity year, by design the adjustment level is modest;
- variant 3 shows exactly one premium level for each number of Basin hurricanes; as expected this increases as the hurricane count increases. Note also that a premium of zero will be charged if no hurricanes are forecast (recall the forecast in this variant is deemed to be 'perfect');
- variants 4 and 5 produce a wide variation of premium for each number of basin hurricanes, as they use more information about landfall rates and severity (and this introduces variability).

Premium relative to the number of landfalling hurricanes Key points from figure 5.8 are as follows:

- The dots form a triangular shape in the plot for variant 3 with white space below the hypotenuse. Note, that should a particular number of landfalling hurricanes occur (n_L say) then at least that number of Basin storms *must* have occurred (i.e. $N_B \geq n_L$ in the pricing formula 5.9). The other terms in the pricing formula are constants and so the lower bound on N_B is also a lower bound on the premium rate, explaining the white space. There is, however, a wide spread of premiums for each landfall number because this pricing method is based on the number of storms generated and the number of landfalls can still vary significantly;
- variant 4 now shows a 1 to 1 relationship between landfalling hurricane count and price (by design) - the more landfalling storms the higher the premium.
- variant 5 also shows an increasing premium for landfalling storm count (as expected), however there is still a spread of premium rates reflecting the variation in severity each season which this variant allows for.

Premium relative to the number of city hits Figure 5.9 shows the premium charged under each of the variants plotted against the number of city hits that arose in the simulated year. The key points arising from this figure are as follows:

- The graphics illustrate that even the most sophisticated methods (4 and 5) produce a wide variation of premium levels for a given number of city hits. The residual uncertainty in the system is still significant and this can lead to a *lower* premium being charged than the control in a year with a city hit. When there are more than 2 city hits, however, the premium charged is never lower than the control.
- The blank lower right triangles for variants 3,4 and 5 arise because a given number of city hits places a lower bound on basin hurricane numbers and landfall numbers - and hence a lower bound on premiums.

Profitability of the variants The boxplot in figure 5.10 shows the mean underwriting profit, relative to the control, for the variants. Key points from this graphic are:

- there is still a wide range of profitability from different samples - even when they are of size 2^{14};
- variants 1 and 5b, are usually more successful than the control. On average variant 1 gives profits that are 2.7% higher and variant 5b 7.9% higher;
- variant 2 is of broadly equal profitability to the control (with sampling error of +/- 5%);
- variants 3,4 and 5 are all *lower* profitability than the control.

The slightly odd outcome noted in the final bullet is explained by the fact that the premium from variants 3,4 and 5 is lower, on average, than for the control. Average underwriting profit is: $E(\pi_j) = E(P_j - \sum_{i=1}^{n_C} S_i)$, for variant j. But, using the additive property of expectations, this is just $E(\pi_j) = E(P_j) - ASL$ where ASL is the 'Average Season Loss' which, crucially, is the same for all variants. So $E(\pi_j) = E(P_j - P_0 + P_0) - ASL = E(\pi_0) + E(P_j) - E(P_0)$ and since $E(P_j) < E(P_0)$ for variants 3,4 and 5 this explains the effect.

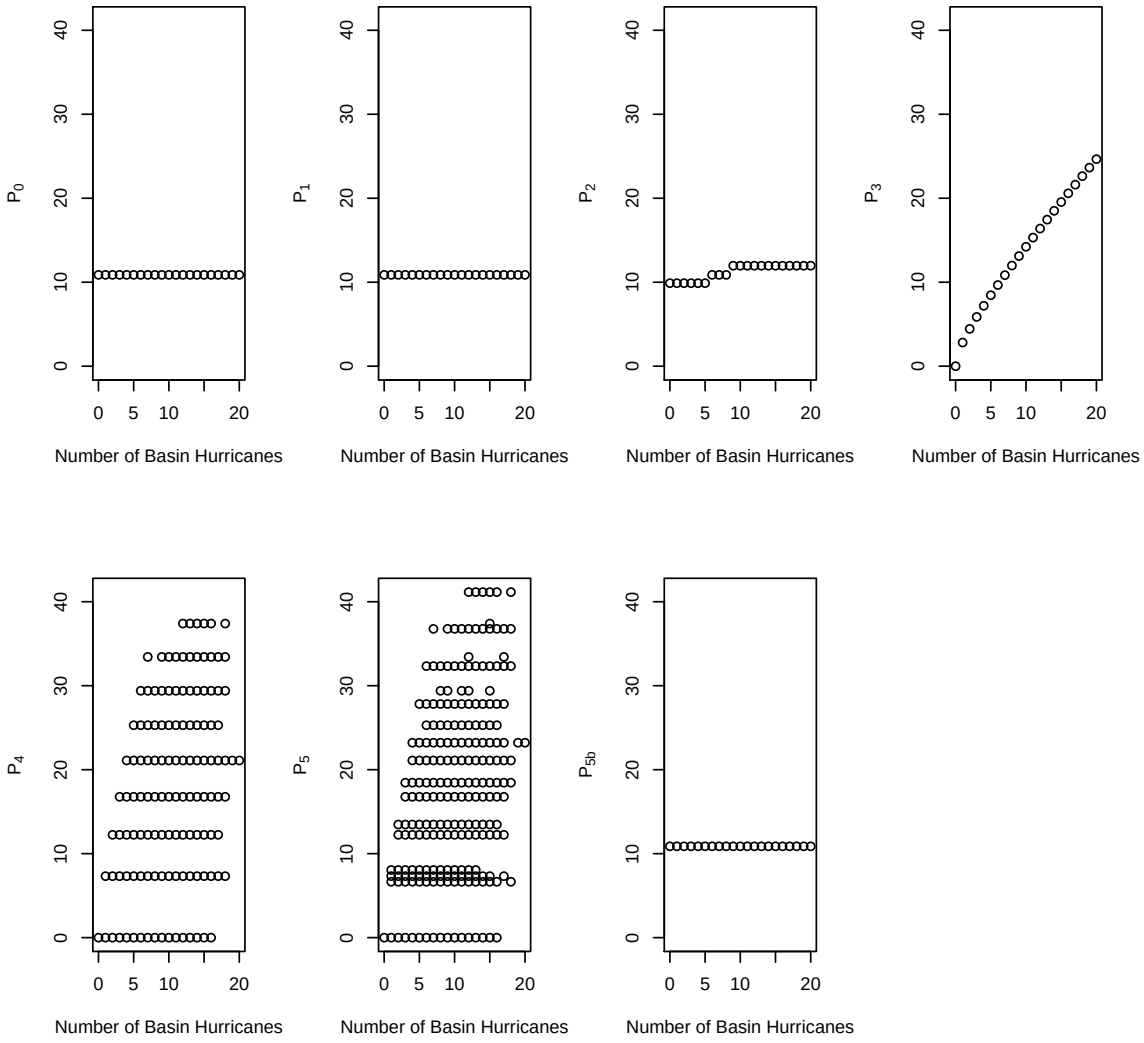


Figure 5.7: Experiment C5.1.x: Premium rates (y-axis) against number of Atlantic Basin hurricanes (x-axis) for each pricing method. The pricing method is indicated by the y-axis label. The control P_0 and variants 1 and 5 charge the same premium rate in all cases (the latter methods scale the volume of business sold). P_1 shows three levels corresponding to the low, medium and high seasons. P_2 by construction has a 1-1 correspondence with the number of basin hurricanes. Variants 4 and 5 show many different rates against number of basin hurricanes since they take account of more forecast information.

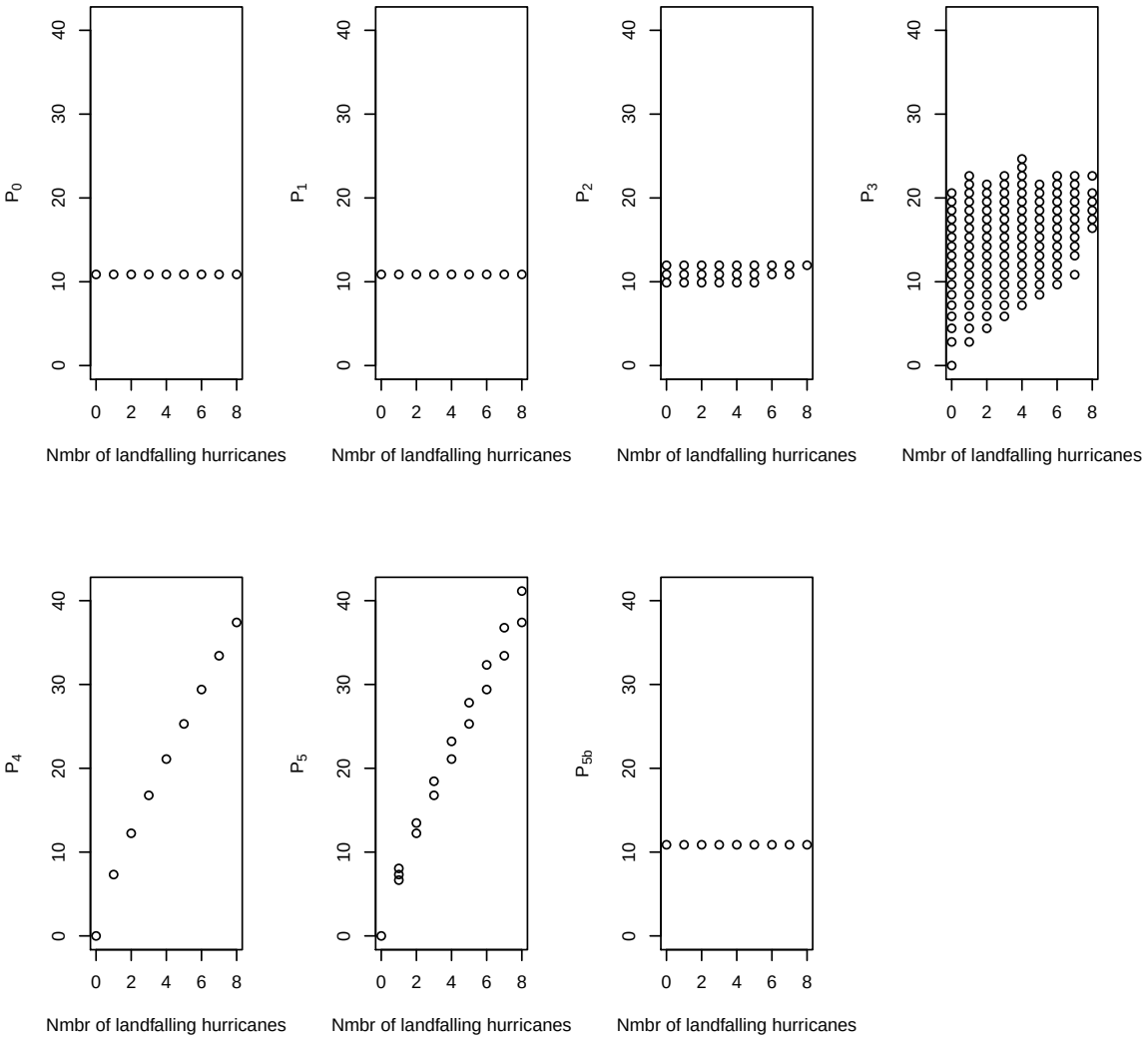


Figure 5.8: Experiment C5.1.x: Premium rates (y-axis) against number of landfalling hurricanes (x-axis) for each pricing method. The pricing method is indicated by the y-axis label. Variant 2 has three premium levels except when the number exceeds 5; in this case it is not possible for the basin season to have been 'low'. Variant 3 has many premium rates but the landfalling number places a lower bound on the basin frequency explaining the white space at the bottom right of the plot. By construction Variant 4 shows a 1-1 relationship with landfalling number. Variant 5 shows three rates (low, medium and high season strength) when the landfalling number is 1; for larger landfalling counts only medium and high strength seasons are possible.

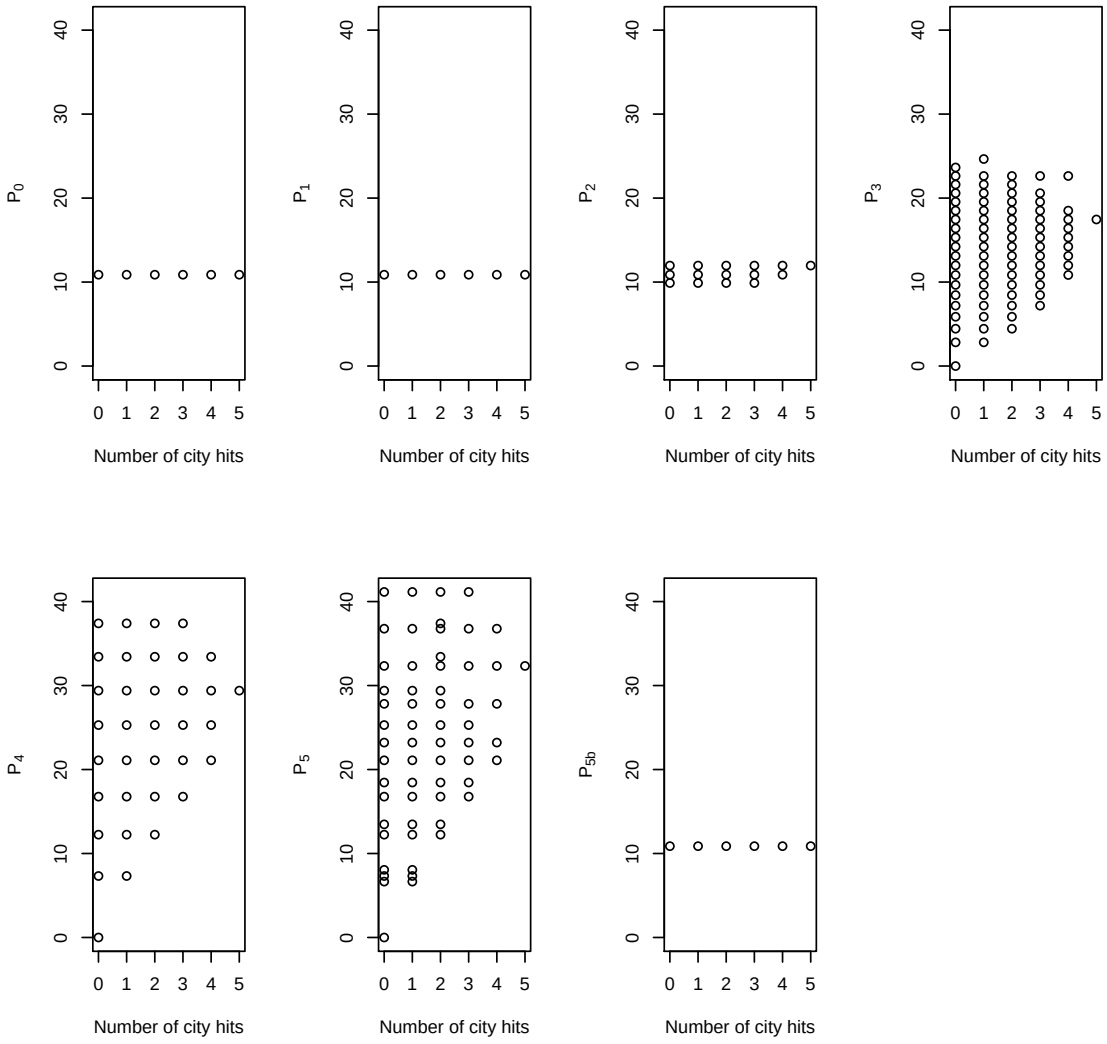


Figure 5.9: Experiment C5.1.x: Premium rates (y-axis) against number of city hits (x-axis) for each pricing method. The pricing method is indicated by the y-axis label. The white space in the bottom right of the figures for variants 3,4 and 5 are caused by the number of city hits placing a lower bound on the basin and landfalling frequencies respectively. Note that under variant 4, when there is 1 city hit it is possible for this variant to have charged less than the control premium rate.

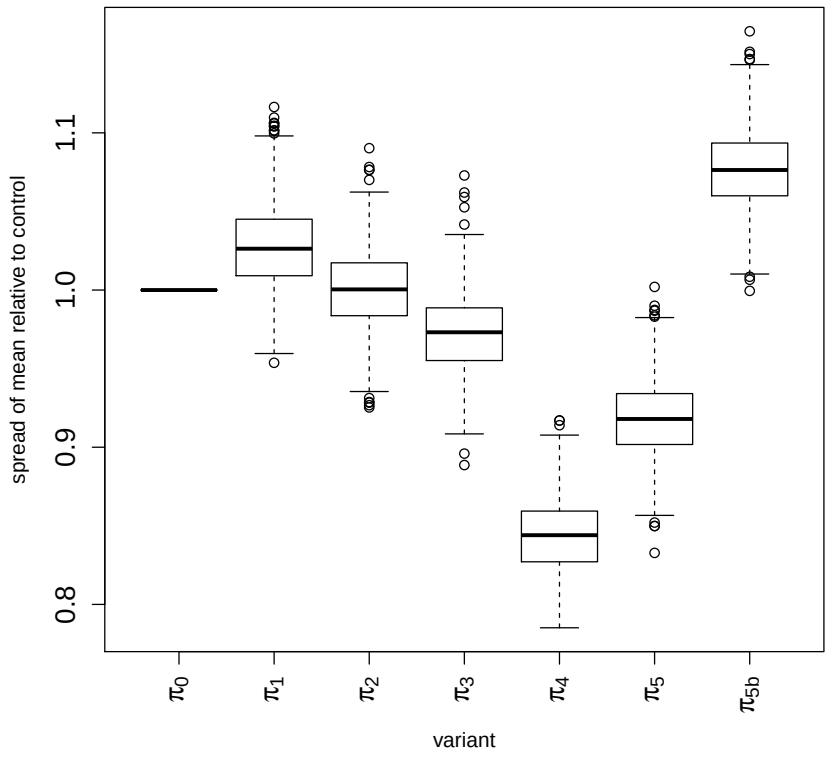


Figure 5.10: Experiment C5.1.x: Boxplot of mean underwriting profit relative to the control (π_0). The mean profit ($E(\pi)$) of each resample is calculated for each pricing method and then divided by $E(\pi_0)$ from the control method. The profitability of variant 4 (π_4), where the number of landfalling storms is known perfectly, is significantly lower than the others despite the additional information used. The profitability of Variant 5b (π_{5b}) which makes use solely of the season strength information is significantly higher than the others. Boxplots produced using sampling method A (taking 2^{10} bootstrap resamples each of size 2^{14} from the 2^{15} underwriting results produced by simulation).

Impact on extreme profitability Define **'extreme conditions'** as negative profits that arise less frequently than 1 in 200 years (0.5%). Figure 5.11 shows the insurer's (negative) profitability, as a proportion of the profitability of the control experiment, in extreme conditions.

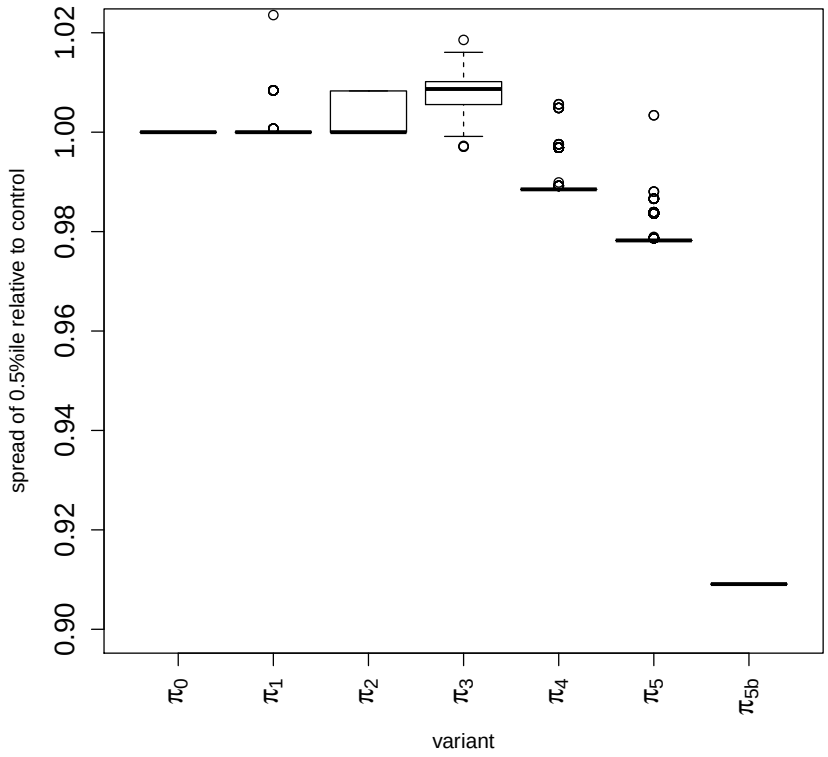


Figure 5.11: Experiment C5.1.x: Boxplot of 1-in-200 negative profit ($Q(\pi, 0.005)$) for each pricing method relative to the control. The quantile for each pricing method is divided by the value for the control experiment. Values greater than 1 indicate the 1 in 200 underwriting loss will be worse under the pricing variant and values less than 1 indicate a better outcome. Each of Variants 4,5 and 5b have significantly lower extreme negative profits than the other methods. Variant 5b (π_{5b}) in particular achieves a 9% reduction in extreme negative profits relative to the control. Created using sampling method A (taking 2^{10} bootstrap resamples each of size 2^{14} from the 2^{15} underwriting results produced by simulation).

Key points from this graphic are:

- Variants 4, 5 and 5b use successively more forecasting information, and have lower extreme negative profits than the control;
- Variants 1,2 and 3 lead to a similar level of negative profitability to the control in extreme conditions;

- Business volume reduction methodology 5b is both more profitable (7.9%) (figure 5.10) and avoids large extreme negative profits (-9.1%) (figure 5.11).

Impact on capital Assume that the insurer is required to hold sufficient capital (assets in excess of reserves) to withstand extreme conditions[12]. Key findings are:

- Control: The company treats each season the same as any other, so holds the same amount of capital in all cases. Under the control method a premium of 10.9 is charged and held as a reserve. The 1 in 200 claim amount is 130.0 and hence the required capital is 119.1.
- Variant 1: the company reduces its exposure by 10% in a high frequency year and increases it by 10% in a low frequency year. It charges the same premium per risk as the control - but both the total aggregate premium and 1 in 200 claim will scale in the same way. Therefore the total capital required will be the same as the control in a medium year, it will be 10% lower in a high year and 10% higher in a low year. The capital per unit of risk is, however, the same in all cases.
- Variant 2: The company receives a high, medium or low basin activity forecast and will charge a differential premium in this case (for the same level of exposure as the control experiment). Figure 5.12(a) shows that the 1 in 200 claim is actually the same in the 'low' and 'medium' case even though the premium is not; hence the capital held actually falls in the 'medium' case as more premium is charged. The level of risk has certainly increased in the medium frequency season, but this doesn't show up until around the 1 in 500 year event. Clearly this is an artefact of the particular assumptions made in this chapter, but it does underscore an important point: whatever regulatory threshold is chosen it is likely that there will be features that appear anomalous. In this case the company will hold lower capital in a riskier year.
- Variant 3: The company knows the number of basin storms perfectly. The regulator requires the capital to flex with the level of risk. Figure 5.12(b) shows that the 1 in 200 level of claim has regions of stability (e.g. when the number of basin storms (n) is between 4 and 8 inclusive) - however the pricing

[12]This is (in essence) the current requirement under the Solvency II framework.

method is linked to the increasing variance of loss which does increase as n increases. Hence the required capital actually falls during these stable zones in line with the increase in premium. After a step up in risk the capital can still fall if the premium increase is greater than the rise in 1 in 200 claim - but in other cases it rises when the level of extreme claims outpaces the premiums. Capital held for $N_B < 4$ is much lower - it is interesting to note that in this case the company could not survive a single cat 5 storm - at this point the regulator might step in and override the 1 in 200 rule and require additional capital to be held.

- Variant 4: The company knows the number of landfalling storms perfectly. Whenever N_L is zero they would hold zero capital. The company will make a certain profit in the year equal to the premium charged. In practice, however, if the forecast really were perfect it is likely this would be generally known and so no insurance would be bought. Figure 5.12(c) illustrates the situation where the number of landfalls is non-zero. Capital has the typical saw tooth pattern initially as the the premium charged sometimes increases faster than the 1 in 200 claim. Eventually the capital increases monotonically in cases where the number of landfalls is very large (these are very rare) - in this case the 1 in 200 claim increases by more than the premium - even though the conditional premium is much higher. Note, however, that the high capital years are extremely rare; on average capital will be lower for this method as shown in figure 5.11.
- Variant 5: Figure 5.12(d) illustrates an interesting feature of the conditional season strength metric chosen. If the number of landfalls is more than 2 but the season is of medium strength then the forecasted hurricanes can only be category 1 or 2 since otherwise the loss will exceed the quantile for the function g. Hence actually the 1 in 200 claim falls even though the number of landfalls has risen (because the company knows the overall season strength is low or medium). In this case the company never needs to hold capital in low or medium years and will make certain profits if the forecast is perfect. Figure 5.12 (e) shows that for this knowledgeable company the capital rises very high in the rare case of multiple landfalls.

- Method: 5b Figure 5.12(f) illustrates the results for a company using method 5b. In a low severity season the 1 in 200 loss is 2 units. The premium charged is based on the control premium however and so exceeds this level, losses would be made very rarely in this situation. Hence theoretically capital requirements are negative! In practice the regulator would not allow a company to write business with no capital held. For a high severity season the capital held is still lower than the control because the premium charged is higher.

5.4 Pricing to reflect target return on capital

The Kreps' formula does not directly reference the capital required to write the business; the σ term is a proxy for this. It is possible (as with equation 4.10) to adapt the pricing formula to target an expected return on capital directly. Recall, that the Capital held (κ) is defined as a regulatory claim amount (Q_j) less the premium charged ($\tilde{P}_j$) for $j \in \{0, 1, 2, 3, 4, 5, 5b\}$ i.e. $\kappa_j = Q_j - \tilde{P}_j$. Then alternative premium rates (shown by a tilde), referred to as the 'target return' method below, are defined as:

$$\tilde{P}_j = E_j(C) + \gamma\kappa_j \tag{5.22}$$

or equivalently:

$$\tilde{P}_j = \frac{E_j(C) + \gamma Q_j}{1 + \gamma} \tag{5.23}$$

Where E_j is the expectation allowing for the forecast and similarly Q_j is the regulatory quantile dependant on the forecast.

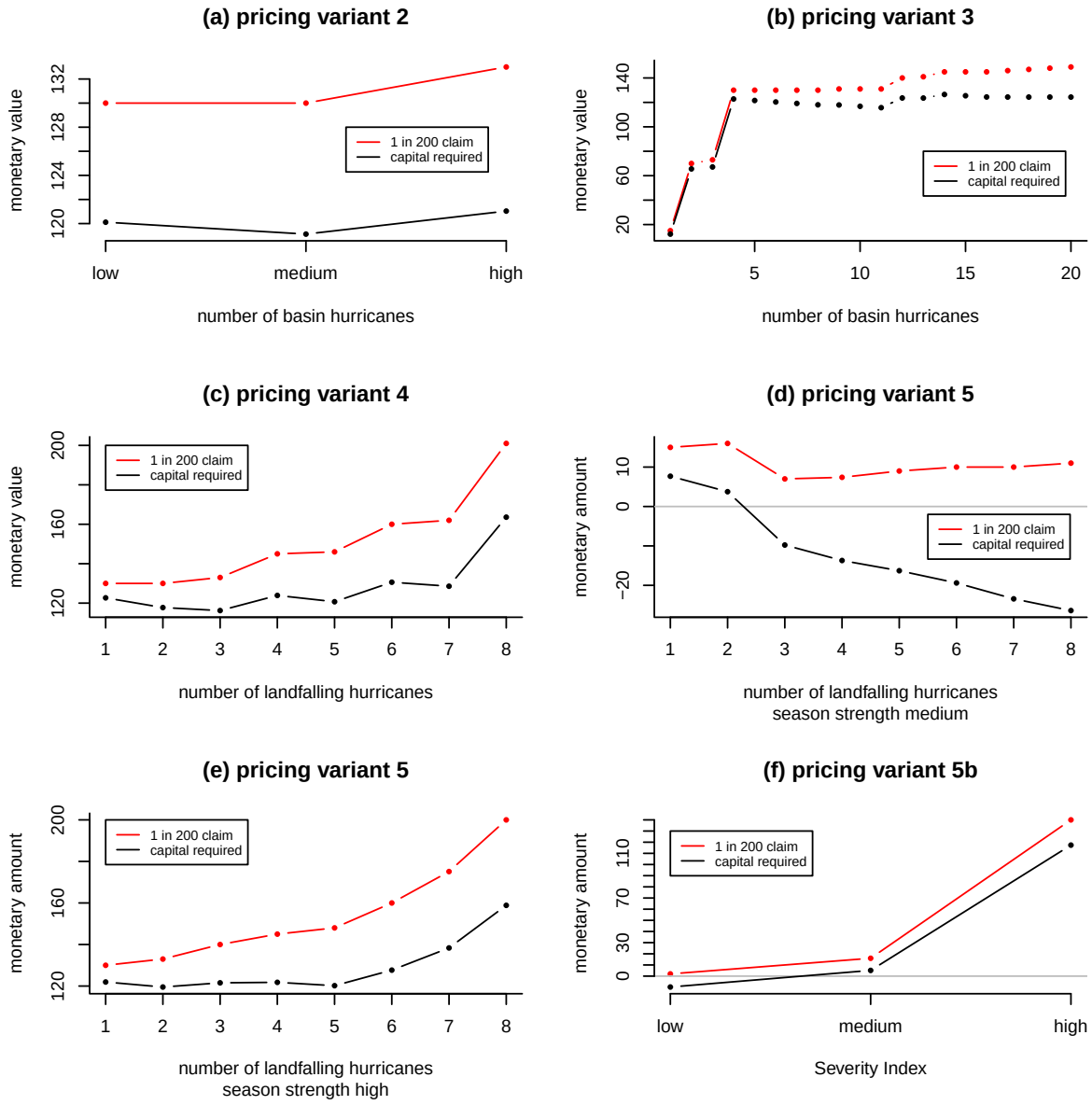


Figure 5.12: Experiment C5.1.2- C5.1.5b: Capital requirements arising under each method. Red lines show the 1 in 200 annual aggregate claim arising black shows the capital arising after deduction of the premium charged from the 1 in 200 claim. Figure (b), Variant 3 shows that when the number of basin hurricanes is 1 or 2 the chances of a category 4 or 5 storm making landfall as a city hit is beyond a 1 in 200 probability so that capital does not (in theory) need to be held for this eventuality. Figure (4) Variant 5 shows that if the season strength is medium and there are more than 3 landfalling storms then no capital is required as the premium rates are more than sufficient. In practice no insurance would be bought under these conditions on the assumption that the forecasts are exact.

Experiment C5.2.x Target return method

All parameters as for experiment C5.1.x except that pricing formula targets a chosen return on capital.

Target Return on capital: The return on capital is chosen so that the premium charged by a company using naive methods will be the same. In the situation when forecasts are not used, the 1 in 200 claim is kept fixed at 130 in each year. Recall that expected losses are 5.42 and the Naive premium (P_0) was 10.88. Hence the required return on capital to equate the more sophisticated method with the Kreps method is derived as:

$$\gamma = \frac{P_0 - E(N_C)E(S_i)}{Q_0 - P_0} = \frac{10.88 - 5.42}{130 - 10.88} = 4.58\% \tag{5.24}$$

Results for experiment C5.2.x The following series of figures compares the premium for pricing variant j under the simple Kreps method (P_j) with the target return method ($\tilde{P}_j$). Variants 1,2 and 5b are not shown since (by design) the premium is the same for the Kreps and Target Return methods. Pricing variants 3,4 and 5 will give different premium rates. Each shows an increasing monotonic relationship between the premium rates as the relevant storm number (and, where relevant, storm strength).

- Figure 5.13(a) illustrates variant 3 and shows that when the hurricane storm count is close to the expected count (i.e. 6,7,8) then the two methods show similar pricing. This isn't that surprising since the target return on capital is forced to give the same premium for the naive pricing method which uses expectations for each part of the process. For $N_B > 7$ the Kreps method produces higher premium rates than the Target Return method.
- Figure 5.13(b) illustrates variant 4 and shows that when the landfall number is close (i.e. 1 and 2) to the expected number (1.7) then the premium rates are similar for the two methods; for similar reasons noted for variant 3. Above 2 landfalling storms the Kreps method produces a higher price than the target return method.
- Figure 5.13(c), illustrates variant 5 and shows similar results - for number of landfalls in excess of two storms the Kreps method shows higher prices. Since

the price is a multiple of the method 4 prices this is quite obvious.

It is not immediately obvious which of the Kreps and Target Return methods will be the most profitable. Although the Kreps method gives higher prices when storm counts are high, such situations are the rarest. In fact in all cases the Kreps method gives higher profitability and has a lower coefficient of variation[13] (CoV) as shown in the table.

Table 5.2: Comparison of profitability between Kreps and Target Return Methods

Pricing method	**Kreps mean (CoV)**	**target return mean (CoV)**
3	5.40 (3.32)	5.37 (3.34)
4	4.68 (3.79)	4.54 (3.90)
5	5.10 (3.47)	4.90 (3.59)

Career performance An underwriter might be thought to have done reasonably well in their career if they can return the expected level of return on capital, on average, over its course. Define a career to be 40 years. It is then simple to calculate the simulated probability of an underwriter achieving this. The method described in this chapter produces 2^{15} simulated years' of underwriting profits; hence these can be split into 819 separate 'careers' (with 8 years to spare). Consider two underwriters, one, underwriter A, that charges the Kreps' price P_0 and another, underwriter B, which undercharges by 10% due to misstatement of the level of risk (i.e. their processes are faulty and they have made an error).

For underwriter A there is a 44.2% chance they will achieve a geometric average return equal to their desired return on capital ($\gamma = 4.58\%$p.a.) or better. Underwriter B has a 33.1% chance of achieving such a career average return. Two situations can therefore occur:

1. an underwriter who is pricing correctly still has a very high probability of returning a less than average return in their career; and

[13]The Coefficient of Variation of a random variable X is defined as $CoV(X) = \frac{sd(X)}{E(X)}$

2. an underwriter who is pricing incorrectly still has a reasonable probability of appearing to provide a decent return over their whole career.

In the first case Underwriter A may be considered to have 'failed' when they did nothing wrong; in the second case Underwriter B may be thought to be a highly successful underwriter when in fact they were underpricing for their entire career. This finding is similar to that in Emmanuel et al [75] who find it can take 10 years to differentiate between a company using past claims averages and a more sophisticated one that uses forecasts, even in the case of a perfect forecast. In practice EP curves are not as extreme as this simple example but this illustrates the difficulties of assessing underwriter performance in the presences of fat tailed loss distributions.

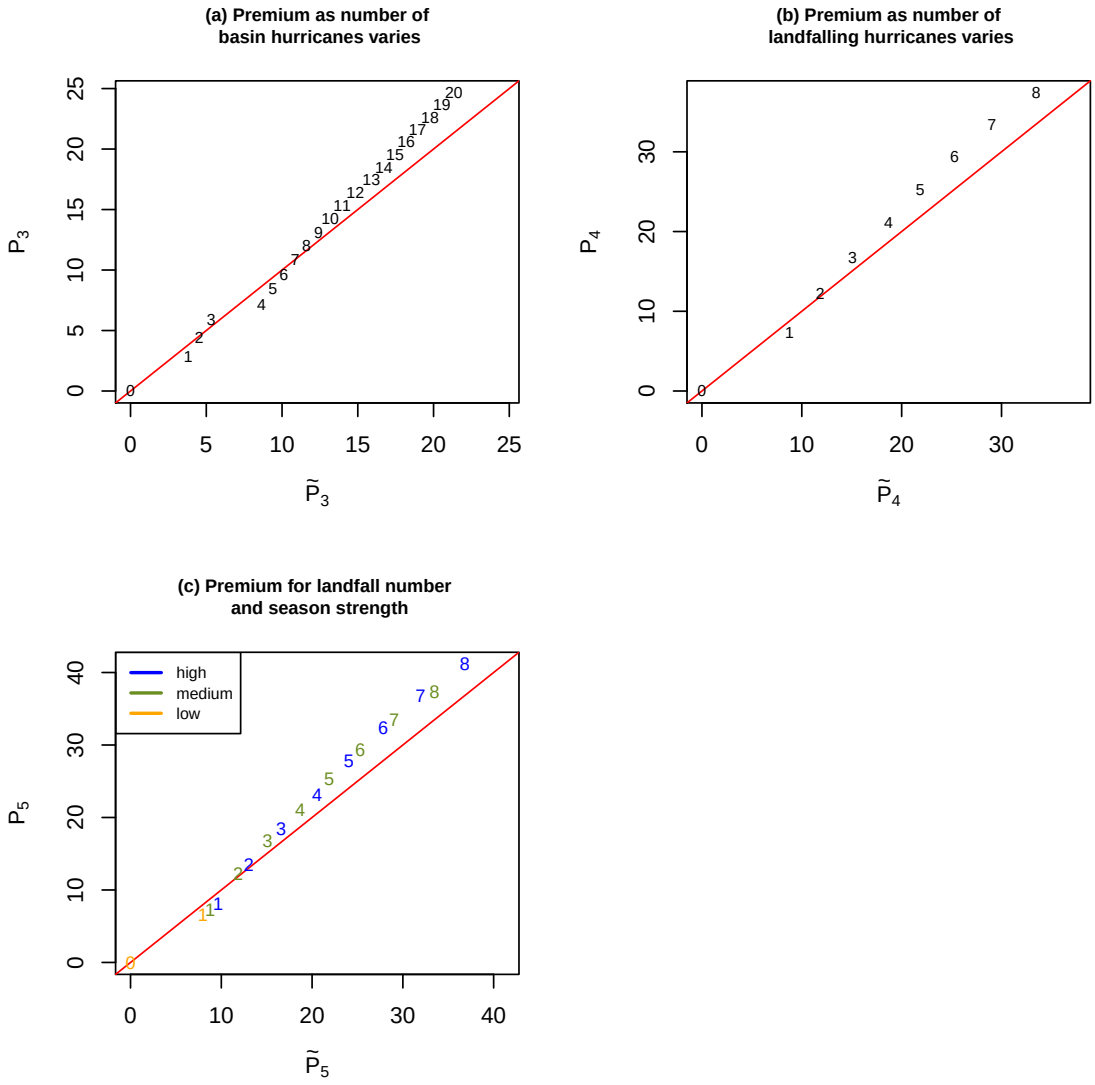


Figure 5.13: Experiment C5.2.3 - C5.2.5. Comparison of premium rates using the Kreps method (P_j on the y-axis) and the Target Return method ($\tilde{P}_j$ on the x-axis) for variants $j \in \{3, 4, 5\}$. The line $y = x$ is shown in red for easy comparison. Figure (a) The plot character is the number of basin hurricanes in the year (noting that there will be one premium rate for each forecast of basin hurricane numbers under variant 3), figure (b) the plot character is the number of landfalling hurricanes in the year (noting that there will be one premium rate for each level of forecasted landfalling number under variant 4) and figure (c) shows variant 5 the plot character is the number of landfalling hurricanes in the year, the plot colour shows the severity strength of the season - the premium rate is sensitive to both these parameters so there is a unique premium rate for each combination of landfall number and season strength.

5.5 Conclusions

The results in this chapter are based on a novel demonstration of a simple yet informative model with qualitative conclusions that are expected to generalise. The model is not intended to produce 'realistic' prices; rather conclusions about the efficacy of forecasting methods. The results of the various investigations suggest that there is value in forecasting information if carefully used.

In the case where the frequency distribution is known perfectly a simple business volume scaling method (where exposure is reduced in high forecast activity years and increased in low forecast activity years) is the most successful compared to the other pricing methods in terms of maximising underwriting profits. Business volume reduction may be an appropriate method for a sophisticated reinsurer, however, but does not work for the insurance industry as a whole unless deductibles are also variable for direct writers[14]. In this experiment, it is found that an underwriter who undercharges systematically can still appear successful with 33% probability showing that mis-pricing can be difficult to detect in the presence of extreme risks.

Pricing variants which vary the premium rates in line with forecast information produce lower premiums on average in the experiments described and therefore lead to lower expected profitability. This is because the benefits of reduced variance (caused by the availability of more information) are passed straight to policyholders. To preserve average profitability a minimum premium could be set equal to that of the naive methods; i.e. only charge more premium in high frequency years but retain the minimum premium in lower frequency years to reflect the high residual volatility. The variable premium methods tested require a few percentage points less capital on average. In high frequency years companies recognising the increased risk would, however, be required to hold more capital, in some cases much more. Unless this practice were widely adopted this may be hard to justify to investors. The next chapter addresses the issue of competition directly by introducing a novel insurance industry model with two competing companies.

[14]Direct writers are insurers that sell directly to the public or businesses; this contrasts with Reinsurers that sell to insurers.

Chapter 6

The insurance industry in-silico

'***With some advice and encouragement from his fellow student and friend Walter Newlyn ... [Phillips] set about designing a machine in which water flowed through transparent pipes and/or gathered in reservoirs, to demonstrate the macro-economy's properties. He described this machine, and the economic interpretation of its workings, in his first publication ... Phillips actually built his machine, and it worked too...***' David Laidler 2000 [130]

The previous chapters have illustrated that forecasting can be useful in an insurance context. The insurance setting in each case has, however, been highly over simplified. This chapter aims to partially redress the balance by describing a model of an insurance industry with many of the key features present in the real world. The complexity of computer modelling in the insurance industry has increased over the past 40 years since models were first used to explore better ways of calculating solvency risks [61,63,218,219,244]. Such models have tended to focus on single companies [50,195,220], however, and it is far rarer to find analysis of insurance markets in the literature [250, 277]. This chapter proposes a novel insurance market model which includes competition between two companies and captures the main processes in the insurance industry [62, 105, 198, 250]: customer loyalty, premium rating, capital setting, investment returns, claims payments and dividend payments. Like the subject of the opening quotation, the MONIAC[1] model of the UK economy based on fluid flow, created by Bill Phillips in 1949 [24,210], the insurance model presented in this chapter is a simplification of a highly complex industry but one that is intended

[1]Monetary National Income Analogue Computer

to give useful insight to insurance relevant questions. As with reality, each company in the model has a limited claims history from which to estimate premium rates and capital requirements [53]. The pricing method and underlying assumptions used by the two companies is specified in advance after which the behaviour and outcomes emerge from the simulation. The model provides a framework to investigate such questions as: will the companies survive for as long as specified by the regulatory test, on average? Does profit sharing (or 'payback'), which is used in some re-insurance markets, lead to a better or worse investment for shareholders? Would a different regulatory method lead to a stronger or weaker industry? Is a company always better off if it sets premium rates using the same distribution family from which the claims arise? These questions are answered, sometimes with, initially, surprising results.

The pricing method used in this chapter is 'climatology' pricing similar to that described in Chapter 4. Future work could combine the methods from Chapters 4 and 5 by introducing exogenous forecasts which the insurers in the model could use. The original elements of this chapter are believed to be:

- Presentation of a novel, agent based insurance market model, and the use of the model in the following investigations;
- Quantitative investigation of efficacy of a payback rule, with implications for profitability and average company lifetime;
- Investigations into the impact of different regulatory regimes on average company lifetime and other metrics: (1) varying Value at Risk (VaR) thresholds and (2) the impact of an equivalent TVaR rule (defined in equations 6.6 and 6.7);
- Quantifying the impact of using incorrect pricing distributions on key metrics;
- Investigation of the impact of different claims distributions as a proxy for different business mix.

6.1 Model Design

This section explains how the competition model has been specified. Ultimately the aim is to simulate the remaining capital held at the end of a year of trading allowing

for profits made during the year, any capital injections and dividends paid out. This quantity affects the amount of business that can be written in the following year and also determines whether the company is still solvent. The capital at the end of the year is defined in equation 6.29 and the following variables are introduced to lead up to that definition.

Each **'experiment'** runs a specified number N_{sims} of **'simulations'** , these are indexed with the subscript i. Each simulation starts with two companies the **'main'** company (this company is the main focus of the experiments) and a **'competitor'**. The simulation continues one year at a time (indexed with t) until the death of the main company up to a maximum duration T_{max}. If the competitor dies it is re-initialised except that the main company takes a portion (equation 6.15) of the prior competitor's market share. In reality, when one company becomes insolvent others are typically created to take its place and the existing companies grow to take up the business [99]. In the following, for ease of notation, the formulae do not show the subscript i except when it is clearer to do so.

Currency For concreteness assume all financial figures are in Great British Pounds Sterling (GBP).

Unit of risk The insurance market is assumed to have a fixed number (N^{market}) of 'risk units'. For example each risk unit could be a single factory at a given location. These are always shared between the main company and competitor. In reality policyholders face differing levels of risk but here it is assumed each are identical. In the following descriptions the phrase 'per unit' always refers to a unit of risk.

Claims per unit of risk (C_t^{market}) The claims per unit of risk are a stochastic variable that is simulated each year from a distribution that will always produce positive real numbers.

Total claims in the market $\mathbb{C}^{market}$ The total claims in the insurance market are defined as:

$$\mathbb{C}_t^{market} := C_t^{market} N^{market} \tag{6.1}$$

Claims initialisation The market level of claims (per unit), prior to time zero, is available $C_0^{market}, C_{-1}^{market}, \ldots, C_{-(n-1)}^{market}$. This is then divided into claims histories for the main company and its competitor using equations 6.16 and 6.17. The initial market share of each company at time zero is an exogenous parameter.

Estimated expected claims per unit The estimated expected claims per unit is defined as:

$$\hat{E}(C_t^{main}) = \frac{1}{Y^{main}} \sum_{j=1}^{Y^{main}} C_{t-j}^{main} \tag{6.2}$$

Where Y^{main} is the number of years of *past* claims (C_t^{main}) that the main company chooses to use in the averaging calculation. Note that claims are available before time zero. The competitor company uses the same formula but possibly with a different averaging period Y^{comp} and using its own claims history C_t^{comp}.

Premium calculation All risk units are assumed to have the same underlying risk profile (i.e. the level of risk per 1 unit of sum insured is the same for each company). In commercial property insurance, for example, this is equivalent to saying each property is built the same way, in a location that is subject to exactly the same level of risk. In practice this is never the case: properties have different designs, different locations and therefore are subject to different levels of hazard, and respond differently when disaster strikes. This simplification is not a major shortcoming for the broad conclusions drawn in this chapter. The premium rate for the main company P^{main} and for its competitor, P^{comp}, are calculated using the same overall methodology; although the specific pricing assumptions (as specified in the experiment) can be different. The process for the main company is:

$$P_t^{main} = e^{-\mu} \left(\hat{E}(C_t^{main}) + M + \text{payback} \right) \tag{6.3}$$

Where,

μ	is the risk free continuous interest rate
$\hat{E}(C_t)$	is the estimated expected level of claims (equation 6.2
M	is an additional capital load (equivalently 'profit margin') which targets a specified investment return (equation 6.8)
'payback'	denotes a variety of profit sharing options, defined in equation 6.5

Claims excess The **'Claims Excess'** is defined as:

$$\text{Claims Excess} = \frac{C_{t-1}}{\hat{E}(C_{t-1}^{main})} - 1 \tag{6.4}$$

It is the excess (or shortfall) of actual claims compared with those expected.

Payback Unless otherwise specified the **'payback'** component of premium is set to zero. An alternative pricing process is considered in one of the experiments using the formula below:

$$\text{payback} = \hat{E}(C_t^{main}) \left(\frac{C_{t-1}}{\hat{E}(C_{t-1}^{main})} - 1 \right) \psi^{main} \tag{6.5}$$

Where ψ^{main} is the proportion of the previous years' Claims Excess (positive or negative) that is passed on to the policyholder[2] in the current year t.

Value at Risk (VaR**))** The **'Value at Risk'** is defined as follows:

$$\text{VaR}_b(X) := \inf\{X | P(x > X) \leq b\} \tag{6.6}$$

where b is a chosen percentile. This measure is not coherent[3] [14].

Tail Value at Risk (TVaR**)** The **'Tail Value at Risk'** is defined as:

$$\text{TVaR}_b(X) := E(X | X > \text{VaR}_b(X)) \tag{6.7}$$

It is the average value of a variable X given that it has exceeded its Value at Risk. The risk measure is coherent. In insurance this is also known as the Conditional Tail Expectation (CTE) [1].

[2]For example, if last years claims were 50% higher than expected last year, and if the share parameter ψ^{main} was set equal to 50%, then the policyholder will receive a premium loading of 25% of the expected claims this year.

[3]'Coherence' (from Artzner et al [14]) is defined as a risk measure (ρ) satisfying the four axioms of translation invariance, sub-additivity, positive homogeneity, and monotonicity. The following notation is particular to this footnote only. 'Translation invariance': for all $X \in G$ and all real numbers α, we have $\rho(X + \alpha) = \rho(X) - \alpha$. 'Subadditivity': for all $X_1, X_2 \in G$, $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$ 'Positive homogeneity': for all $\lambda \geq 0$ and all $X \in G, \rho(\lambda X) = \lambda\rho(X)$. 'Monotonicity': for all $X, Y \in G$ with $X \leq Y$, $\rho(Y) \leq \rho(X)$.

Capital Load (M) The **'capital load'** is calculated as follows:

$$M = \gamma^{main} \hat{K}^{main} \tag{6.8}$$

Where, γ^{main} is the return on capital required by the shareholders of the main company and $\hat{K}^{main}$ is the capital requirement per unit of risk required by the regulator but estimated by the company. This loading therefore ensures that if the claims are as expected then the premium charged will be enough to pay the claims and provide the desired return on capital to the shareholders (a standard approach as described in Taylor [250]). The capital load is calculated using the same method for the competitor but using a (possibly) different return on capital requirement γ^{comp} and based on the estimated capital based on their own claims history. This is the Target Return method discussed in Chapter 5 (equation 5.22). Two definitions for the capital requirement, $\hat{K}^{main}$, are considered in this chapter: (1) a VaR measure (equation 6.6) and (2) a TVaR measure (equation 6.7). As discussed in section 4.5 on page 222 an extreme level of claims is calculated (Q^{main}):

$$Q^{main} := \text{VaR}_{\tilde{b}}(\hat{C}_i^{main})) \tag{6.9}$$

Where $\tilde{b}$ is the desired regulatory percentile.

$$\hat{K}^{main} := \frac{Q^{main} - \hat{E}(C_t^{main}) - \text{payback}}{1 + \gamma^{main}} \tag{6.10}$$

where $\{\hat{C}_t^{main}\}_i$ are N_{sim}^{main} and N_{sim}^{comp} simulated claims from the claims distribution assumed by each company respectively (these may be different). The parameters for the distribution are estimated from the past claims for the relevant company. As such, as in the real world, they are estimated from very sparse data; this is one of the main reasons for mis-pricing both in the model and in reality.

Market share as function of premium (ζ) This section defines how competition is allowed for in the model. The market share of the main company ζ_t^{main}, as a function of premium rate, is calculated as follows. First, define a relativity factor α_t^{main}:

$$\alpha_t^{main} = \left(\frac{P_t^{comp}}{P_t^{main}} \right)^{\nu_t^{main}} \tag{6.11}$$

where,

$$\nu_t^{main} = \begin{cases} -1 & \text{if } P_t^{main} > P_t^{comp} \\ 1 & \text{otherwise} \end{cases} \tag{6.12}$$

Then define,

$$\delta_t^{main} = \overline{\delta}^{main} \left(\frac{\min(\alpha_t^{main}, \overline{\alpha}^{main}) - 1}{\overline{\alpha}^{main} - 1} \right)^{\sigma^{main}} \tag{6.13}$$

Where the **'customer loyalty'** parameter $0 < \sigma^{main} < \infty$ determines the sensitivity of the market share to the relative premium rates of the two companies. The **'Maximum relativity'** factor $1 < \overline{\alpha}^{main} < \infty$ quantifies the level of α_t^{main} at which the maximum change in market share occurs. When $\alpha_t^{main} > \overline{\alpha}^{main}$ we have $\delta_t^{main} = \overline{\delta}^{main}$ (i.e. the change in market share has reached its maximum permissible value $\overline{\delta}^{main}$). A graph of δ_t^{main} as a function of P^{main} is shown in figure 6.1. Finally, let the **'Market Share'** be defined as:

$$\zeta_t^{main} = \min \left(\zeta_{t-1}^{main} \left(1 + \delta_t^{main}\right)^{\nu_t^{main}}, \overline{\zeta}^{main} \right) \tag{6.14}$$

Where, $\overline{\zeta}^{main}$ is a parameter denoting the maximum market share that the main company can take. The only exception to the above formula is on the death of the competitor in which case the main company's market share is determined in the following way:

$$\zeta_t^{main} = \min \left(\max \left(\zeta_{t-1}(1 + \delta^{death}), \underline{\zeta}^{main} \right), \overline{\zeta}^{main} \right) \tag{6.15}$$

Where δ^{death} is a parameter denoting the growth in market share on competitor death, subject to the overall maximum share $\overline{\zeta}^{main}$ and also subject to a minimum value $\underline{\zeta}^{main}$. This jump in share occurs regardless of the new premium rate (which is likely to have increased) and allows the main company to make supernormal profits after a major event in which the competitor dies.

The above definition is summarised in words as follows: Except on the death of the competitor, starting from a given market share the main company will gain business if its premium rate is lower than the competitor and lose it if higher. The rate of gain/loss is determined by:

1. the function δ whose shape is adjusted by a parameter σ which reflects the assumed level of customer loyalty. Higher values of σ reflect greater levels of

loyalty (to both the main company and competitor) and hence the smaller the jump in market share.

2. the premium difference $\overline{\alpha}$ at which the maximum change in market share $\overline{\delta}$ occurs.

Having calculated a proposed change in market share δ_t the final market share ζ_t is subject to a maximum level $\overline{\zeta}$ but not a minimum. This prevents the main company from taking all the market share. On competitor death the main company seizes significant additional market share through a jump of at least δ^{death} but to no less than a minimum new market share of $\underline{\zeta}$ though still subject to the overall maximum share of $\overline{\zeta}$.

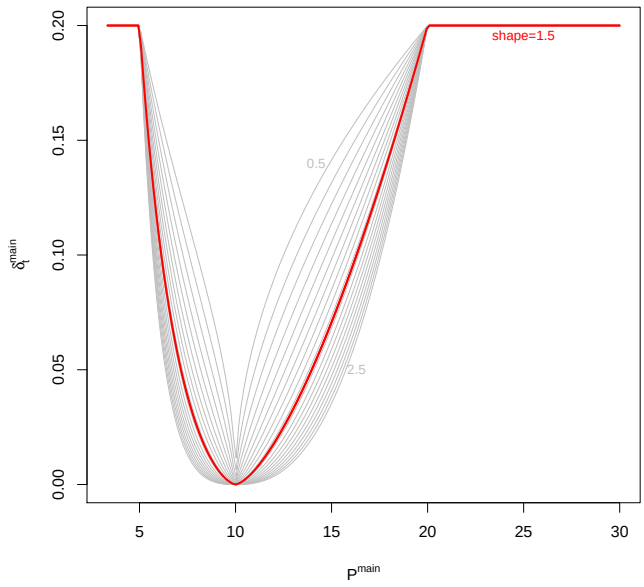


Figure 6.1: Impact of competition on market share. The graphic shows the change in market share of a given level of main company premium P^{main} relative to a fixed competitor premium $P^{comp} = 10$. The Market share change factor δ_t^{main} (equation 6.13) is shown as a function of P^{main}. The maximum change in share $\overline{\delta}^{main} = 0.2$ and $\overline{\alpha}^{main} = 2$, the red line illustrates a shape factor $\sigma^{main} = 1.5$, the grey lines show various other values of σ^{main} ranging from 0.5 to 2.5. Note that when $P^{main} > P^{comp}$ the volume of business will fall for the main company as the prior market share is multiplied by $\dfrac{1}{1+\delta}$

Number of risk units for each company N^{main} and N^{comp} The number of risk units insured by the main company at time t is:

$$N_t^{main} := \zeta_t^{main} N^{market} \tag{6.16}$$

The new competitor is allocated the remaining number of risk units:

$$N_t^{comp} := N^{market} - N_t^{main} \tag{6.17}$$

Since N^{market} is fixed in all experiments the number of risk units for a given company is equivalent to its market share and the two terms are used interchangeably below. The number of risk units for the main company defined above assumes that the company has enough capital to write all the risks that its premium rate would allow it to. This may not be the case and this is discussed in more detail in the Written Premium paragraph on page 292.

Capital at beginning of year ($\mathbb{K}_t$) The capital at the start of the year is the monetary amount of shareholder assets in excess of any reserves or other liabilities on the balance sheet. In this model all claims are assumed settled at the end of the prior year, so reserves are zero. No other liabilities are considered. So the capital at the start of the year will be the assets carried over from the end of the prior year (defined in equation 6.29). At the start of the simulation each company is initialised with an amount of capital $\mathbb{K}_0^{main}$ and $\mathbb{K}_0^{comp}$ respectively.

Required capital The **'Required Capital '** is defined as:

$$\mathbb{K}_{t,R}^{main} = e^{-\mu} \hat{K}_t^{main} N_t^{main} \tag{6.18}$$

Recall that claims are deemed to occur at the end of the year; the capital invested at the start of the year will grow at the risk free investment rate (e^{μ}) hence the amount needed at the start of the year is reduced by the reciprocal of this growth.

Free capital The **'Free Capital '** is defined as

$$\mathbb{F}_{t,R}^{main} = \mathbb{K}_t^{main} - \mathbb{K}_{t,R}^{main} \tag{6.19}$$

Written premium ($\mathbb{P}$) If the capital held ($\mathbb{K}_t^{main}$) is sufficient to write all the risk units the premium rate would support, the total written premium ($\mathbb{P}_t^{main}$) is calculated as:

$$\mathbb{P}_t^{main} := N_t^{main} P_t^{main} \tag{6.20}$$

This may not be the case, however, since each unit of risk written requires a certain amount of capital to support it. If the the total amount of Required Capital is more than the actual capital held (i.e. $\mathbb{K}_{t,R}^{main} > \mathbb{K}_t^{main}$) then there are two options (in practice): (1) the company must reduce the number of risk units it sells or (2) it must seek additional capital from the investment markets ('recapitalise'). The model presented here assumes option 2 is followed in which case the companies will cover the maximum number of risk units (N^{main}) that the premium rates will allow. The recapitalisation approach is defined as follows:

Recapitalisation ($\mathbb{J}$) If necessary a capital 'injection' $\mathbb{J}_t^{main}$ will be made to top up the required capital as follows:

$$\mathbb{J}_t^{main} := \max(\mathbb{K}_{t,R}^{main} - \mathbb{K}_t^{main}, 0) \tag{6.21}$$

In words, where the premium rates would permit more business to be written but the level of capital in the company is not sufficient to write this level of business a capital injection will be made[4].

Claims share as function of market share (η) In practice it is typical that the proportion of the total market's claims paid by a given company is not equal to their market share (ζ_t^{main} defined in equation 6.14). This is not modelled here so that $\eta_t^{main} = \zeta_t^{main}$.

Total claims for each company The main company's total claims are calculated as $\mathbb{C}_t^{main} := \eta_t \mathbb{C}^{market}$ and the competitor's claims are defined as $\mathbb{C}_t^{comp} := \mathbb{C}_t^{market} - \mathbb{C}_t^{main}$. The claims per unit of risk (C_t^{main}) are then calculated as:

$$C_t^{main} := \frac{\mathbb{C}_t^{main}}{N_t^{main}} \tag{6.22}$$

[4]This annual recapitalisation can be thought of as the similar to the annual Lloyd's business planning process ('coming into line') [143] where names provide sufficient capital to write the proposed business plan.

Since $\eta_t^{main} = \zeta_t^{main}$ we have $C_t^{main} = C_t^{market}$, the notation retains the superscript *main* to emphasise the company in which the claims occur.

Investment return ($\mathbb{I}$) The company is assumed to invest in risk free investments only. Premiums and capital injections are assumed to occur at the start of the year, and claims (for simplicity) are assumed to occur at the end of the year. The investment return in the year ($\mathbb{I}_t^{main}$) is then calculated as:

$$\mathbb{I}_t^{main} := \left(\mathbb{P}_t^{main} + \mathbb{K}_t^{main} + \mathbb{J}_t^{main}\right)(e^{\mu} - 1) \tag{6.23}$$

In practice a company may have different investment strategies for its capital and premium funds; this is not modelled here.

Profit (π) The profit in the year is defined as:

$$\pi_t^{main} := \mathbb{P}_t^{main} - \mathbb{C}_t^{main} + \mathbb{I}_t^{main} \tag{6.24}$$

Note that capital injections are not included in the calculation of profit.

Dividends ($\mathbb{D}$) In reality dividend setting involves much human interpretation and is therefore hard to express mathematically. The following discussion motivates the dividend calculation proposed below. Dividends to shareholders are paid out of profits, or possibly from capital if sufficient profits are not available (see Chapter 1). In practice dividends are at the choice of the Directors who seek to meet many, sometimes conflicting, objectives. A relatively steady dividend stream is often thought to be a sign of a well run company, so Directors often see the prior years' dividend level as a starting point for the current year in order to incorporate a degree of **'smoothing'**. Shareholders will typically expect a larger dividend if profits have been higher than average in the year and will accept a lower dividend if lower than average. A dividend is limited by total capital. As a final constraint Directors would not wish the free capital of the company rise too high since this can be seen as inefficient and subject to high opportunity costs. The Directors are, however, likely to retain *some* additional capital as a **'buffer'** against the risk of regulatory insolvency as described in Hitchcox et al [105]. In practice any capital thought to be in excess of a desired buffer would be used to either grow the business by broadening

into other lines of business (which is not modelled here), or returned to shareholders as a special dividend (as Royal and Sun Alliance did in the late 1990s, [201]). Let θ^{main} denote the proportion of last years' dividend that the Directors would seek to pay as a *minimum* dividend in the current year, in line with the comments about smoothing above. Let ω^{main} denote the proportion of profits the Directors would seek to return to shareholders. Recall that the desired return on capital is denoted by γ^{main}. Finally let ϵ^{main} denote the additional buffer capital that will be retained if any special dividend is paid. Then define:

$$\mathbb{D}_t^{main} := \max(\mathbb{D}1_t^{main}, \mathbb{D}2_t^{main}) \tag{6.25}$$

where:

$$\mathbb{D}1_t^{main} := \max\left(\theta^{main}\mathbb{D}_{t-1}^{main}, \min\left(\omega^{main}\pi_t^{main}, \gamma^{main}\mathbb{K}_t^{main}\right)\right) \tag{6.26}$$

and $\mathbb{D}2$ is designed to stop the capital growing too fast, as follows:

$$\mathbb{D}2_t^{main} := \left(\mathbb{K}_t^{main} + \pi_t^{main} + \mathbb{J}_t^{main}\right) - \mathbb{K}_{t,R}^{main}(1 + \epsilon^{main}) \tag{6.27}$$

Note, in this model it is therefore possible for a special dividend to be paid, only to find at the beginning of the next year that a capital injection is sought. In practice Directors would seek to avoid this by considering next years' business plan and consequent capital requirements before setting the dividend. The approach adopted here is due to the model design where next years' premium is not known at the time the dividend is set. The timing of these two cashflows are effectively the same in the model so the present value of the dividend stream (less capital injections) is not affected.

Company Value ($\mathbb{V}$) In practice, company stock market value is determined by the market forces of supply and demand. Here, the **'Present Value[5]'** of the dividend stream less capital injections is taken to define the **'Company Value'**. Note that this excludes the initial capital held by the company which are shareholder assets and make up part of the value of the company. If any capital injections are made it is assumed that the same shareholders from the prior year provide all

[5]For a continuous rate of interest μ, 1 GBP receivable in 1 years time has a 'present value'(PV) of $PV(1) = e^{-\mu}$. A series of payments $A = \{a_1, ...a_n\}$ receivable in years $1, ...n$ has a present value $PV(A) = \sum_{t=1}^{n} a_t e^{-\mu t}$. This process is known as 'discounting' or 'deflating'.

the capital for the next underwriting year. If new shareholders were permitted to provide a proportion of the capital injection then the share capital of the existing shareholders would become diluted. The model effectively assumes that new capital, if required, can be borrowed at the risk free rate. The Company Value ($\mathbb{V}$) is therefore formally defined as:

$$\mathbb{V} := \frac{1}{N_{sims}} \sum_{i=1}^{N_{sims}} \sum_{t=1}^{T_{main}(i)} (\mathbb{D}_{t,i}^{main} - \mathbb{J}_{t,i}^{main} e^{\mu}) e^{-\mu t} \tag{6.28}$$

where $T_{main}(i) = \min(T_{lifetime}(i), T_{max})$ and $T_{lifetime}(i)$ is the number of years the main company lives in simulation i.

Capital at end of year ($\mathbb{K}_{t+1}$) The capital at the end of the year (which becomes the capital at the beginning of the following year) is then calculated as follows (noting that the investment income on any capital injection is included in profits):

$$\mathbb{K}_{t+1}^{main} = \mathbb{K}_{t}^{main} + \pi_{t}^{main} + \mathbb{J}_{t}^{main} - \mathbb{D}_{t}^{main} \tag{6.29}$$

Control Experiment A series of experiments will be described and one of these is designated the 'control' against which the others are compared. See section 6.3.

Importance of various parameters Table 6.1 summarises the importance or otherwise of the various parameters and distribution assumptions in the model.

Table 6.1: Importance of various model parameters and distribution assumptions

Assumption	Variable name/choice	Importance
Underlying distribution	Lognormal etc	Not explored here, expected to be important.
Underlying claims parameters	$E(C)$,var(C)	Important, especially the relationship of the variance parameter to the mean. (Figure 6.11)
Pricing distribution	Lognormal etc	Important, may be helpful to choose heavier tailed pricing distribution than reality. (Figure 6.9)
Target Return on capital	γ	Important, lower target returns lead to shorter company lifetime and lower Company Value. (Figure 6.3)
Payback proportion	ψ	Important, use of the payback rule proposed shortens company lifetime. (Figure 6.6)
Dividend parameters	θ, ω, ϵ	Not explored here. Essential in determination of Company Value.
Regulatory target return period	b	Important, key determinant of expected company lifetime. (Figure 6.7)
Regulatory risk measure	VaR etc	Important, key determinant of expected company lifetime. (Page 319)
Market share, customer loyalty	$\overline{\alpha}, \overline{\delta}, \overline{\zeta}, \underline{\zeta}, \delta, \sigma$	Not explored here. Expected to be important.
Past claims included	Y	Important (fewer years included lead to shorter company lifetime) - results not presented but have been tested.

6.1.1 Some Simplifications

Single Currency The model assumes the insurer operates within a single currency environment under a single regulatory framework that does not change over the lifetime of the company.

No investment decision rules The modelled company has a single asset basket - and no investment decisions are modelled.

No explicit expenses All expenses are assumed to be included within the definition of a claim.

No tax Tax is not included, a complication that provides no additional insights. If it were, the premium rate would be adjusted to seek a desired post tax return and the effect would broadly cancel out.

Short tail business All claims occur and are settled at the end of the year or, equivalently, 'Reinsurance To Close' is purchased annually and included in the claims cost. In other words there are no 'long tail' policies that could further deteriorate or produce profits so 'reserve risk' is not included explicitly (see glossary for definition of terms)

Zero Inflation Inflation is the tendency for prices to increase over time; this model assumes zero inflation to avoid difficulties in comparing multiple years.

6.1.2 Possible model extensions:

Alternative Claims processes The claims generation processes used in this chapter are standard, positive definite statistical distributions: Gamma, Lognormal and Pareto. The distributions used are common in insurance practice [88,199]. Any appropriate random process generating claims distributions could be used, however. For example the Lorenz 96 index from Chapter 4 could be used and this would allow forecasting to be brought into the pricing method. Similarly the simple hurricane claim generator described in Chapter 5 could be used.

Pricing extensions: If the claims process included forecastable elements then the premium calculation could be adapted to use some of the methods described in Chapters 4 and 5. The key learning point from Chapters 4 and 5 is that this would have to be done carefully to avoid passing all benefits to policyholders and retaining too much insolvency risk as a consequence. For example ϕ-transformed pricing methods from Chapter 4 could be used to adjust both the $\hat{E}(C_t^{main})$ component (equation 4.15) and also the 'capital load' (not considered in this thesis). Business volume scaling could also be considered (described in both Chapters 4 and 5).

Recapitalisation not permitted: If recapitalisation is not allowed then the number of risk units will be constrained by the available capital as follows. Let the number of risk units be denoted $\tilde{N}_t^{main}$. Then the number of risk units allowing for capital constraints is calculated as:

$$N_t^{main} = \min(\frac{\mathbb{K}_t^{main}}{e^{-\mu}\hat{K}_t^{main}}, \tilde{N}_t^{main}) \tag{6.30}$$

6.2 Plot descriptions

The following describes three types of plot that are used in this chapter. The 'Quantile Boxplot', which shows how the mean value of a chosen statistic varies between experiments, the 'Time Mean' plot which shows how the mean of a chosen statistic varies over time and the 'Specific Simulation' plot which shows the results for a chosen statistic for a specific simulation i.

Quantile boxplots Given a 'result' from the simulation (for example the average lifetime). The 'Quantile Boxplot' is used to show how the average result varies between experiments; but also to illustrate the size of sampling error in the result. The figure is produced as follows:

- Given results from a number of experiments $n \in \{1, ...N_e\}$; there is one result per simulation in each experiment where the total number of simulations is N_{sim}.
- For samples $j \in \{1...N_J\}$

  - Sample N_K simulations $s_1, ...s_{N_K}$ from the N_{sim} simulations, with replacement and calculate the mean result ($E_{j,n}$) over these simulations.

- For one of the experiments (n=1 say) determine specified quantiles $\{Q_i^1\}_{i=1}^q$ of $E_{j,1}$; for each quantile Q_i^1 determine which sample $j(i)$ it arose from.
- For the other experiments (n=2,...N) calculate the result for each sample $j(i)$ and determine what quantiles Q_i^n these represent.
- Draw lines (call these **'Quantile Lines'**) between $Q_i^1, Q_i^2, ...Q_i^n$.
- Draw grey boxes around $\min(E_{j,n})$ and $\max(E_{j,n})$.

In this chapter $N_{sim} = 2^{10}$, $N_K = 2^9$ and $N_J = 2^9$. See figure 6.3 for an example where: $N_e = 5$, the 'result' is the length of lifetime of the main company in each simulation, $n = 3$ relates to the control experiment with quantiles $Q_i^3 = \frac{i}{10}$ for $i \in \{0, ...10\}$. For example the control experiment quantile $Q_2^3 = 10$ relates to the following quantiles of the other experiments: $Q_1^2 = 11$, $Q_2^2 = 6$, $Q_2^4 = 8$ and $Q_2^5 = 12$: a red line joins these together.

If the grey boxes in the Quantile boxplots do not overlap significantly then it is reasonable to conclude there are significant differences between the means arising from the experiments. However, as with figure 6.3 there may still be a significant difference even if the boxes do overlap. This is highlighted when Quantile Lines all show a similar slope and do not overlap very frequently. Such behaviour indicates that whilst there is sampling error in the value of the mean (shown by the grey box) - the relative behaviour of different experiments in different simulations is consistent and hence the difference is significant.

Time mean plot Figure 6.2 is an example of the 'Time Mean' plot. The year is shown on the x-axis; the value of the chosen statistic is shown on the y-axis. Let a statistic of interest be denoted $X_{i,t,e}$ where i denotes the simulation number ($i \in \{1, ...N_{sim}\}$), e denotes the experiment type (in this case 'control' and 'known distribution') and t denotes the year. As described in the model design section, each simulation is run until the main company has died. Let $I(i, t, e)$ denote the status (alive or dead) of the main company at time time t, for simulation i and experiment e; then I is equal to 1 if the main company is still alive at time t and equal to 0 if it is dead at this time. Then $M(t, e) = \sum_{i=1}^{J} I(i, t, e)$ denotes the

number of main companies that are still alive by time t out of the N_{sim} simulations. The value of $M(t,e)$ may differ between companies ($m_{control}(t)$ and $m_{perfect}(t)$ say) the labels at the top of the plot show the minimum number of simulations ($m(t) = \min(m_{control}(t), m_{perfect}(t))$) over which the average is taken. For a given experiment e the line shown in the plot is defined by $L(e,t) = \frac{\sum_{i=1}^{J} I(i,t,e)X_{i,t,e}}{M(t,e)}$, the average value of the statistic for the main companies that are still alive by time t. The value of M is shown for particular times at the top of the plot. A 95% confidence interval is shown around the line based on the Gaussian approximation of the distribution of the mean.

Specific Simulation Plot Figure 6.4 is a 'Specific Simulation Plot' which is a figure type that is used several times in this chapter. Taking the right hand plot as an example the x-axis shows the simulation year and the y-axis shows the statistic of choice. The figure shows a chosen statistic (in this case the premium per unit of risk) for a set of chosen experiments e (in this case the control experiment and various different payback levels). The statistics are shown over the maximum lifetime over all the experiments, in this example the lifetime exceeds 70 years in two of the experiments but is under 50 for the rest. The left hand plot shows the claims per unit of risk, the simulated claims history prior to time zero is shown by a dotted line.

6.3 Choice of control experiment

The following describes how the control experiment was chosen. The two companies described in section 6.1 are set up so that certain desirable characteristics arise but other formulations are possible and no less appropriate. The point of the control experiment is to give a stable point from which the impact of changing assumptions can be compared. The underlying claims are a sample from a Lognormal distribution with a mean of 1 and variance of 0.65. The control experiment was chosen so that the following desirable characteristics arose:

- Assumptions that market practitioners would feel are reasonable, in particular both companies exhibit realistic looking behaviours;

- The main company retains a material market share until death i.e. there isn't a steady decline (or increase) in market share in each simulation; as this would suggest an underlying bias;
- The competitor doesn't die too often before the main company, but nor does it outlive the main company most of the time;

Initially both companies were chosen to be identical, including their pricing strategies. In this case the premium rates set by the two companies are always identical so the market share stays constant at 50% this is a consequence of both companies having the same underlying claims experience: if this was varied they would likely set different premium rates. The competitor and main company always die together in this case and have a mean lifetime of 71.6 years and a 90%ile range of (70.0, 74.3). The average loss ratio[6] for the main company is 79.8% (as for the competitor in this case as they charge the same premium rate). The behaviour of this experiment does not look realistic, in particular no interesting competition effects arise. So this approach was discarded.

In experiment 2 the competitor uses a different distribution for pricing, in this case Pareto (being a heavier tailed distribution this leads to more expensive premium rates on average). In this case the main company lives on average for 71.2 years with a 90% confidence interval of (68.0, 73.7) and the competitor dies just 2.2% of the time before the main company and 61.4% of the time at the same time. As such, the competitor is 'too strong' in the sense that it is outliving the main company too often to be a realistic insurance market, given they have the same claims experience. As the competitor charges higher premiums on average (1.335 vs 1.318), the main company gets a greater market share (57.6% on average) and an average loss ratio of 79.8% compared to 78.1% for the competitor (lower as expected because it charges a higher premium on average). To 'weaken' the competitor a third experiment was run where the competitor only uses 9 years of past claims data for pricing (compared to 15 years for the main company) which tends to make the competitor 'overreact' to new claims information. This third variant was used as the control case and is described in the next paragraph. The use of a shorter claims horizon for the main company of 5 or 10 years led to shorter lifetimes and premium rates (not

[6] $\text{loss ratio} := \dfrac{\text{claims}}{\text{premiums written}}$

shown). A number of different pre zero claims histories were investigated, the relative strength of the companies are affected by the initialisation because their pricing strategies differ, the effects can mainly be explained by premium rate differences. If a practical and realistic model were ever built, with companies representing their actual counterparts in the real world, then this feature would not be a constraint because actual claims histories would be used.

6.4 Experiment descriptions

The following describes the key parameters of all experiments. There are eight groups of experiments in all: (1) The control, (2) Perfect pricing where the claims distribution is known, (3) Alternative target returns on capital, (4) Use of various Payback parameters, (5) Impact of different strengths for the regulatory test, (6) Impact of using TVaR for the regulatory test, (7) Use of imperfect pricing distribution assumptions and (8) Impact of different underlying claims distributions. Note that the assumptions for the competitor are kept the same as in the control experiment unless explicitly noted. It is stressed that the statistics arising in each experiment allow comparisons between experiments and to explore relationships; the statistics are useful in a relative sense.

Experiment C6.1 Control experiment.

Global parameters .

Number of simulations: $N_{sims} = 2^{10}$ simulations of the company lifetime process were run, each until either the main company dies or the maximum duration is reached.

Maximum duration: $T_{max} = 2000$ years

Market size: $N^{market} = 10000$ risk units.

Initial market shares: $\zeta_0^{main} = 0.5$, $\zeta_0^{comp} = 0.5$

Risk free interest rate: $\mu = 0.035$

Recapitalisation method: Option 2 (annually), both companies.

Initial company capital: $\hat{K}^{main} = 10000$, $\hat{K}^{main} = 10000$

Claims parameters

Underlying claims process distribution: $C^{market} \sim$ Lognormal

Underlying claims process parameters: $E(C^{market}) = 1$, $var(C^{market}) = 0.65$

Market share parameters: $\overline{\alpha}^{main} = 2$, $\overline{\delta}^{main} = 0.2$, $\overline{\zeta}^{main} = 0.75$, $\underline{\zeta}^{main} = 0.5$, $\delta^{death} = 0.2$, $\sigma^{main} = 1$

Regulatory Parameters .

Risk measure for regulatory test: VaR (Value at Risk) (equation 6.6)

Return period for regulatory test: 200 years (i.e. maximum 0.5% per annum failure probability)

Pricing parameters .

Assumed claims distribution: $\hat{C}^{main} \sim$ Lognormal, $\hat{C}^{comp} \sim$ Pareto

Past claims used in pricing: $Y^{main} = 15$, $Y^{main} = 9$

Target return on capital: $\gamma^{main} = 0.15$, $\gamma^{comp} = 0.15$

Payback proportion:: $\psi^{main} = 0$, $\psi^{comp} = 0$

Number of simulations to determine capital: $N_{sim}^{main} = 10000$, $N_{sim}^{comp} = 10000$.

Dividend parameters .

Target proportion of last years dividend: $\theta^{main} = 0.5$, $\theta^{comp} = 0.5$

Target proportion of profits to return: $\omega^{main} = 0.6$, $\omega^{comp} = 0.6$

Max additional capital buffer after dividend: $\epsilon^{main} = 0.3$, $\epsilon^{comp} = 0.3$

Experiment C6.2 Underlying Claims distribution known.

All parameters as for Experiment C6.1, except for:

$\hat{E}(C_t^{main}) = 1$, the true underlying expected mean claims is used in the calculation of capital and premium.

$\hat{var}(C_t^{main}) = 0.65$. The true variance is used in in the calculation of capital and premium.

Experiments C6.3.x Alternative Target Returns on Capital.

All parameters as for Experiment C6.1, except for:

Target return on capital: Four experiments ($x \in \{a, b, c, d\}$) are carried out with $\gamma^{main} = \{0.05, 0.1, 0.2, 0.25\}$ respectively.

Experiments C6.4.x Use of various Payback parameters.

All parameters as for Experiment C6.1, except for:

Payback proportion: Two experiments ($x \in \{a, b\}$) are carried out with $\psi^{main} = \{0.1, 0.5\}$ respectively.

Experiments C6.5.x Impact of different strengths for the regulatory test.

All parameters as for Experiment C6.1, except for:

Return period for regulatory test: Five experiments ($x \in \{a, b, c, d, e\}$) are carried out with $Q_{reg} = \{50, 100, 150, 250, 500\}$ respectively. Note this applies to both the main company and the competitor.

Experiments C6.6 Impact of using TVaR for the regulatory test.

All parameters as for Experiment C6.1, except for:

Risk measure for regulatory test: One experiment is carried out using a TVaR measure.

Return period for regulatory test: $Q_{reg} = 69$ chosen so that the same level of extreme claim arises for the true underlying claims distribution (see derivation on page 319). Note this applies to both the main company and the competitor.

Experiments C6.7.x Use of incorrect pricing distribution assumptions.

All parameters as for Experiment C6.1, except for:

Assumed claims distribution: Two experiments ($x \in \{a, b\}$) are carried out with $\hat{C}^{main} \sim \{\text{Gamma}, \text{Pareto}\}$ respectively.

Experiments C6.8.x Impact of different underlying claims distributions.

All parameters as for Experiment C6.1, except for:

Underlying claims process distribution: Two experiments ($x \in \{a, b\}$) are carried out with $C^{market} \sim \{\text{Gamma}, \text{Pareto}\}$ respectively. Note the claims affect both the main company and competitor.

6.5 Results

Control (C6.1) The main company in the control experiment will be referred to as the **'Control Company'**. The mean lifetime of the Control Company was 72.7 years, an 80% confidence interval[7] is (69.1, 76.6) years. The 10% ile lifetime was 12 years and 90%ile was 160 years. The Company Value is GBP 48,777. This will be taken to be the value of the Control Company against which others can

[7]Confidence intervals around the expected lifetime are created by the following process: For $i \in \{1, ..N_r\}$, sample the lifetime with replacement from N_b of the N_{sims} simulations and calculate the average lifetime L_i over those N_b values; calculate quantiles from the set $\{L_1, ...L_{N_r}\}$. In this chapter $N_r = 2^{10}$ and $N_b = 2^9$. An '80% confidence interval' is defined as the interval between the 10th and 90th quantile of $L_1, ...L_{N_r}$

be compared. Company value is relevant to the hypothetical shareholders; another measure, perhaps more relevant to regulators, is the (deflated) size of the deficit when the company eventually dies. For the Control Company this is GBP 1,108. The competitor dies 31% of the time before the main company and 42% of the time in the same year as the main company, due to a large claim that neither survives, and hence 27% afterwards. On this measure the strengths of the two companies are similar.

Underlying Claims distribution known (C6.2) The Control Company does not know the underlying claims distribution parameters. It is interesting to note that because of this the mean company lifetime is 72.7 years, even though regulatory capital requirements are set to give a failure rate of no more frequently than 1 in 200 years. This discrepancy arises because, in the control experiment, the main company is trying to estimate the $\text{VaR}_{0.005}$ (equation 6.6) of a relatively fat tailed distribution from a small amount of data.

What would the mean lifetime be if the company *did* know the underlying claims distribution[8] and its parameters? Call such a company **'perfect'**. The mean lifetime of the Perfect Company is 392.9 years with an 80% confidence interval of (375.8, 407.0) years. Note, however, that the maximum duration (2000 years) occurs eight times in the 1024 simulations, so the true mean lifetime would be somewhat greater than this. The mean lifetime is greater than the regulatory minimum of 200 years because the company holds capital in excess of the minimum (on average 20.3% more in this case). Of the 1024 simulations, the 10% ile lifetime is 40 years and the 90% ile is 926 years, illustrating that even when the underlying distribution is known perfectly there is significant probability that the company lifetime is short lived. Regulators are often blamed for company failures [38,213,251,255] but a large claim within expectations would not illustrate a failure of the regulatory regime, just the underlying volatility of the claims process. The more capital that companies are required to hold the larger extreme claims they can survive but the higher premiums they must charge to provide the target return on capital to shareholders. The chosen level of regulatory minimum capital (equivalently the return period in

[8]In Hooker et al [106] this would be equivalent to zero parameter uncertainty and zero specification error.

the VaR calculation) is therefore a political choice that has to balance affordability with expected solvency. To determine whether a regulator has failed one would have to show that they did not spot a company that was miscalculating its risks; this is not determined by one large event that may be within the distribution of expected outcomes. The Company Value is GBP 57,844, some 18.6% more valuable than the control; and illustrates the value in a better understanding of the underlying risk (for example if claims data can be coupled with other predictive exogenous variables). If, however, only dividends paid up to the time that the Control Company dies are included (or up to the earlier death of the Perfect Company, which is rare in this experimental design) then the value is only GBP 47,003 which is actually lower than the Control Company. So in general, the dividends paid are lower from the Perfect Company, but, since it lives so much longer, the discounted value of dividends after the death of the Control Company more than compensate, leading to a higher company value overall. Ironically, since it would take many years for this to become apparent, the Control Company would appear (to the media for example) to be the more profitable and better run company.

The control experiment prior claims levels are samples from the underlying distribution but are chosen so that the initial mean and variance lead to roughly the 'right' price (samples which don't achieve this are rejected) per unit of risk (1.45) in year 1, this explains why the red line in figure 6.2 starts at the correct underlying level (which of course the blue perfect company calculates correctly). After this time, however, in the control experiment, both the main company and its competitor underprice on average (charging 1.32 and 1.30 respectively). Figure 6.2 shows the premium rate averaged over all simulations, by year for both the Control Company and the Perfect Company. The Perfect Company calculates the premium broadly correctly each year. The residual variance is due to the fact that the Perfect company still estimates capital requirements by simulating from the claims distribution - and sampling error remains. The calculated premium by the Perfect Company falls in the interval (1.42, 1.48) 90% of the time (note figure 6.2 does not display this level of variance because it shows the premium rate averaged over all simulations, so sampling error is largely, but not completely, smoothed out).

The competitor in experiment C6.2[9] makes the same pricing assumptions as in the control experiment. The Perfect Company's market share falls to 27% on average because it charges a higher premium than Competitor:C6.2. Given the Perfect Company knows the true claims distribution, it is not surprising that the Competitor:C6.2 dies 82% of the time before the main company and 14% at the same time (and just 4% afterwards).

Alternative Target Returns on Capital (C6.3.x) In the control experiment the shareholders of both companies require a return on capital of 15%. In this subsection the target return for the main company is varied from 5% to 25%. The premium rate is lower when the target return on capital is lower due to the pricing formula; this means that the required capital per unit is higher (it has to compensate for the fact that the premium provides less protection against losses). As expected the market share rises as the target return falls due to lower premium rates. The effect on average lifetime, figure 6.3(left), was not easy to predict in advance. The premium rate is lower, so one might expect this to imply a shorter lifetime. The lifetime is shorter, but that is not the reason. Although the premium is lower, the capital per unit is set to be higher to compensate exactly for this. The lifetime decreases for lower target returns because the free assets per unit is lower on average when the target return is lower (figure 6.3(right)) so the company is less likely to survive shocks in excess of the regulatory minimum.

Figure 6.3(bottom) shows Company Value for the various experiments. Would a lower target return lead to a higher or lower company value? The lower target return will make premium rates more competitive increasing the number of policies being sold; but conversely each policy will be less profitable. Figure 6.3(bottom) shows that for a target return of 5% and 10% the Company Value is lower than the control. When the target return is 20% and 25% there is little difference in value relative to the control suggesting that whilst individual policies under these assumptions are more profitable the lower market share counteracts this. The difference in behaviour for low and high target returns shows that the impact in general will be subtle and dependent on the specific competitive details assumed.

[9]In general the main company and competitor will be referred to by their experiment number - so in experiment C6.2 it is referred to as: 'Competitor:C6.2'.

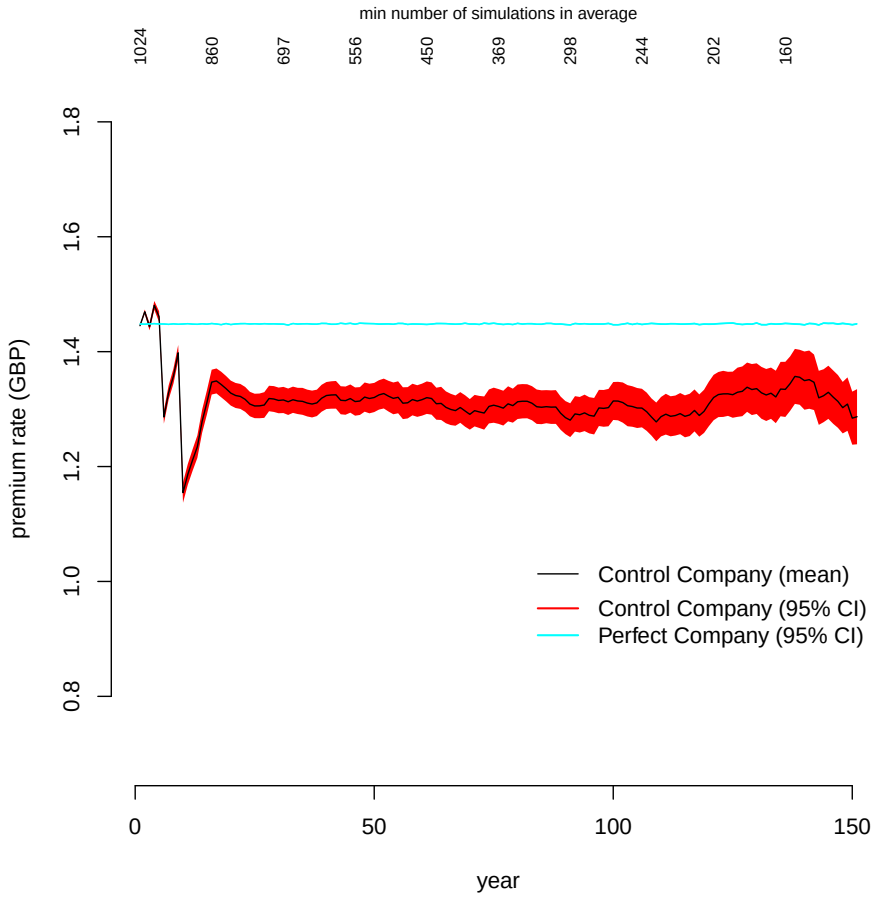


Figure 6.2: Experiment 6.2: Time Mean plot showing the average premium rate per unit of risk for hypothetical Perfect Company (shown in cyan) compared to Control Company (shown in red). The minimum number of simulations included in the average starts at 1024 (all simulations) and ends at 160 indicating that in many simulations one or both of the companies is typically dead before this time. The red region for the Control Company shows a 95% confidence interval around the mean (black line) based on a Gaussian approximation ($\bar{P}(t) \sim N(E(P(t)), \frac{sd(P(t))}{\sqrt{n(t)}}$). The Cyan region is also a 95% confidence interval but is thin because the Perfect Company sets almost the same premium in each simulation (slight differences arise from sampling error in calculating the capital requirement $\hat{K}$). Note that the Control Company's red confidence interval falls below the Perfect Company's cyan region: the underpricing of the former is therefore significant at the 95% level.

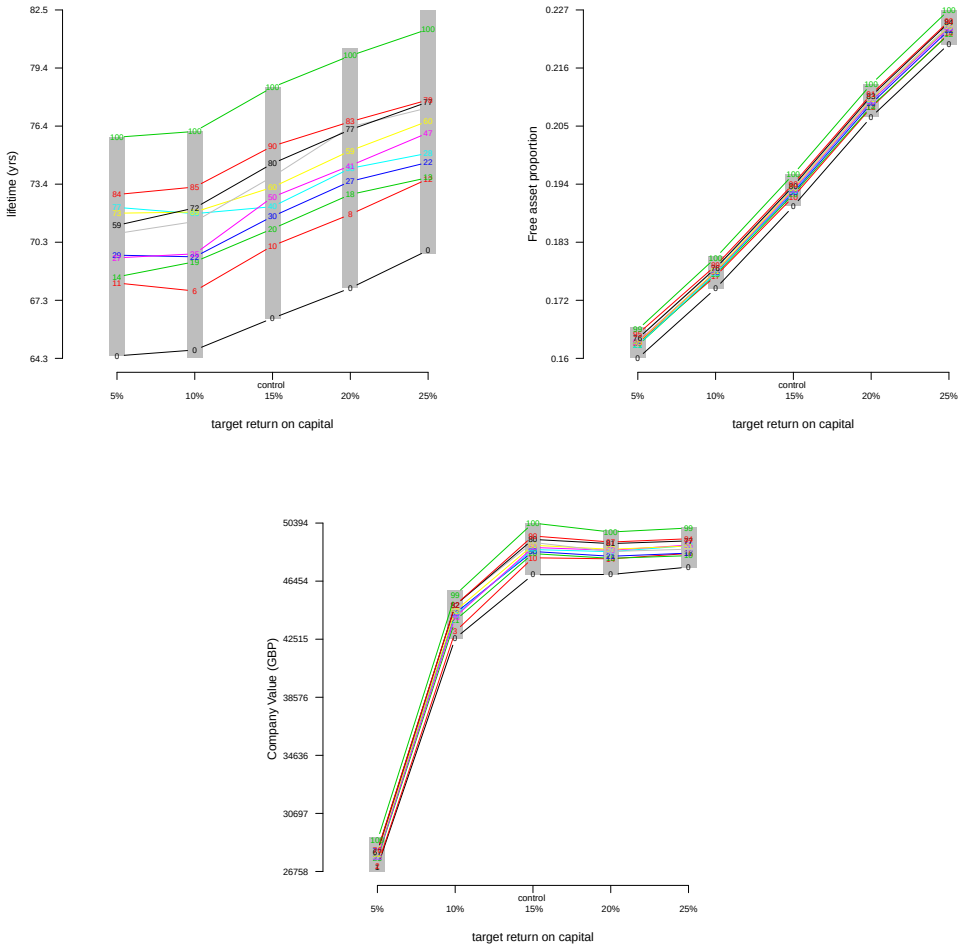


Figure 6.3: Experiment C6.3.x: Quantile Boxplots showing the mean lifetimes, free asset proportion, and present value of dividends of the main company for different target return on capital (5-25%). In the mean lifetimes plot the width of the grey boxes illustrate sampling error in the calculation of the mean. Whilst these grey boxes overlap it is still clear that the increasing trend in lifetime is robust as the target return increases because the quantile lines almost all show an increasing trend and do not overlap very frequently. The increase in free asset proportion as target return increases is clear to see. Note, the Company Value initially increases as the target return increases but this levels off for target returns in excess of 15%.

Variation of Payback parameters (C6.4.x) Recall that the premium rule is defined in equation 6.3. In all the experiments described so far the payback parameter is set to zero. The payback rule tested in this subsection is defined in equation 6.5 for $\psi \in \{0.1, ...0.5\}$. The following series of graphics illustrate the outcome for simulation number 753 (chosen because it is one of the examples when the lifetimes of the various payback rates differ). Figure 6.4(left) shows the claims per unit of risk in the chosen simulation and 6.4(right) the corresponding premiums charged. Note that the Control Company (red line) with no payback has a much smoother premium (as expected). As the ψ parameter increases the premium rate gets less smooth. For most values of ψ, companies using payback die in year 57 (except when $\psi = 10\%$) following a larger claim relative to low claims in previous years. Note that the premium rate for companies using payback is low as they are sharing the profitability in the years running up to year 57, so they are unable to survive the large claim in year 57. Figure 6.5 zooms into years 25 to 35 to examine behaviour around year 29 when there is a large claim. The figure shows that after the large claim the premium rises in all cases in year 30. But, as expected, the higher the ψ parameter, the greater the price rise in year 30. These are followed by significant premium reductions in year 31 caused by the very benign loss environment in year 30. The price rise for companies using the payback rule causes them to lose market share (because their prices are greater than their competitors) whereas the Control Company grows its market share. Companies using the payback rule pay a large dividend in year 30, but the Control Company steadily increases its dividends and is more successful in years 31 to 35.

Figure 6.6(middle left) shows that the premium rate is about the same on average for all the payback rules. This is to be expected as the payback rule is symmetric with respect to losses in the prior year. The average number of risk units 6.6(top left) follows a clear trend, however, the number of risk units increases with ψ. At first it may appear odd that the premium rate is the same on average, but the number of risk units is different. In the previous experiments higher premiums have led to lower market share and vice versa, so one might expect similar premium levels to lead to a similar market share. Closer investigation, shows that the difference occurs for the following reason (illustrated when $\psi = 0.5$):

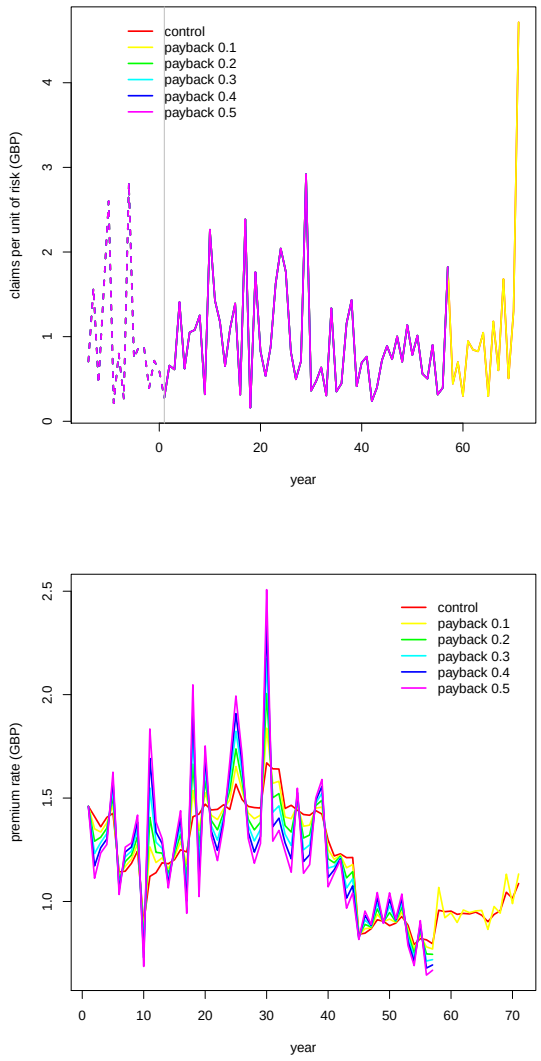


Figure 6.4: Experiment C6.4.x: Specific Simulation Plots for the main company claims and premiums per unit of risk for simulation 753. Results are shown for the control experiment (red) and several levels of payback percentage (between 10% and 50%)

- When payback is used ($\psi \neq 0$) the premium rate is less than control experiment more often than it is greater. For example when $\psi = 0.5$ the payback premium (P_p) is typically less than the control (P_c) around $q = 59\%$ of the time. The claims distribution is Lognormal with mean 1 and variance 0.65 and so the probability of losses falling below the mean is 64%, hence it is more likely that the payback rule will give a discount to the prior premium rate;
- Let $\bar{P}_p$ be the average premium for a company using payback and $\bar{P}_c$ be the average for the Control Company. Let $\bar{P}_{p,G}$ be the average premium for a company using payback in the cases when $P_p > P_c$. Similarly $\bar{P}_{c,G}$ is the

average premium of the control company when $P_p > P_c$. Let $\bar{P}_{p,L}$ be the average premium for a company using payback in the cases when $P_p \leq P_c$ and define $\bar{P}_{c,L}$ similarly. Let $\bar{R}_G = \frac{\bar{P}_{p,G}}{\bar{P}_{c,G}}$ and $\bar{R}_L = \frac{\bar{P}_{p,L}}{\bar{P}_{c,L}}$. Then, for example, when $\psi = 0.5$, $\bar{R}_G(0.5) = 1.19$ and $\bar{R}_L(0.5) = \frac{1}{1.15}$.

- Putting this together explains why the average premium is similar (since $\frac{\bar{P}_p}{\bar{P}_c} \approx 1$):

$$\begin{aligned} \frac{\bar{P}_p}{\bar{P}_c} &\approx q\bar{R}_L + (1-q)\bar{R}_G \\ &= 59\% \times \frac{1}{1.15} + (1 - 59\%) \times 1.19 \\ &\approx 1 \end{aligned} \tag{6.31}$$

- The average market share on the other hand varies with ψ (figure 6.6(top left)). Due to the model design, market share does not change that quickly, there is 'loyalty' so companies using payback do not see sudden reductions in market share when they impose large premium increases;
- The average number of risk units over all simulations when $\psi = 0.5$ is 4668 compared to 4311 for the Control Company. Thus the average market share is around 8% higher when payback is used in this example (i.e. $\frac{4688}{4211} \approx 1.08$ (see also Figure 6.6(Top left))
- The market share of companies using payback is usually higher than the Control Company (since $q = 59\%, above$) . Loyalty ensures that this is not quickly eroded when premiums rise after a bad year;
- For example when $\psi = 0.5$ the proportion of time for which the payback company number of risk units (N_p) is less than the Control Company (N_c) is just $r = 21\%$.
- Let $\bar{N}_p$ be the average number of risk units written by a company using payback (proportional to its market share) and $\bar{N}_c$ be the average for the Control Company. Let $\bar{N}_{p,G}$ be the average number of risk units for a company using payback in the cases when $N_p > N_c$. Similarly $\bar{N}_{c,G}$ is the average premium of the control company when $N_p > N_c$. Let $\bar{N}_{p,L}$ be the average premium for a company using payback in the cases when $N_p \leq N_c$ and define $\bar{N}_{c,L}$ similarly. Let $\bar{S}_G = \frac{\bar{N}_{p,G}}{\bar{N}_{c,G}}$ and $\bar{S}_L = \frac{\bar{N}_{p,L}}{\bar{N}_{c,L}}$. Then, for example, when $\psi = 0.5$, $\bar{S}_G(0.5) = 1.12$ and $\bar{S}_L(0.5) = \frac{1}{1.05}$.

- Putting this together explains why the number of risk units is around 8% higher when $\psi = 0.5$

$$\begin{aligned} \frac{\bar{N}_p}{\bar{N}_c} &\approx r\bar{S}_L + (1-r)\bar{S}_G \\ &= 21\% \times \frac{1}{1.05} + (1 - 21\%) \times 1.12 \\ &\approx 1.08 \end{aligned} \tag{6.32}$$

When $\psi = 0.5$ the Company Value is around 4% less than the control, but such a company lives for around 15% less time than the control, on average. At death companies using payback typically owe more to policyholders than the Control Company so from a regulatory point of view such companies provides less protection for policyholders. Arguably the current payback rule is broadly neutral for policyholders (since premiums on average the same); worse for shareholders and worse for regulators.

Alternative payback rules are left for future investigations. These include asymmetric rules that raise premiums after a major loss but do not lower them for low losses; or time-average-rules that consider a longer period, or only payback when multi year profits are considerable.

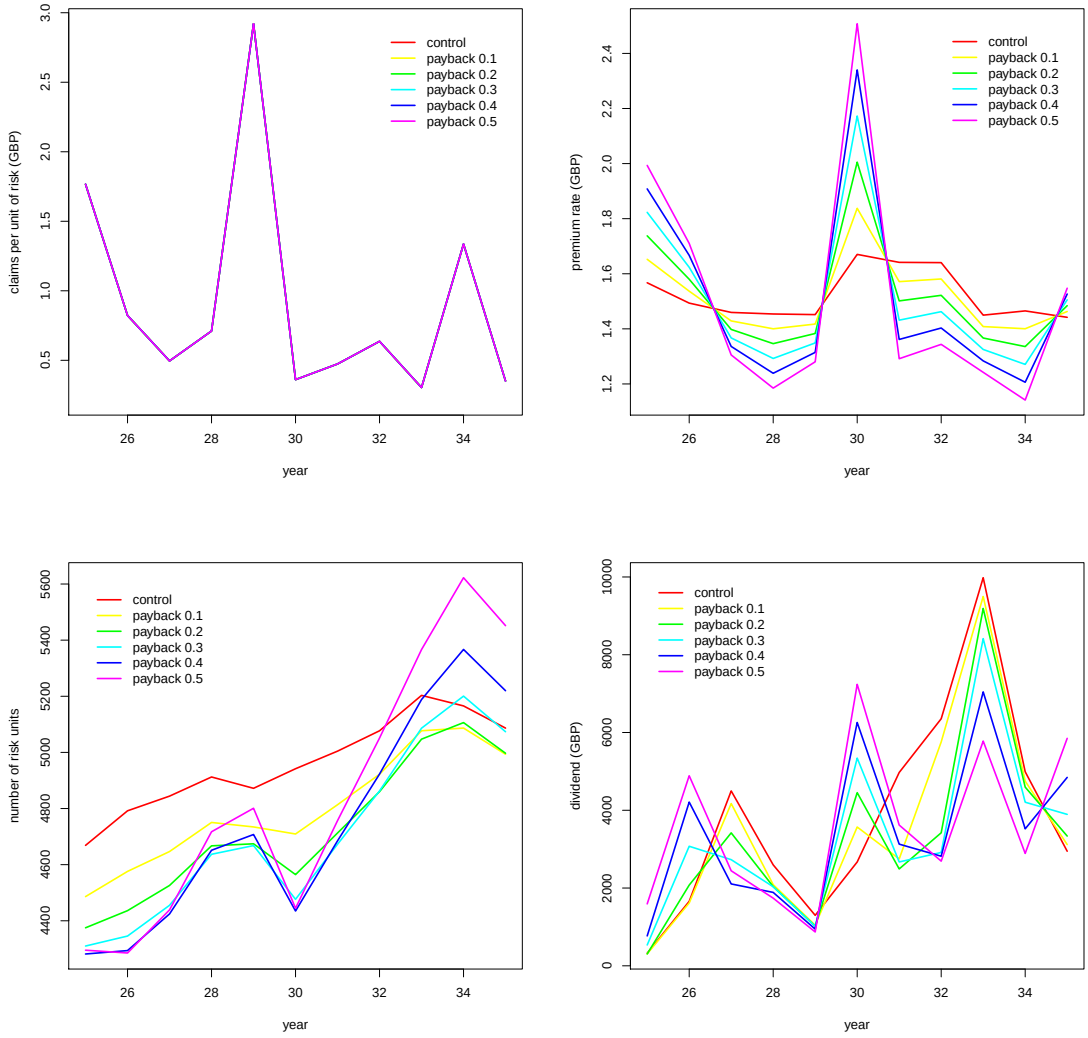


Figure 6.5: Experiment C6.4.x: Specific Simulation Plots for the main company in simulation 753 zoomed into years 25 to 35. Top left shows claims per unit of risk, top right shows premiums per unit of risk, bottom left shows the number of risk units sold by the main company by year and the bottom right shows the dividend paid in each year. Results are shown for the Control Company (red) and several levels of payback percentage between (10% and 50%).

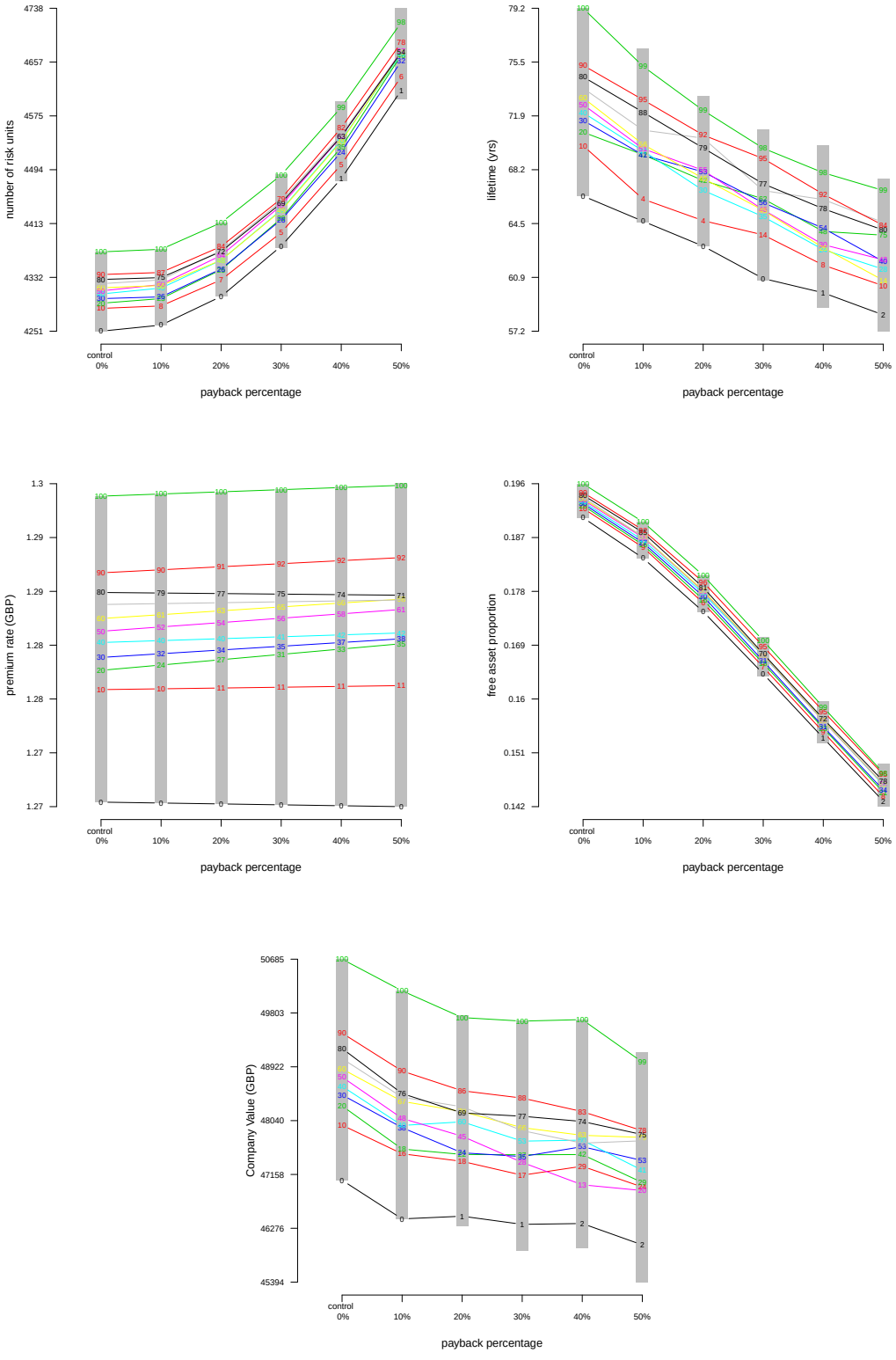


Figure 6.6: Experiment C6.4.x: Quantile Boxplots of key statistics (y-axis) for the control (0% payback) and other values of payback percentage from 10% to 50% shown on the x-axis. Top left shows the mean number of risk units across all simulations, top right shows average lifetimes, mid left shows the premium rate per unit of risk and mid right shows the average value of free assets, the bottom plot shows the present value of dividend payments for different payback rules.

Impact of different strengths for the regulatory test (C6.5.x) Let the **'Regulatory Return Period'** (τ_R) be defined as the reciprocal of the quantile which defines the regulatory test. For example in the Control experiment $\tau_R = 200$ years (i.e $b = 0.005$ in equation 6.6). This subsection investigates the impact of different Regulatory Return Periods. The experiment assumes the regulator has set consistent requirements for the whole market so that both the main company and its competitor are subject to the same rules.

Figure 6.7(top left) shows that, as expected, the lower Regulatory Return Period (and hence lower capital held), the lower the premium charged. Unsurprisingly, the less capital held the shorter the company lifetime (6.7(bottom right)). When $\tau_R = 50$ the average lifetime is around 25 years (or 50% of the Regulatory Return Period of 50 years), whereas when $\tau_R = 500$ the average lifetime is around 130 (just 25% of the Regulatory Return Period). This feature reflects the increasing difficulty of estimating percentiles from a limited set of data points.

The impact on the number of risk units sold (equivalently market share) is less obvious. One might expect that as the premium increases the market share should *fall*. In fact the market share actually rises as the Regulatory Return Period rises (6.7(bottom left)), apart from when $\tau_R = 50$ where the opposite occurs. As τ_R increases the main company premium rate grows slower than than the competitor (not shown) due to their different pricing distribution assumptions; therefore the market share of the main company rises. The atypical behaviour when $\tau_R = 50$ arises because, in this case, the competitor dies more frequently than the main company and this leads to regular additions in market share for the main company (see equation 6.15) which increase the average number of risk units. Figure 6.7 shows that as τ_R increases the Company Value increases.

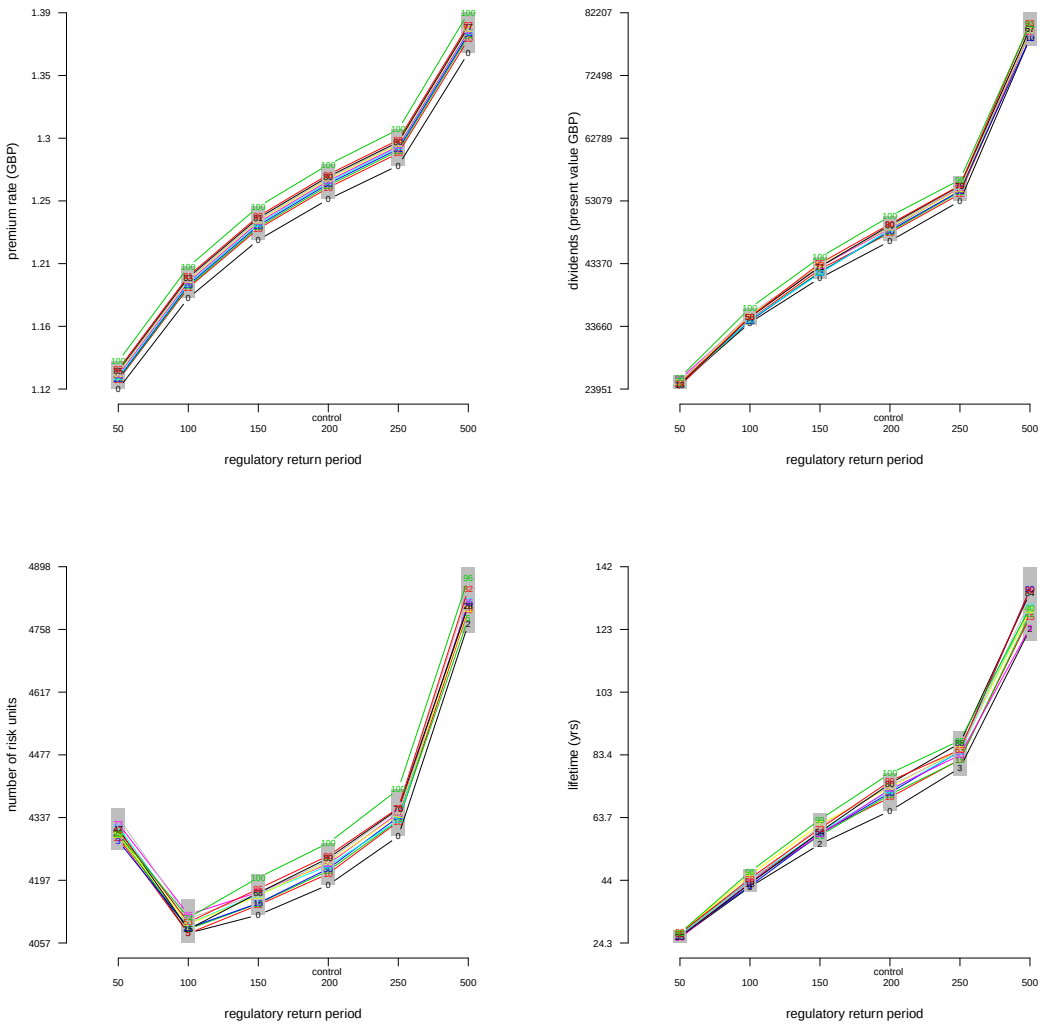


Figure 6.7: Experiment C6.5.x: Quantile Boxplots for main company showing key statistics for different regulatory capital requirements (VaR levels from 50 years to 500 years). Top left shows mean premium rate per unit of risk, top right shows the mean present value of dividends paid, bottom left shows the average number of risk units sold and bottom right shows the average lifetime.

Impact of using TVaR **for the regulatory test (C6.6)** Recall that in the control experiment the claims process follows a Lognormal distribution with mean 1 and variance 0.65. In order to compare a TVaR risk measure (equation 6.7) with the VaR measure used in the Control Experiment it is necessary to find which quantile value (b_{TVaR}) leads to the same capital levels in theory: specifically we wish to solve $\text{TVaR}_b(X) = \text{VaR}_{0.005}(X)$. A return period of 69 years (i.e. $b = \frac{1}{69}$) was determined by simulating 1,000,000 claims (call this the 'VaR/ TVaR Equivalence Calculation' below). This was repeated 10 times and 7 out of 10 cases led to a return period between 68.5 and 69. The Regulator is assumed to require this test of the whole industry so both the main company and its competitor use a TVaR to calculate capital. Call the main company in this experiment the 'TVaR Company'.

The average lifetime of the TVaR Company is 71.7 years compared to 72.7 for the Control Company. Hitchcox et al [105] argue that a TVaR measure is more stable than a VaR measure so it is a surprise that companies using this approach actually live for shorter periods. The two companies live for the same amount of time in almost all experiments (976 out of 1024). When the lifetime is different, however, the Control Company (using VaR) lives longer in 33 cases and shorter in just 15 cases. To see why this occurs the following algorithm was run:

TVaR Comparison Algorithm

- For $j \in \{1, ...N_j\}$;
- for each j, simulate 15 years of past claims data from the true underlying distribution;
- **Determine estimated distribution parameters** compute the mean ($\hat{\mu}_j$) and variance ($\hat{\sigma}_j^2$) from this 15 years;
  - for $k \in \{1, ..N_k\}$ (and given j)
  - Estimate $\text{VaR}_{j,k}$ and $\text{TVaR}_{j,k}$ by sampling a further 10,000 values from the estimated claims distribution (Lognormal with mean $\hat{\mu}_j$ and variance $\hat{\sigma}_j^2$);
  - let $I_{j,k} = 1$ if $\text{VaR}_{j,k} > \text{TVaR}_{j,k}$ and zero otherwise.
- Determine the proportion $F_j = \frac{\sum_{k=1}^{M} I_{j,k}}{M}$ that the estimated VaR exceeds TVaR.

The TVaR Comparison Algorithm was run with $N_j = 10000$ and $N_k = 2^7$. Each coordinate $(\hat{\mu}_j, \hat{\sigma}_j^2)$ is plotted in figure 6.8; the point is coloured purple if $F_j < 0.5$ and orange if $F_j \geq 0.5$. The figure illustrates the joint probability density of the estimator of the mean and variance of the claims process and shows that even though the underlying mean is 1 and variance is 0.65 the estimated parameters vary considerably. Values of 0.5 to above 2.0 are observed for the mean and from just above 0 to as much as 15 for the variance (NB the value in the graphic for variance is on a log2 scale). The number of cases for which the VaR exceeds TVaR over 50% of the time is 7178 out of 10000. Although the VaR and TVaR should be the same, it is more likely that, with just a 15 years claims history, the estimated parameters will cause the VaR to be *greater* than the TVaR due to parameter estimation error. It was shown (page 317) that when the estimated capital requirement is higher the company charges more premium and lives longer thus explaining why the Control Company has a longer average lifetime than the TVaR company. From a regulatory perspective VaR leads to errors in a more prudent direction.

The average number of risk units of the TVaR Company is 5335 compared to 4496 for the Control Company. This is because the competitor of the TVaR Company is relatively *less* competitive than the competitor of the Control Company, for the following reason. The 69 year return period was chosen to give approximate equivalence between TVaR at that return period and VaR at a 200 year return period. This relationship is specific to the Lognormal distribution (with mean 1 and variance 0.65). The competitor assumes a Pareto distribution when calculating capital. For a Pareto distribution with mean 1 and variance 0.65 the VaR / TVaR Equivalence Calculation leads to $b_{TVaR} = \frac{1}{56}$. By definition $\text{TVaR}_{\frac{1}{69}} > \text{TVaR}_{\frac{1}{56}}$ hence the use of a 69 year return period by the competitor leads them to estimate higher capital requirements than they would by using the $\text{VaR}_{0.005}$ measure. These higher capital requirements lead to higher premium rates and therefore lower competitiveness on average. Consistent with this, the proportion of simulations where the competitor dies after the main company rises from 26% to 32% due to the additional capital held by the competitor. The Company Value of the TVaR Company is GBP 52,011 compared to GBP 48,777 for the Control Company due to its higher average number of risk units, despite its average lifetime being shorter. This result highlights that

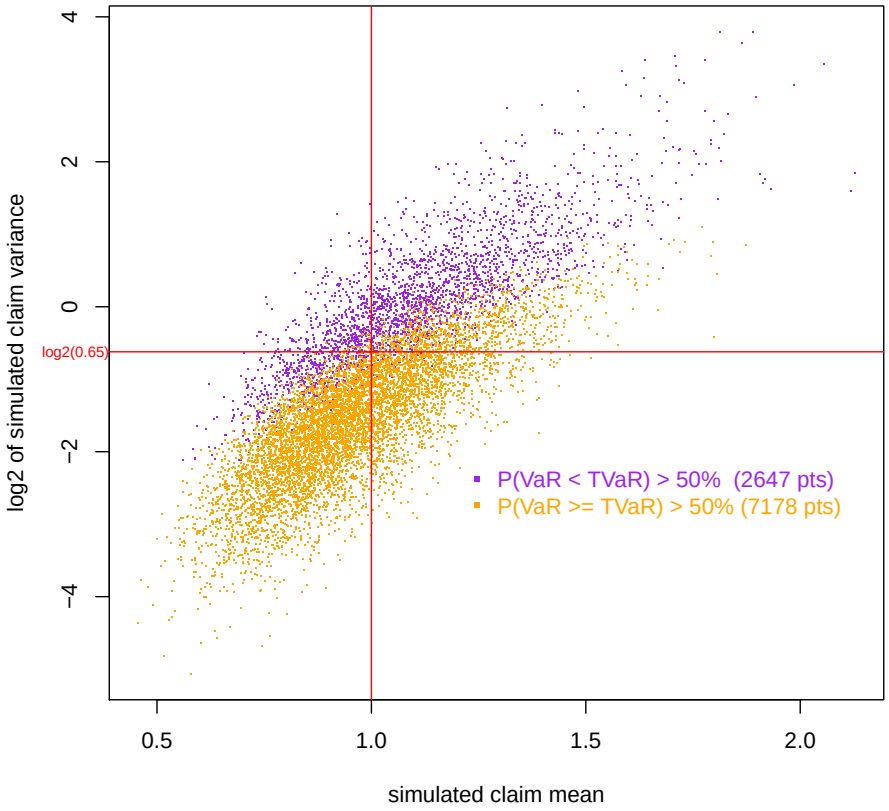


Figure 6.8: Experiment C6.6: Scatter plot showing estimated mean $\hat{\mu}_j$ (x-axis) and variance $\hat{\sigma}_j^2$ (y-axis on log2 scale) of claims process from $j \in \{1, ...10,000\}$ 15 year samples. For each point $(\hat{\mu}_j\ ,\hat{\sigma}_j^2)$, 128 estimates of the $\text{VaR}_{0.005}$ and $\text{TVaR}_{\frac{1}{69}}$ are produced by sampling 10000 values from the Estimated Claims Distribution. The orange dots show the pairs for which the VaR $\geq$ TVaR more than 50% of the time and vice versa for purple. In 7178 cases the proportion of VaR estimates that exceed the TVaR is greater than 50% demonstrating that the capital calculation is more likely to result in a higher VaR calculation than TVaR in this experiment.

the impact of a regulatory change on a given company depends on how the change impacts its competitors as well as the effect on itself.

Use of incorrect pricing distribution assumptions (C6.7.x) The Control Company does not know the true mean and variance of the claims distribution but does know its true form (Lognormal). In practice the company's claims analysis may cause them to choose an incorrect distribution family[10]; this risk is called

[10]Where a 'distribution family' is taken to mean one of the standard statistical distributions: Gaussian, Gamma, Pareto etc

'specification error' in Hooker et al [106] . This experiment illustrates the impact on company lifetime of choosing either a Gamma or Pareto distribution for pricing. The main companies in these two experiments will be referred to as the Gamma Company and the Pareto Company. Other parameters are the same as the Control Experiment, in particular the main company still uses only 15 years of data for pricing and the competitor's pricing approach is wholly unchanged. Figure 6.9(top left) illustrates that using a Gamma distribution shortens the average company lifetime, and assuming a Pareto distribution lengthens it. Figure 6.9(top right) shows that average levels of capital in the Pareto Company are lower than the Control Company and figure 6.9(bottom right) shows higher premium rates on average. Lower capital, other things equal, would typically lead to a shorter lifetime but this is offset by higher premium rates. Closer analysis of the capital calculations show that average levels of capital do not explain the simulation specific behaviour. Recall that the true mean of the claims process is 1 and its variance is 0.65. Let $\text{VaR}_{b,d}$ denote the Value at Risk for quantile b and distribution d. Then 95% confidence intervals are: $\text{VaR}_{0.005,Lognormal} \in (4.78, 4.84)$ and $\text{VaR}_{0.005,pareto} \in (4.69, 4.78)$ estimated from a sample of size $N_a = 2^{20}$; so with 95% confidence we can say $\text{VaR}_{0.005,pareto} < \text{VaR}_{0.005,Lognormal}$. Recall that the main companies estimate the VaR from 15 years of data leading to significant sampling error. The following algorithm was used to assess this:

- for $j \in \{1, ...N_j\}$
- **Sample 15 years of claims data** Let $C_1, ...C_{15}$ be a sample from a Lognormal distribution with mean 1 and variance 0.65.
- **Sample claims from pricing distribution** For a given pricing distribution d, with mean $E_C = \frac{1}{15}\sum_{k=1}^{15} C_k$ and standard deviation $\sigma_C = \sqrt{\frac{1}{14}\sum_{k=1}^{15}(C_k - E_C)^2}$ sample 10000 modelled claims $\hat{C} = \{\hat{C}_1, ...\hat{C}_{10000}\}$.
- **Calculate** VaR Let $V_j = \text{VaR}_{0.005}(\hat{C})$

This algorithm was run (with $N_j = 2^{15}$) for $d = Pareto$ and $d = Lognormal$ showing that the capital estimated by the Pareto company is actually larger than the Lognormal Company 66% of the time even though the mean estimated capital

is higher from the Lognormal distribution. Since the capital estimate at the start of the year is independent of the claim arising in the year there is therefore a greater probability that the Pareto company will have both a higher premium and more capital to survive a large claim than the Lognormal company. This situation is illustrated by considering simulation 820 (figure 6.10). Note the relatively large claims spike at year 25 after a run of below average claims for the previous 7 years. The Control and Gamma Companies die at this point as they have charged a lower premium and are holding less capital; the Pareto pricing company just survives. Note that the claim in year 45 which eventually kills the Pareto pricing company is much smaller than the claims it survived in year 25. The reason for this is that there is a really long run of below average claims prior to this; in particular the large claim in year 25 has been 'forgotten' in the 15 year moving average pricing approach. This illustrates the importance of retaining institutional memory; and also cautions against assuming that long runs of benign claim years represent a shift in risk. Note the difference in this setting than the payback example considered previously. In the payback example although the premium rate differed between companies the total of capital charged to the policyholder and premium was constant between them - as the premium went up, the required capital went down. This was because all the main companies (regardless of payback rule) used the same underlying claims distribution assumption. Conversely when different assumptions of the underlying claims are used, the estimated capital requirements will also differ.

Figure 6.9(bottom) illustrates the impact of pricing distribution choice on company value. The Pareto Company value (GBP 51,776) is greater than the Control Company (GBP 48,777) despite using the wrong distribution family and writing less business. This is due to the Pareto Company's longer lifetime (111.8 years vs 72.6 for the Control Company); the value of discounted dividends from the Pareto Company up to the point the Control Company dies is GBP 47,756 which is less than the Control Company.

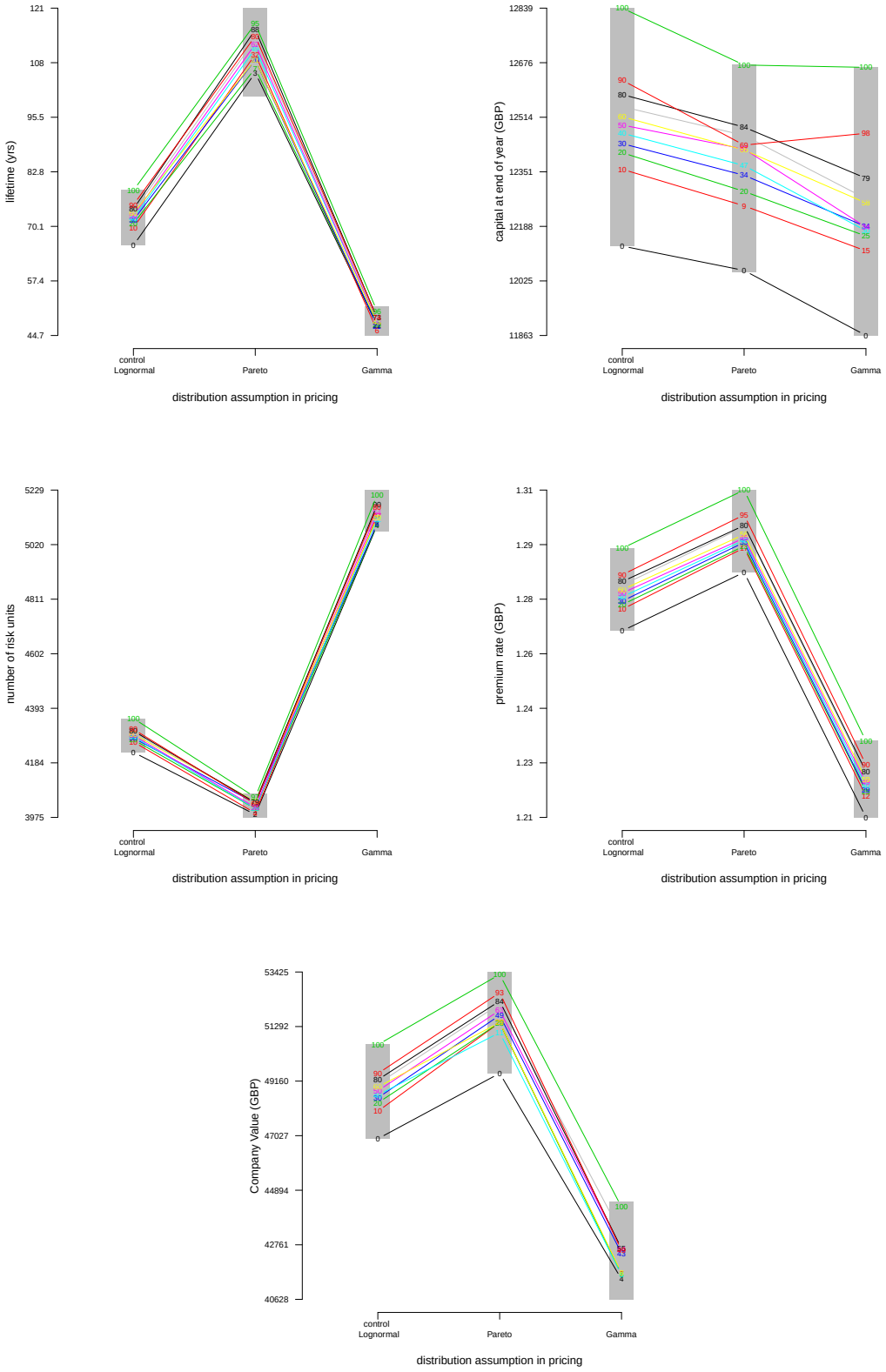


Figure 6.9: Experiments C6.7.x: Quantile Boxplots for the main company for different underlying claims distribution assumptions used in calculating the premium rate. Shown for the control (Lognormal), Pareto and Gamma distributions. Top left shows average lifetime, top right the capital held at the end of the year, bottom left shows the number of risk units sold on average and bottom right shows the premium rate per unit of risk.

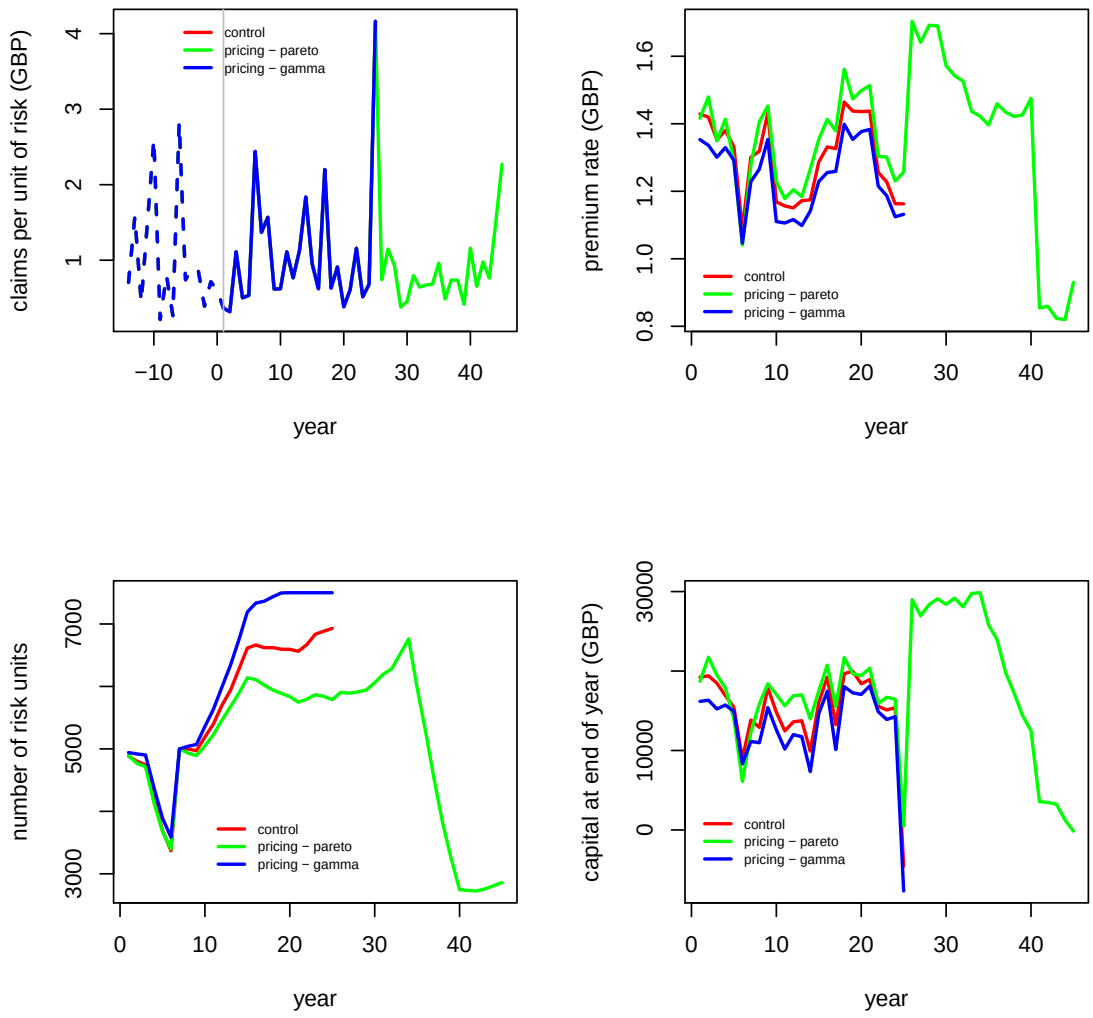


Figure 6.10: Experiments C6.7.x: Specific Simulation Plots for the main company and simulation 820. Shown for each pricing distribution assumption: Lognormal (control, red), Pareto (green) and Gamma (blue). Top left shows claims per unit of risk, top right shows the premium rate, bottom left shows the number of risk units sold and bottom right the capital held at the end of the year.

Impact of different underlying claims distributions (C6.8.x) So far the underlying claims distribution is Lognormal with mean 1 and var 0.65. This section illustrates the impact of distributions with different variance (specifically var = 0.5 and var=0.8) . For example this would illustrate how different classes of business might behave. Note that the underlying claims affects both the main company and the competitor - they both continue to price using the same approach as before.

Figure 6.11 shows that as the variance of the underlying claims increases the average lifetime decreases. At first look it might seem obvious that a more volatile claims process will lead to more company deaths . Note, however, that each company holds capital to survive a 1 in 200 level of risk. For a riskier claims process more capital will be held. Hence, other things equal there should be no impact on lifetime. As the variance of the underlying claims increases, however, it becomes more difficult to estimate the distribution parameters and this explains the difference. Indeed the probability of the company calculating capital that is too low when the underlying variance is 0.5 is around 69% compared to 78% when it is 0.65 (the control) and 83% when it is 0.8.

As expected the premium rate rises as the underlying claim variance rises. The companies may mis-estimate the capital requirement more often when the volatility increases, but they do estimate *higher* variance (if not the correct value) and hence higher capital requirements (and hence higher premiums). Not all profits are paid out in dividends, some are retained to bolster the free assets - and this is why the free asset proportion is higher in the high variance case.

The policyholder (and regulatory) perspective is not as positive however. As the underlying variance rises, the average shortfall at company death (i.e. the amount by which the company fails to pay 100% of the claim) rises from 19.9% of the claim to 23.6%.

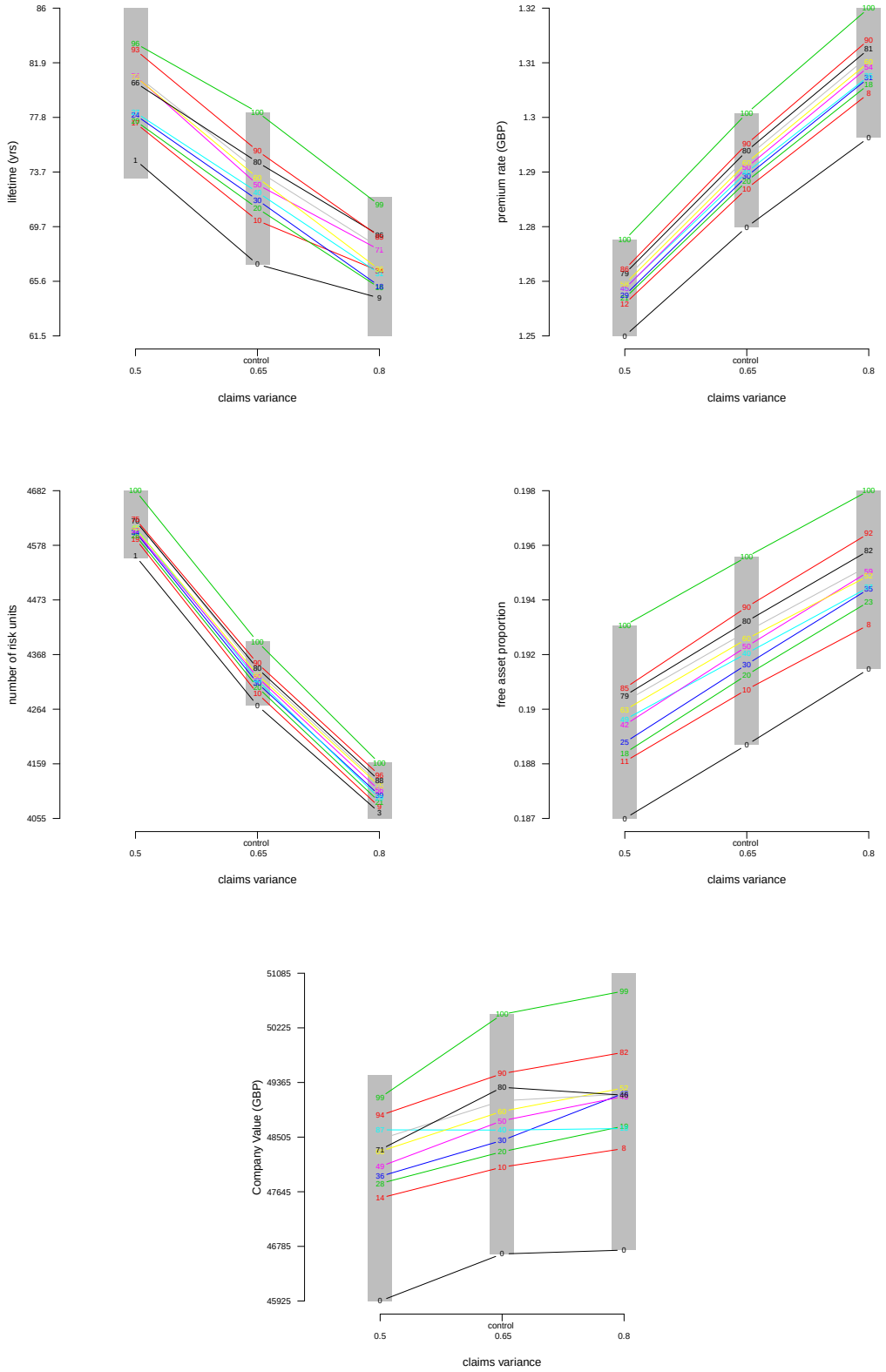


Figure 6.11: Experiments C6.7.x: Quantile Boxplots for main company for different levels of underlying claims variance 50%, 65% (control) and 80%, where distribution is Lognormal with mean 1 in all cases. Top left shows average lifetime, top right premium rate per unit, middle left shows the number of risk units sold, middle right the proportion of the company assets that are free and bottom shows the Company Value

6.6 Summary and conclusions

This chapter presents a quantitative examination of a qualitative insurance industry model; like the MONIAC, the model presented here is informative but not an accurate representation of reality. Both the model and the experiments carried out in this chapter are all believed to be new. Computer modelling within the insurance industry has increased materially in its complexity and power over the last three decades [61, 63, 218, 219, 244]. Models used in practice within insurers, however, are still optimised to answer questions over short time horizons [83], and typically exclude non-trivial longer term effects and omit competition effects. These limitations constrain the questions that can be asked of current operational models resulting in misleading impressions of the impact of future strategies. This chapter describes a simple model of an insurance industry with two competing companies to address these issues. Competition is defined by a functional relationship of relative premium rates taking account of loyalty. The simplifying assumptions are made are not thought to significantly affect the main conclusions and certainly of much lower importance than the introduction of competition within the model. The main conclusions are as follows:

1. Even when the Regulatory Return Period is 200 years the control main company lives for just 73 years on average due to the impact of parameter uncertainty when estimating capital;

2. Even if the true claims process is known by the main company there is still a wide variance in the central 80% of lifetimes from 40 to 900 years. The average, however, does exceed 200 years in this case;

3. When the shareholders target return is lower, the average company lifetime is shorter due to a lower capital buffer being retained;

4. For lower target returns, company value is lower despite attaining a higher market share. Company value does not appear to increase when the target return is higher than the control experiment, however, because of reduced volumes of business sold;

5. Payback (as defined in equation 6.5) increases premium volatility, leads to shorter lifetimes and increases market share while lowering company value;

6. For all tested variations in the Regulatory Return Period (50,100,150,200,250 and 500 years) the actual lifetime falls short of the target lifetime. As the VaR threshold increases lifetimes increase, however, the proportion to target lifetime decreases;

7. In an equivalent $TVaR$ regulatory regime the company lifetime is shorter than for VaR. Note that typical industry views [105] are that $TVaR$ is a better and more resilient metric;

8. When the true claims distribution is not known: the use of a Gamma distribution in pricing leads to a shorter average lifetime and the use of a Pareto distribution to a longer lifetime;

9. Some classes of insurance business (e.g. reinsurance) typically have a higher variance than others (e.g. motor insurance). To approximate companies with different mixes of business different underlying claims variance levels were, therefore, tested. In each case the companies are subject to the same Regulatory Return Period and are therefore expected to have a similar average lifetime; this is not observed, as variance increases the average lifetime decreases;

10. If a company uses five years of past data for estimating pricing, its average lifetime falls to just 14 years compared to 73 years when 15 years of data are used.

It is noted in conclusion 2 that even in a perfect model situation the company lifetimes vary dramatically. This is not a failure of the regulatory regime, but may be perceived so. Criticism of regulators (by politicians or the media for example) is unjustified if company insolvencies arise from the natural variability of the claims process.

The asymmetric effect of raising or lowering the target return on capital is noted in conclusion 4. In the former case premiums rise, business volumes fall and company value remains broadly constant; in the latter market share rises but annual profits

fall and in this case the company value falls. This was an interesting competition effect presumably linked to the fact that under the control experiment the main company and the competitor are chosen to have similar length of lifetimes.

Several of the effects listed above arise due to poor estimation of parameters given sparse data, a non-trivial problem for insurers. For example conclusion 7 relates to $TVaR$ a measure that incorporates an estimate of all claims above a threshold, including in the deep tail of the distribution. Large, low probability claims are the most uncertain part of the estimated claims distribution. For the heavy tailed Lognormal distribution it turns out that it is more likely the capital estimated using a VaR measure will be higher than using $TVaR$ (even though the capital values should be the same theoretically). This is why companies in the VaR regime live slightly longer. This was an initially unexpected effect as $TVaR$ is generally considered to be a good risk measure with desirable coherence [14] properties. Company lifetime decreases as volatility increases (conclusion 9). For each level of volatility the company is still operating in the same regulatory regime, however, and the regulatory aim is that all companies should hold sufficient capital to achieve the same maximum failure rate (1 in 200 years). Lifetime falls as volatility increases because of poorly estimated parameters in the capital calculation. The chance a company understates capital increases with the underlying variance.

This chapter does not include explicit consideration of real options [168] or optimal control [41] discussed in section 1.3. The investment model is not the focus of this chapter and so a fixed investment return is assumed rather than including state price deflators [12] accompanied by a more typical asset mix of equities, corporate bonds etc. This is not thought to be flaw, particularly because the insurance payouts are not dependent on the investment returns achieved (i.e. the model considers non-systematic risk rather than systematic risk [43]). This contrasts with With-Profits life assurance where fair valuation techniques are essential to capture the embedded option values [118, 235]. Inclusion of stochastic investment returns is one of several areas in which the model could be extended. Management actions (such as reviewing prices each year according to incoming claims information) are implicitly included in the model presented here. Other path dependent [280] management actions (real options) could be included in the model and the Company

Value would adjust accordingly. Such methods could be used to assess the value of potential actions such as: the use of forecasts taking account of their costs, the value of any policyholder options included in the contract, or the acquisition of new data to improve understanding of risk, leading to better pricing.

The insurance industry model presented in this chapter assumes that the claims process is stationary. It would be informative to introduce a non-stationary process such as a trend or cycle and asses the impact on company lifetime using climatology pricing. It would then be possible to investigate the impact of a variety of pricing techniques that made allowance for trends on average lifetimes and profitability. The first step towards allowing for trends in pricing is detecting them and some thoughts on trend detection are included in appendix C. Bayesian techniques are not explored in this thesis and it would also be useful to consider how they can be incorporated into the insurance workflow; such techniques could also be explored and quantitatively assessed using the model presented in this chapter.

Chapter 7

Concluding remarks

This thesis considers a number of statistical topics with particular relevance to insurance. A primary focus has been extreme risks arising from dynamical systems and methods to forecast and account for them, to explore whether scientific models can improve insurance pricing.

The concept of a 'skill score' was considered and properties of skill scores were listed and discussed in an insurance context. The properties of Propriety [274] and Locality [165] are considered to be important. A number of skill scores have been suggested in the reviewed literature and those wishing to assess forecast skill have to choose which ones to use. This thesis suggests that skill scores should be avoided if they repeatedly give a good score to forecasts that ascribe low probability to events that actually arise. In the context of insurance such forecasts could lead to mis-pricing and hence skill scores that penalise them should be used in favour of scores that do not. A new property, named Feasibility, is introduced to highlight scores that do not have this perceived flaw. The Ignorance [69, 97, 216], Brier [20] , Power Rule [224] and Spherical [86] scores all have the Feasibility property whereas the commonly used CRPS [82] and Mean Squared Error [82] scores do not. The thesis concludes that the CRPS score should not be used in an insurance context and suggests other users should consider whether its lack of Feasibility constrains its usefulness in their context.

Three experiments were then carried out to quantitatively test different skill scores and thereby create a ranking in each case. The first experiment sampled from a known distribution and then used various skill scores to evaluate a family of

forecasts indexed by a bandwidth parameter and containing the true distribution. For each score, Optimal Score Estimation [69,94] was used to select the best scoring forecast which could then be objectively compared with the true distribution. A ranking amongst scores was then produced by selecting the score that picked a forecast whose bandwidth was closest to that of the underlying distribution over multiple samples. In this example the Ignorance score was the highest ranking. The non-proper scores performed worst with the Mean squared error being shown to fail systematically. The CRPS score performed worst amongst the Proper scores.

The second experiment considered Gaussian underlying distributions where the number of observations are sparse. Similar to experiment 1 a limited family of forecasts was considered one of which is correct and the number of times the skill scores identify the true forecast is counted and compared. Again the Ignorance score performed best.

The third experiment uses the Skill Gap (an extension of the Information Deficit [69, 216] to other scores) to measure how quickly a structurally incorrect forecast can be 'rejected'. Rejection does not mean that a given forecast has no useful information, just that it has been shown to be inconsistent as a representation of the underlying pdf of outcomes. For a stated level of confidence the time to Reject a forecast (for a stated level of probability) can be determined. Skill scores can then be ranked objectively by comparing their rejection times. This experiment was only carried out for the Spherical, Proper Linear, Naive Linear and Ignorance scores. Unexpectedly the rejection time for the first three scores was the same and the reason for this was explained. This experiment showed that the best skill score was dependent on the underlying distribution, leading to the conclusion that the use of multiple proper scores to assess forecast skill can be useful rather than relying on one only.

It is my view that further skill score comparison experiments should be carried out in future to establish a preferred list of skill scores; such a programme of work can be used to determine in which situations different skill scores should be used. It is hoped that the work presented here can add to that overall goal.

Insurance contracts often cover risks arising from dynamical processes including: floods, large scale windstorms and hurricanes [138,175,248]. Many of these processes

are forecasted by meteorological offices or other organisations [169, 181]. In order to explore whether such forecasts can be useful in insurance a suitable proxy system, based on the Lorenz 96 models [155], was explored. Lorenz 96 system II was used as a proxy for reality (the 'System') and then Lorenz 96 system I (with extensions) was used to produce various 'Models'. Multiple parameterisations of the Lorenz system II were considered and in each case five different models were developed. The models included three examples with fixed forcing terms, one with a deterministic dynamic functional relationship and one with a stochastic parametrisation [11, 129, 239, 271–273]. Reduced coupling for a given level of forcing was expected to lead to increased predictability but this did not arise because the effective forcing in the system actually increases when coupling decreases and this is shown to be the dominant effect. Climatology blending [30] was used to derive the best scoring blend between the forecast and climatology at each forecast lead time over a chosen period of 24 observations. The blending parameters are observed and their behaviour illustrated and explained. As expected the weight put on the forecast diminishes as lead time increases and the kernel width increases; however the fact that these appear to operate in series rather than parallel was not anticipated.

Armed with Lorenz 96 as a proxy System, an insurance index [5, 6, 39, 185, 269] is developed which is then priced using traditional techniques similar to the method proposed by Kreps [128] which incorporates an expected cost and a return on equity component. The Models are then used to forecast the index. A best scoring relationship between the Models and System is developed using a technique, named ϕ-transformation, that is thought to be novel. This method uses Optimal Score Estimation methods to determine the Score Optimal Piecewise Linear Relationship between observed Index values and Modelled values. ϕ-transformation is shown to work in two idealised situations and then used in the Lorenz 96 example. Two pricing methods then make use of the ϕ-transformed forecasts to adjust prices period by period. The first method updates the expected index value (described as the 'Updated Expectation' method) and the second carries out climatology blending to update both the expected index and also the return on equity element of the price. A simple rule is introduced to cater for competition effects and, under these assumptions, the Updated Expectation method gives a lower price than the

traditional method, on average, yet leads to fewer company insolvencies and higher profitability. As such, this shows that the use of forecasts in this example has benefitted Policyholders, through lower average premiums; shareholders, through greater profits and, arguably; regulators, through a lower risk of insolvency. The method assumes that the hypothetical company will reduce the amount of insurance sold when large losses are forecast and, therefore, those companies not using the proposed method are assumed to take up the excess business. The latter companies will be more likely to fail as a consequence. Ultimately it would be hoped that many companies would adopt the more sophisticated pricing methods and then competition effects would not limit the price, in which case the insurance industry as a whole would be less likely to see insolvencies following major events. The new approach would lead to increased price volatility, however, but this would provide an annual indicator of expected risk levels which, it is hoped, may lead to adaptive behaviour from policyholders which overall is thought to be a positive outcome.

In the Lorenz 96 example the index payouts are closely related to the behaviour of the system and therefore skillful models can add value. To test the value of forecasting for insurers in a more challenging setting a simple model of hurricane losses is created. Basin hurricane counts are sampled from a Poisson distribution, then successively sub-sampled to create landfall numbers and major city hits. The severity of landfalling storms is also simulated as are insurance losses which are modelled by assuming a simplified 1-1 relationship between hurricane strength and loss. This model is not sophisticated when compared to operational catastrophe models [211]; but such detail was not deemed necessary to discursively explore the role of forecasts within pricing in this setting. It is shown that simple business volume scaling methods that react to forecast information can improve expected profitability. More sophisticated forecasts, if used in pricing, can lead to reduced capital requirements in quiet years, but would lead to lower profitability unless steps were taken to ensure premium levels are on average no lower than traditional pricing methods. Where pricing is changed it is assumed in this example that the whole market have adopted the pricing method. Finally it is noted that, given the natural variability of this system, it is very difficult to distinguish between an underwriter who is good or just lucky.

The final chapter of the thesis explores the impacts of competition more thoroughly. The complexity of computer modelling in the insurance industry has increased over the past 40 years since models were first used to explore better ways of calculating solvency risks [61,63,218,219,244]. Such models have tended to focus on single companies [50,62,195,219,220], however, and it is far rarer to find analysis of insurance markets in the literature, though some have been carried out [250,277]. This thesis proposes a novel insurance market model which includes competition between two companies and captures the main processes in the insurance industry [61,105,250]: customer loyalty, premium rating, capital setting, investment returns, claims payments and dividend payments. The pricing method and underlying assumptions used by the two companies is specified in advance after which the behaviour and outcomes emerge from the simulation. The model provides a framework to investigate such questions as: will the companies survive for as long as specified by the regulatory test, on average? Does profit sharing (or 'payback'), which is used in some re-insurance markets, lead to a better or worse investment for shareholders? Would a different regulatory method lead to a stronger or weaker industry? Is a company always better off if it sets premium rates using the same distribution family from which the claims arise? These questions are answered, sometimes with, initially, surprising results. For example the $TVaR$ measure (equation 6.7) is shown to reduce expected company lifetime due to estimation difficulties in the presence of parameter uncertainty, at odds with common industry views [105] that this is a better and more resilient metric. A profit sharing or 'payback' rule is shown to reduce expected company lifetime and lowers Company Value. For all tested regulatory return periods the company lifetime is shorter than the regulatory target lifetime due to errors in estimating the parameters of the underlying claims distributions. It is not claimed that these features would be replicated in practice; but the findings suggest that regulators would be advised to invest in the development of comprehensive insurance market models covering multiple classes of business. Such models would allow the advance testing of regulatory changes and would enable the resilience of the insurance industry to systemic threats to be assessed. The model shows, for the chosen parameters, that companies exposed to more volatile risks are likely to have shorter lifetimes despite their regulatory targets being the same strength, due

to a magnification of the effect of parameter uncertainty in such cases. The model can be developed in future to incorporate the work of earlier chapters to consider different dynamic claims processes, and also the use of forecasts in a competitive environment. The impact of trends and cycles could also be considered, following appropriate trend detection methods as discussed in appendix C. There is no discussion in this thesis of Bayesian methods and it may be possible to re-express the work within a Bayesian framework which could be a useful and appropriate setting for future work.

This thesis concludes that the use of Proper, Feasible skill scores can help insurers to use forecasts to improve pricing. But this must be done carefully and, in particular, model outputs should be transformed to have the highest average skill when compared to the system.

Appendix A

CRPS favours median observations

The following gives a proof that the CRPS score will always give the best score when an observation arises that is at the median of the distribution of possible outcomes. The CRPS is defined in equation 2.10. Let $P(v)$ represent the Cumulative Density Function of the forecast p, so that:

$$P(v) = \int_{-\infty}^{z} p(t)dt$$

Then the CRPS is equivalently defined as:

$$S(p, v) = \int_{-\infty}^{\infty} (P(z) - H(z - v))^2 dz$$

The derivative of this with respect to the observation v is defined as:

$$\frac{dS}{dv} := \lim_{\delta \to 0} \frac{S(v + \delta) - S(v)}{\delta}$$

Now,

$$S(v + \delta) - S(v) = \int_{-\infty}^{\infty} (P(z) - H(z - (v + \delta)))^2 dz - \int_{-\infty}^{\infty} (P(z) - H(z - v))^2 dz$$

Using the rules of integration the two integrands can be brought within a single integral with the same limits. Squaring the bracketed terms and cancelling the repeated $P(z)$ terms and noting that $H^2(x) = H(x)$ by definition of the Heaviside function, the above equation reduces to:

$$S(v+\delta) - S(v) = \int_{-\infty}^{\infty} (-2P(z)+1)(H(z-(v+\delta)) - H(z-v))dz$$

Note that:

$$H(z-(v+\delta)) - H(z-v) = \begin{cases} 0 & \text{if } z < v \\ -1 & \text{if } v \leq z < v+\delta \\ 0 & \text{if } z \geq v+\delta \end{cases}$$

so that the integrand and limits of integration can be changed as follows:

$$S(v+\delta) - S(v) = \int_{v}^{v+\delta} (2P(z)-1)dz \approx (2P(v)-1)\delta$$

Therefore the derivative of S with respect to the observation v is as follows:

$$\frac{dS}{dv} := \lim_{\delta \to 0} \frac{(2P(v)-1)\delta}{\delta} = 2P(v) - 1$$

As noted in the section 2.3.1 this derivative is zero when $P(v) = \frac{1}{2}$ i.e. when the observation v is at the median of the forecast p.

Appendix B

Experiment C4.1.x ϕ' definition

Table A.1: Experiment C4.1.x - definition of ϕ'

x	$y = \phi'(x)$	x	$y = \phi'(x)$	x	$y = \phi'(x)$
0.000000	5.403489	34.878013	42.425648	73.429516	69.575231
3.806653	12.807920	36.891898	43.483424	77.457285	71.338191
5.245142	17.039024	38.042689	43.483424	80.909659	73.101151
5.532840	20.917536	39.481178	44.188608	85.225125	75.216703
6.971328	22.680496	40.344271	46.304159	86.088219	76.274479
9.272911	25.148640	42.358156	50.887855	88.102103	78.742623
12.149889	25.853824	42.645854	54.766367	89.252894	80.152990
14.451471	27.264192	44.947436	58.292287	90.115988	82.973726
17.040751	28.674560	48.399809	60.760431	92.417570	86.147054
19.342333	29.379744	50.701391	60.760431	93.856059	89.320382
21.356218	31.142704	54.153765	62.875983	95.294548	90.730750
22.507009	32.553072	57.318440	64.638943	96.733037	92.141118
24.233195	34.668624	59.620023	65.344127	97.020735	94.256670
25.671684	36.078992	62.209303	65.696719	98.459223	96.724814
28.260964	38.547136	63.647792	66.401903	99.610015	97.192958
30.850244	41.015280	67.100165	67.459679	100.000000	98.000000
32.864129	41.720464	70.840236	69.222639		

Appendix C

Trend detection

'***The rapid development of computer speeds and storage capacity should by now have relegated most of estimation theory to footnotes about numerical approximations and refocused attention on all of the issues surrounding methodology, inference, model formulation and equation selection***' Hendry, 1980 [102]

'Trends' are important to the insurance industry and so it is unsurprising that trend detection algorithms abound [263]. Traditional pricing methods rely on past claims data [193] and it is useful to know the extent to which these are still relevant for predicting risk levels in the future in order to maintain profitability and preserve solvency [85]. Increases in price are difficult to introduce into a competitive marketplace [139], as argued in Chapter 6. Misinterpreting a trend detection algorithm could therefore lead to loss of market share or the acquisition of unprofitable business. Extrapolation is, however, not discussed in this appendix. This appendix focusses rather on what it means to say a trend has been observed in a data set.

What is a trend? There are many methods to detect trends, but the concept of trends has multiple definitions [98,217,225,256,263]. Arguably the simplest method, discussed by Merriman [167] in 1884 is to use the data to determine some well defined line, for example one that minimises the squared residuals. A trend might then said to be 'detected' if the slope is non-zero. For any set of observations, however, the slope will be either positive or negative with probability one, leaving the decision relevant significance of the analysis not assessed.

At the other extreme, statisticians have constructed general frameworks within

which significance is easily determined. Assume, for example [40], that each observation $y(t)$ is a realisation of a random variable $Y(t)$ with mean $\mu(t)$ and 'innovations' arising from Additive Observational Noise (typically from an Independent and Identically Distributed (IID) Gaussian process, $N(0, \sigma^2)$). For example, a linear model takes the form $Y(t) = mt + c + \epsilon$ where the innovations $\epsilon \sim N(0, \sigma^2)$. When the Gauss Markov assumptions (described in section C.1) are satisfied, and provided the innovations are Gaussian, it is possible to carry out a t-test (described in section C.2) which then allows for the significance of the slope to be calculated. Within this framework, a trend is said to be detected if the slope significance exceeds an arbitrary, pre-chosen, threshold. This and other general frameworks assume a great deal of machinery and the assessed significance is relevant only conditional on the assumptions. In many situations relevant to insurance it is clear that these are not satisfied and it is then inappropriate to use the t-test if the level of assessed significance is meaningless. Other approaches exist such as the non-parametric approach of Theil-Sen [225] or those listed by Gray [98] such as using Kendall's τ [256] or Likelihood ratios [217] and differences in detail to the above descriptions may also arise, these frameworks are not explored further.

This appendix returns to the basic question of whether or not a significant 'trend' appears in a time series, quantifying the probability of the observed result if, in fact the data were drawn from an IID process. This is arguably the original intention of trend detection. The method, described in section C.3, determines the probability that a random resampling of the observations could lead to a time series with parameters of a fitted trend defining function rarer than those observed. Trend defining functions include for example monotonic increasing polynomials and exponential functions[1]. The method is illustrated in this appendix by considering the least squares regression line where the probability of the observed slope is estimated and defined as the 'slope probability'. A time series is then said to have a significant trend if the slope probability exceeds a stated significance threshold. It is possible to explore the sub-structure of components of the time series using this method as the data can be subdivided into small subsets to test whether any significant trends

[1]For example, in the case of polynomials the 'parameter' of interest would be the coefficient of the highest order variable; in the case of exponentials it would be the value β in $\hat{y}(t) = e^{\beta t}$.

exist at smaller scales. Three new visual methods are introduced in section C.3 which display the slope probabilities for different subdivisions of the data.

Smith [236] defines **'Surrogate Data'** as: Non-deterministic time series constructed to be similar in appearance to the original data. He then comments that 'A common objection to the dynamical systems analysis of data from poorly understood systems is that the significance of a given result is rarely established.... This objection can be addressed directly by considering a class of non-deterministic surrogate signals. The significance of a result is then established by comparing it with the outcome of the same test applied to these surrogate data sets.' The permutation approach explored here is based on a similar idea, that the significance of trend can be assessed by considering many surrogate time series, known by construction to be IID, and determining their slope. Hotelling [107] proposed a permutation method in 1936 for exploring the degree of rank correlation between two variables without the need to make assumptions of normality and Efron [72] considered bootstrap estimates of regression parameters in 1977. The idea presented here, therefore, is not new, indeed it is implicit in the original idea of what a significant trend is; this appendix is then a return to basics augmented by new graphical presentations and arguments for the importance of a return to relevant trend detection (dropping the t-test along with other needlessly assumption laden tests commonly found in the literature).

The graphical methods provide complementary views and are used for various illustrative and actual data sets in the rest of the appendix to illustrate features of the time series. Section C.4 illustrates the slope probability method in a setting in which the t-test is appropriate (i.e. data arises from a linear trend with homoskedastic, non-correlated Gaussian noise). Section C.5 then explores a situation where the data generating process follows a cosine wave, in this case it is inappropriate to apply the t-test to a linear fit. Section C.6 considers another non-linear process whose mean values follow a $t^{1.5}$ relationship. This type of 'acceleration' leads to a 'wedge' feature in the graphical plots. Similar features can be expected to arise in any such convex situation. Sections C.7.1, C.7.2 and C.7.3: (1) consider three real atmospheric hazard datasets highlighting some interesting features in the data and (2) criticise a common misapplication to trend detection illustrated by Neumayer

and Barthel [186], but evident in many other sources[2] [10, 45, 172, 192, 275]. Section C.8 illustrates the method for the x time series of the Lorenz 63 model (to be briefly introduced in this appendix). This shows that a rich structure of slope significance can be identified by the new graphics. Section C.9 then illustrates the new method for tide gauge data in New York which shows a wedge feature as described in the $t^{1.5}$ example earlier, this result prompted a deeper investigation for a collection of tide gauges around the world producing evidence for accelerating sea level rise. Finally section C.10 considers sunspot data and the new graphics clearly reveal the long term behaviour and the shorter term 11 year sunspot cycle.

In summary, this appendix returns to the original intuitive idea [89] of a significant trend; it exploits modern compute power to evaluate the chance that any particular 'trend statistic' of interest is significantly larger than would be expected given independent draws from an identically distributed distribution. The original elements of this appendix are thought to be as follows:

- The concept of 'slope probability';
- Three novel graphical methods of the slope probability to reveal the nature of any trends in the data;
- Illustration of the new graphics:
    1. a perfect Gauss-Markov setting with a linear trend and Gaussian innovations, and;
    2. three synthetic examples where the usual t-test assumptions are failed;
    3. three real data sources: insurance losses arising from atmospheric hazards, tide gauge data and sunspot data;
- Analysis of global tide gauge data.

C.1 Gauss-Markov assumptions

This section defines notation and gives background information; nothing in this section is new (see for example Larson [131] or Ramanthan [207]). With linear

[2]Some of these sources use other trend calculation methods but still show no evidence of testing for normality or other model assumptions, it is possible such testing was carried out but not reported.

models it is assumed that a response variable (Y) is related to explanatory variables $X_1, ...X_n$ via a linear relationship in the presence of some Additive Observational Noise ϵ_i. Such a model is written over multiple observations $i = 1, ...p$, where $x_{i,j}$ denotes the i^{th} observation of the j^{th} explanatory variable:

$$y_i = \alpha + \beta_1 x_{i,1} + ...\beta_n x_{i,n} + \epsilon_i. \tag{C.1}$$

This can be written in matrix notation by adjoining a constant explanatory variable $X_0 = 1$ and denoting $\beta_0 = \alpha$. In equation C.2, Y is now a column vector of observations, X is a matrix, $\beta = (\beta_0, ...\beta_n)$ a vector of parameters and ϵ a vector of Additive Observational Noise:

$$Y = X\beta + \epsilon. \tag{C.2}$$

Different choices of β lead to different linear models. For a given choice, the $\hat{Y}(\beta) = X\beta$ are a vector of 'fitted' values. The residuals are the values $\epsilon = \hat{Y} - Y$. The sum of squared residuals $Q = \sum \epsilon_i^2 = \sum((\hat{Y} - Y))^T(\hat{Y} - Y)$ quantifies how far the observed data is from the mean values assuming the Euclidean norm. This uses the symbol A^T to denote the transpose of the matrix A. The least squares approach seeks to find the parameter $\hat{\beta}$ that will minimise Q and this is known as least squares regression. Defining $\hat{\beta}$ as follows minimises Q [68]:

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{C.3}$$

Note that the probability distribution of the Additive Observational Noise does not need to be specified in order to calculate $\hat{\beta}$, it is simply the model that minimises the quantity Q for the observed sample $y_1, ...y_p$ and $x_{1,j}, ...x_{p,j}$ (for each explanatory variable j).

Note also that the ordinary least squares estimator $\hat{\beta}$ is of the form $\hat{\beta} = \sum a_i y_i$, a linear sum of the response variables. Such an estimator is known as a 'linear estimator'. An estimator is 'unbiased' if $E(\hat{\beta}) = \beta$. An estimator with a high variance, even if unbiassed, is quite likely to produce a value far from the true parameter. For this reason given two unbiassed estimators the one with the lower variance is preferred. An obvious question is whether there is a linear estimator that has a lower variance than that produced by the ordinary least squares method.

There are conditions under which the answer to this question is 'no'. These are called the Gauss-Markov assumptions [207] which require the following:

1. The response is linear in parameters to the explanatory variables (for example $Y = \alpha + \alpha\beta^2 X$ would fail this requirement);
2. Errors average to zero, $E(\epsilon_i|X) = 0$, $\forall i$;
3. The explanatory variables (X_t) are given and non-random (this implies $\text{cov}(X_t, \epsilon_t) = 0$, $\forall t$);
4. Homoscedasticity, that is $\text{var}(\epsilon_i|X) = \sigma^2$, $\forall i$;
5. Serial Independence, that is $\text{cov}(\epsilon_i, \epsilon_j) = 0$ $\forall i, j$;

With these assumptions the Gauss-Markov theorem establishes that the ordinary least squares estimator has the lowest variance amongst all unbiassed linear estimators and is thus described as BLUE - the Best Linear Unbiassed Estimator [207].

The discussion above relates to any n-dimensional linear model. From now on the discussion is restricted to time series, and the much simpler case of two dimensions where $Y = \alpha + \beta X + \epsilon$. For a given time series the derived slope parameter $(\hat{\beta})$ will be nonzero with probability one. A question that arises naturally is whether the slope $\hat{\beta}$ is statistically different from zero.

C.2 The t-test

This section presents the t-test for slope significance which is typically used to provide evidence for trends; nothing in this section is new and the development follows standard presentations [121, 131, 207].

Suppose that x and y are related by the following equation:

$$Y = \alpha + \beta X + \epsilon \tag{C.4}$$

where $\epsilon \sim N(0, \sigma^2)$ is an Additive Observational Noise term.

Note that such a model would satisfy the Gauss-Markov assumptions; but that an additional strong assumption on the statistical form of the Additive Observational

Noise term has been made. The following describes a standard test which determines whether the sample slope estimator $\hat{\beta}$ is actually statistically non-zero. Let

$$s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}, \tag{C.5}$$

denote the sample standard deviation of x with the s_y defined analogously for y. Define the sample correlation coefficient r to be:

$$r^2 = \frac{\sum((x_i - \bar{x})(y_i - \bar{y}))^2}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2} \tag{C.6}$$

and note that

$$r = \hat{\beta}\frac{s_x}{s_y}. \tag{C.7}$$

Let

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}. \tag{C.8}$$

Let

$$\tau = \frac{r}{s_r} \tag{C.9}$$

Then $\tau \sim T(n-2)$, has a t-distribution with $n-2$ degrees of freedom [131]. Let

$$s_{\hat{\beta}} = \frac{s_y}{s_x}\sqrt{\frac{1 - r^2}{n - 2}}. \tag{C.10}$$

Substituting equations C.7 and C.10 into C.9 an alternative formula for τ can be constructed as follows:

$$\tau = \frac{\hat{\beta}}{s_{\hat{\beta}}}. \tag{C.11}$$

The above equations all relate to sample statistics. Analagous to equation C.7 it is also true for populations (ρ is population correlation, β is slope of relationships between bivariate normals and σ_x is the population standard deviation of X).

$$\rho = \beta\frac{\sigma_x}{\sigma_y}. \tag{C.12}$$

Unless X is constant, it is clear that $P(\rho \neq 0 | \beta \neq 0) = 1$. Putting this together:

- Equation C.12 shows that a test for non-zero correlation is equivalent to a test for non-zero slope;
- Equation C.9 gives a statistical test for correlation where the derived statistic τ will have a t-distribution if certain conditions are met;

- Equation C.11 gives an alternative formula for τ with respect to the sample slope $\hat{\beta}$, this also allows error bars for the slope to be created.

The t-test is not meaningful if the Gauss-Markov assumptions are violated or if the residuals are not normal. The following describes some standard statistical tests for these features:

Test for normality There are many tests for normality. Razakit and Wah [187] shows that the Shapiro-Wilk test [227] has the most power in a series of normalised monte-carlo tests. This test compares the order statistics from a normal distribution to the order statistics that are in the data - it creates a test statistic that can be looked up on test tables to determine the p-value.

Test for homoscedasticity The Breusch-Pagan test [28] tests whether the data is heteroscedastic by fitting a linear model to the squared residuals as a function of the explanatory variables and testing whether the parameters are significant.

Test for serial correlation The Breusch-Godfrey [96] test can be used to test whether the residuals exhibit serial correlation. In this test the residuals are tested to see if they have an $AR(p)$ $p \geq 1$ process using standard methods. If an $AR(p)$ process is a good model for the residuals then the assumption of zero serial correlation can be rejected.

This section has listed conditions which must be satisfied in order for the t-test to be validly applied. As noted already the p-values are sometimes calculated when this necessary condition is not satisfied which motivates an alternative approach which might prove more deployable at the cost of being less informative. Such an approach is suggested in the next section.

C.3 Graphical Methods of Trend Detection

The following subsections illustrate a definition of 'trend' which neither requires the Gauss-Markov assumptions, nor makes any assumptions about the Additive Observational Noise (see Glossary), (or indeed whether the data conform to a particular

linear model). To recap, given a time series $X = \{x_1, ...x_N\}$ (see Glossary) made at times $T = \{t_1, ...t_N\}$ ($\{X, T\}$) a least squares line can be fitted through the data. The slope of the line will almost always be positive or negative. The question is whether the size of the slope is a surprise, indicating a trend in some sense.

Definition: Implied Linear Trend Given a time series $\{X, T\}$ define the **'Implied Linear Trend'** to be the slope $\text{ols}(X, T)$ of the ordinary least squares line through the data, as follows:

$$\text{ols}(X, T) = \frac{\sum_{i=1}^{N}(x_i - E_X)(t_i - E_T)}{\sum_{i=1}^{N}(t_i - E_T)^2} \tag{C.13}$$

where $E_X = \frac{\sum_{i=1}^{N} x_i}{N}$ is the average of the values of X and E_T is similarly defined for T. The approach to trend detection below is easily broadened to a 'trend' defined by any monotonic function (e.g. 'implied exponential trend').

For example, let series $A = \{X_A, T_A\}$ be such that $t_i = x_i = i$ for $i \in \{1, ...N\}$. This has an implied linear trend of 1. Let series B be such that the last two values are swapped, so the series reads $1, 2, .., N, N-1$ then the slope through this data will be lower (for example if $N = 20$ the slope would be 0.99850). In fact, series A has the largest positive slope among all permutations of the values in X_A. There are $N!$ permutations of the first N integers (for $N = 20$ this is approximately $2.4329 * 10^{18}$). Since all other time series in this example have slopes that are less than A it would be highly significant if A was observed.

For a given time series $\{X, T\}$ the probability of observing a slope less than or equal to the Implied Linear Trend $\text{ols}(X, T)$ can be derived. This suggests the following definition:

Definition of Slope Probability Given a time series $\{X, T\}$ of length N, let Φ denote the set of all permutations[3] of the set $\{1, ...N\}$. Let $I(A)$ be an indicator function which is zero when A is false and 1 otherwise and let $|S|$ denote the cardinality of the set S. For $\phi \in \Phi$ let the **'permuted series'** be denoted $X^\phi = \{x_{\phi(1)}, ...x_{\phi(N)}\}$, where T is not permuted. Let $m = \text{ols}(X, T)$ and let $m_\phi = \text{ols}(X^\phi, T)$. Define the

[3]For example the 6 permutations of the numbers $\{1, 2, 3\}$ are: $\phi_1 = \{1, 2, 3\}$, $\phi_2 = \{1, 3, 2\}$, $\phi_3 = \{2, 1, 3\}$, $\phi_4 = \{2, 3, 1\}$, $\phi_5 = \{3, 1, 2\}$ and $\phi_6 = \{3, 2, 1\}$.

'slope probability' p_s as:

$$p_s(X,T) = \frac{\sum_{\phi\in\Phi} I(m_\phi < m)}{|\Phi|} \tag{C.14}$$

The summation in definition C.14 is finite and therefore in theory the slope probability can be calculated exactly. In practice, however, for time series with $N > 15$ this requires the calculation of more than a trillion slopes and time constraints becomes prohibitive. To address this the following Monte Carlo method is adopted:

Estimate the Slope Probability of the Implied Linear Trend

- STEP 1: Calculate the the implied linear trend $m = \text{ols}(X,T)$;
- STEP 2: Sample a permutation $\phi_i \in \Phi$ where each is equally likely and create the permuted series $\{X^i, T\}$ (equivalent to sampling N values from X without replacement and noting that T is unchanged);
- STEP 3: Calculate the implied linear trend $m_i = \text{ols}(X^i, T)$ from this new sample;
- STEP 4: Repeat steps 2 and 3 to generate N_{smpl} realisations;
- STEP 5: Define the estimated slope probability as $\hat{p_s} = \frac{\sum_{i=1}^{N_{smpl}} I(m_i<m)}{N_{smpl}}$.

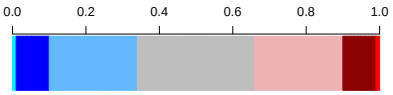


Figure C.1: Colour key used for most plots. Y axis shows probability of slope occurring from random resampling of points

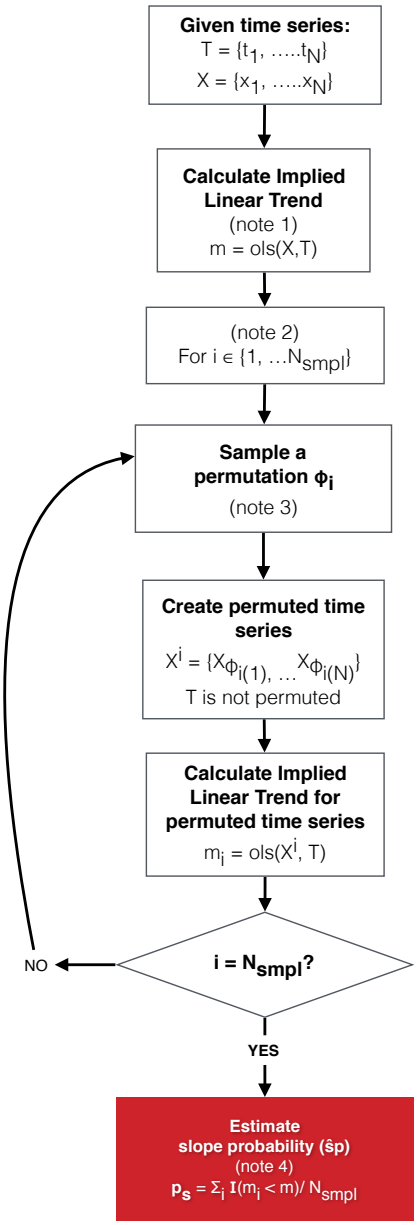
Integer segmentation If a trend is caused by an exogenous driver (for example increasing CO_2 concentrations leading to increasing atmospheric temperatures) then the trend it is likely to be present if the data is subdivided into shorter intervals, although sampling error is generally expected to feature more strongly the smaller the subdivision. Evidence there is such a driver may therefore be found by considering multiple subdivisions of the data, which can be carried out using the following method. Where $[x]$ is used to denote the integer part of the real number x:

Algorithm for Integer Segmentation Plots (see also figure C.3)

- STEP 0: Calculate the slope probability of the full time series (as above)
- STEP 1: Divide the data into two equal sized blocks at the beginning and end of the data set, discard any residual data in the middle. Specifically, keep the ordering from the original time series to produce a subdivision $1 = x_1, ...x_{[\frac{N}{2}]}$ and subdivision $2 = x_{N-[\frac{N}{2}]}, ...x_N$. Keep the remaining data point, if any, separate. Recalculate the implied linear trend in each portion of the data and calculate the slope probability for each. Choose a colour scale (figure C.1) for the slope probability and colour the block relating to the data accordingly.
- STEP i: Divide the data into i equal sized blocks, set aside any residual data. Calculate the slope probability of each sub-series. For the residual data, if there are less than 4 terms colour the block black, otherwise calculate the slope probability of the residual data and its corresponding colour.
- Repeat until $i = N_{max}$.

Figure C.1 shows the colour key used in this appendix. Brilliant red (and cyan) is reserved for slopes that have a slope probability of greater than 99% (or less than 1%) respectively. Dark red (and dark blue) shows slopes that have slope probability between 90% and 99% (or between 1% and 10%) respectively. The number of permutations is always even and any permutation leading to positive slope has a mirror image with negative slope. Since the colour scale chosen is also in symmetric bands either side of the 0.5 probability we can say that red colours correspond to positive slopes and blue to negative.

Running Windows The choice to consider the first half and last half of the data (or any other subdivision) in the integer segmentation plot is arbitrary. It is possible to take blocks of a chosen size from anywhere within the data set. The following suggests this as a complementary graphical method. A **'window'** of size M is defined as a series $x_j, ...x_{j+M-1}$ of data points from the original time series occurring at times $t_j, ...t_{j+M-1}$.



Notes:

1.Define the Implied Linear Trend as: **ols(X,T) = $\Sigma_i(x_i - E_X)(t_i - E_T) / \Sigma(t_i - E_T)^2$** where E_X and E_T are the average of the values in sets X and T respectively.

2.**N_{smpl}** is the number of permutations that will be sampled without replacement to estimate the slope probability.

3.There are N! permutations of the numbers 1,...N. Sample one permutation where each is equally likely.

4. **I**(A) is an indicator function which takes the value 1 when A is true and 0 otherwise

Figure C.2: Flow chart for the calculation of the slope probability.

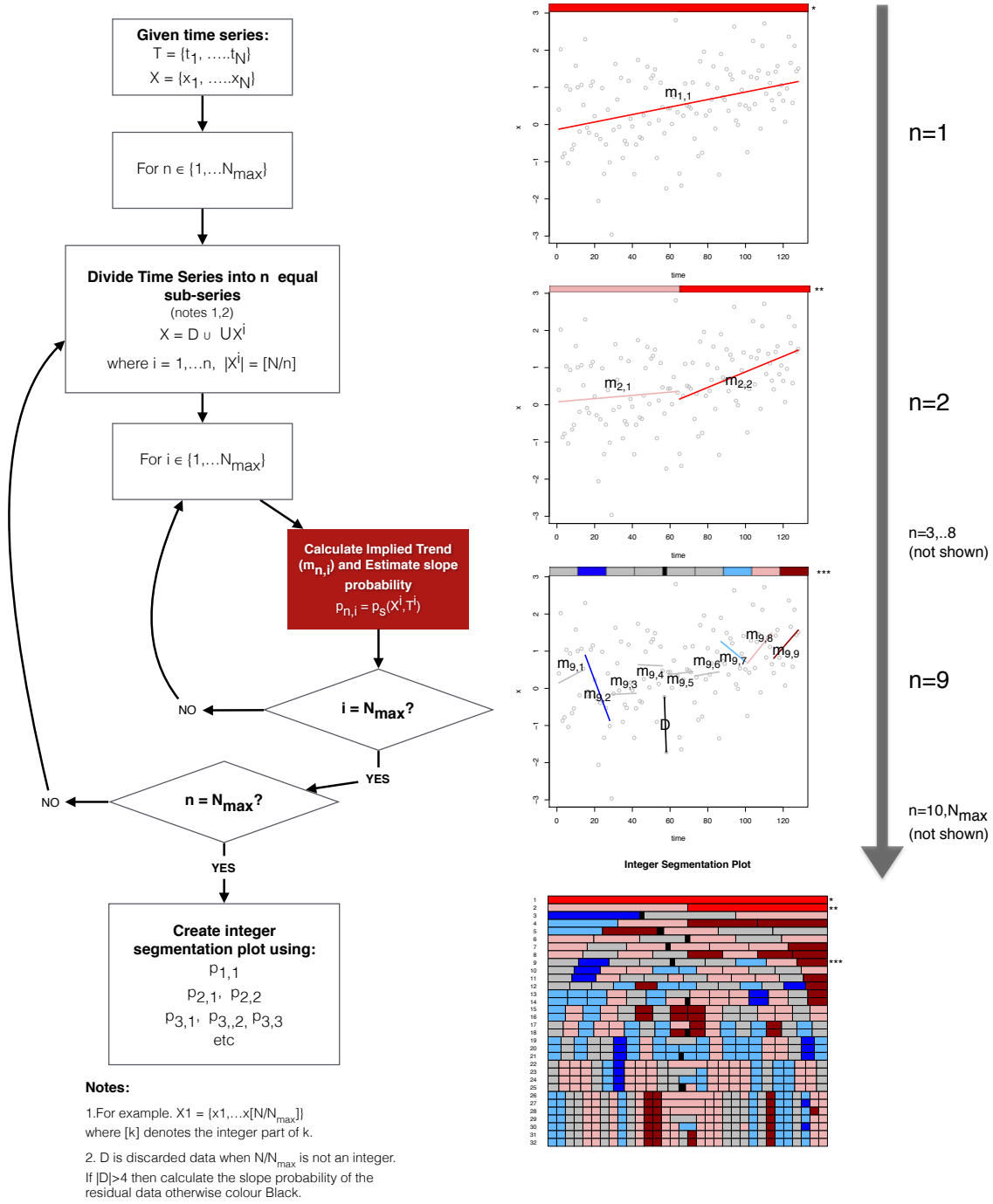


Figure C.3: Flow chart for Integer Segmentation Plot. Right hand column illustrates the steps for n=1,2, and 9 and also shows the resulting plot in the case of a linear trend. Data set is C7.1.3 for illustration.

Algorithm for Running Window plots (illustrated in figure C.4)

- For a given time series $\{X, T\}$ and chosen window sizes $\{M_i\}_{i=0}^{R_{wnd}}$, such that $M_0 = N$;
- For $i \in \{0, ...R_{wnd}\}$
  - For chosen window positions $j \in \{j_{i,1}, ...j_{i,W}\}$
  - Define the sub-series $X^{i,j} = \{x_j, ...x_{j+M_i-1}\}$ and $T^{i,j} = \{t_j, ...t_{j+M_i-1}\}$
  - Calculate the Implied Linear Trend $m_{i,j} = \text{ols}(X^{i,j}, T^{i,j})$ and estimate the slope probability $\hat{p}_{i,j}$.

Two plots have been created using the running window method. The first will be referred to as a **'triangle plot'** and the second as a **'stacked square'** plot. These are described in more detail as follows.

Running window - triangle If the full time series has N data points then the triangle plot displays the slope probability in the case where $M_i = N - i$. There are $i+1$ positions of M_i in this case as the window runs over the full data in increments of one data points at a time. The resulting slope probabilities are colour coded as before and plotted in a triangle of cells (illustrated in figure C.4). The result for the full data is shown as a single cell at the top of the plot at the apex of the triangle, the next two results (where the window is of length $N - 1$ and either starts at the first data point or the second) plotted as the two cells below the top one; and so on 3,4,5 etc. Note, therefore, that when the window is of length $N - i$ this can occupy $i + 1$ positions - and the total number of slope probabilities is $\frac{N(N+1)}{2}$ the sum of the first N integers. As such, the number of calculations is $o(N^2)$ which can take a long time to run. Due to the run-time constraints the graphic has been limited to $N \leq 50$ in the longer data sets so only part of the data is illustrated which is a shortcoming of this method. Another (possible) behavioural bias with this graphic is that arguably the most important row is the first one where the slope probability of the full data set is shown - yet this has just one cell. Smaller window sizes are afforded more ink and the overall colour could be misinterpreted.

Running window - stacked square The stacked square plot addresses two of the shortcomings of the triangle plot (1) The behavioural colour-bias towards less

significant subsets of the data and (2) the runtime issues when data sets are large. For each window size M_i there are $N - M_i + 1$ potential positions for the window - this is the source of the runtime issues mentioned above. Let W equal the maximum number of windows that will be considered for each window size. The first and last windows are fixed to cover the first and last M_i data points; the rest are chosen to jump through the data in approximately equal increments. Note if $N - M_i + 1 < W$, the windows are incremented in steps of 1 data point at a time. In this case the height of the strip is adjusted to fit a fixed block height. In summary, this plot (also illustrated in figure C.4) displays the selected strips for each window size choice within a block of a fixed height, the height is the same in each case. This reduces the behavioural colour-bias in the triangle plot because each window group is afforded the same about of plot space. The full data (i.e. where no data is removed) is therefore a single strip which takes up one complete block. Run times are reduced by restricting the size of K and D.

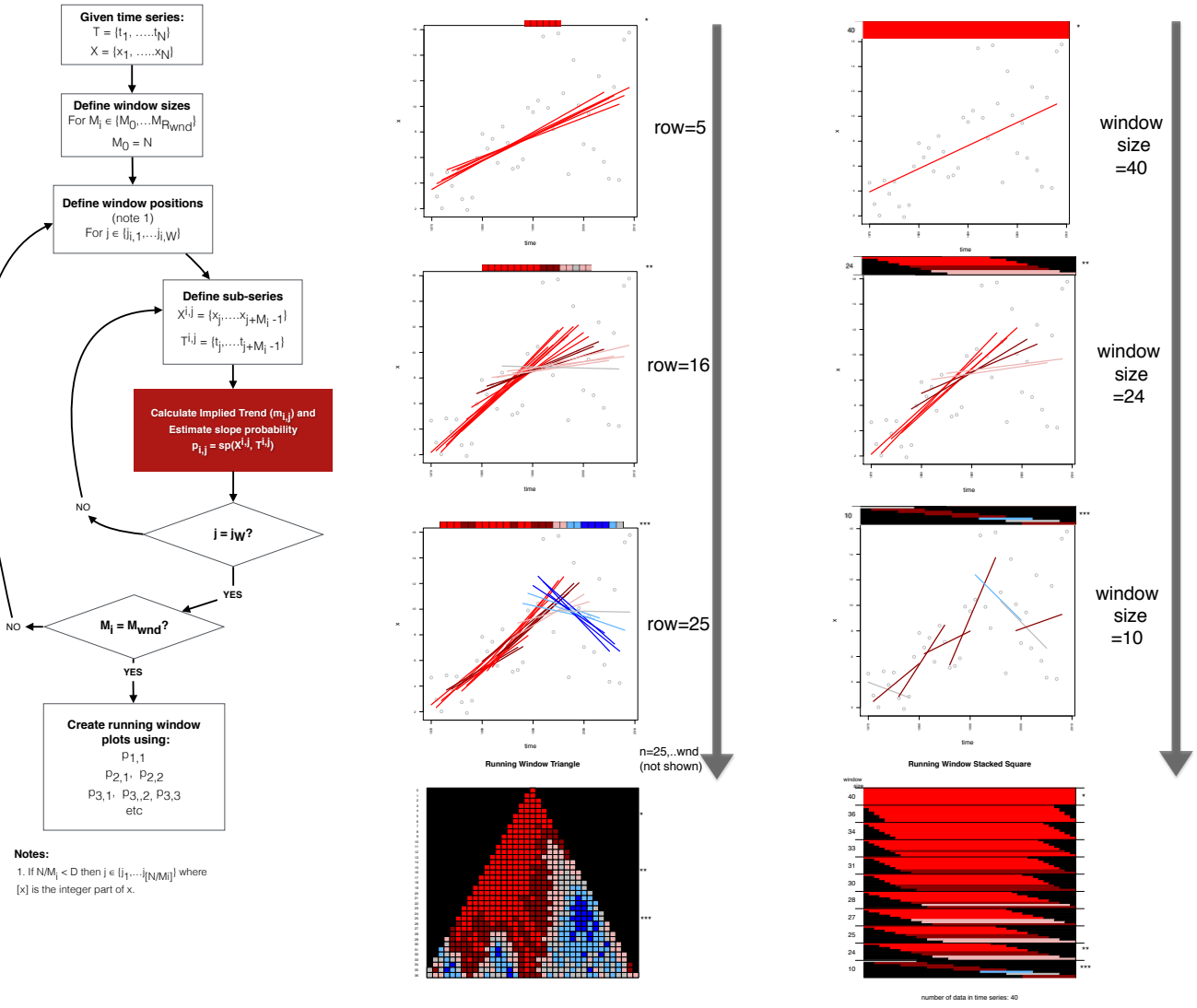


Figure C.4: Flow chart of Running Window plots: Graphics in middle column illustrates the Running Window Triangle and the right hand column illustrates the Stacked Square method in the case of a linear trend. Data set is C7.5.1 (Convective events in the USA) for illustration.

C.4 Perfect Guass-Markov examples

In order to illustrate the concept of Slope Probability described above the following well behaved examples have been produced. Examples C7.1.x and C7.2.x described below, all satisfy the Gauss-Markov assumptions. The mean of the Additive Observational Noise term is chosen so that in examples C7.1.x and examples C7.2.x all have the same mean in expectation. These are illustrated in a series of figures as follows. Most figures in this appendix retain this format: Top left is the time series and statistical test results, top right shows the Integer Segmentation Plot, Bottom left the running window stacked square and bottom right the triangle.

Example C7.1.x: Gaussian Additive Observational Noise about a line with slope $\frac{1}{N}$

Time series $y_t = \frac{t}{N} + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma)$ for $t = 1, ...N$

- Case C7.1.1: $\sigma = 0.01$, $N = 2^7$
- Case C7.1.2: $\sigma = 0.1$, $N = 2^7$
- Case C7.1.3: $\sigma = 1.0$, $N = 2^7$

7.2.x Gaussian Additive Observational Noise about a line with zero slope

---

Time series $y_t = \epsilon_t$, $\epsilon_t \sim N(0.5, \sigma)$, for $t = 1, ...N$

- Case 7.2.1: $\sigma = 0.01$, $N = 2^7$
- Case 7.2.2: $\sigma = 0.1$, $N = 2^7$
- Case 7.2.3: $\sigma = 1.0$, $N = 2^7$

Figure C.5 shows the results for one realisation of example C7.1.1. The time series plot includes (in red text) the results of the various standard tests. The Shapiro-Wilk test is carried out against a p-value of 5% and also 1% and the normality of the residuals is not rejected. The Breusch-Godfrey test is carried out on the residuals and as expected does not reject the assumption of zero serial correlation.

The Breusch-Pagan test does not reject homoscedasticity, again as expected. The t-test is carried out and, unsurprisingly the slope parameter is found to be significant so zero slope is rejected. The data has been shown to have a trend, using the the standard statistical definition of this term.

The Integer Segmentation Plot within C.5 shows that the least squares slope in the data occurs less frequently than 1% of the time from random resamples of the data. The data can therefore be said to contain a trend according to the definition proposed in this appendix. In general, in this appendix, a slope will be described as **'significantly positive'** if it has a slope probability of less than 10%[4] (or **'significantly negative'** if greater than 90%), this significance threshold is arbitrary. This first and last half of the data set also includes such a significant slope and this is the case for each subdivision until the data is divided into 13^{ths} when one of the segments has between a 10% and 1% chance of occurring from a random sample. As the data continues to be divided the slopes lose their significance although it is notable that the entire plot is some shade of red. Duplicate rows can arise due to the treatment of excess data after subdivision. For example there are 128 data points in the current plot. Divide by 15 and you get 15 blocks of 8 - with the remainder also 8 - so 16 blocks of 8. As an alterative example, division by 16 goes exactly into the 128 data points with no remainder and so 16 blocks of 8 arise. Therefore rows 15 and 16 in the plot (numbered on the left of the plot) are duplicated.

The apex of the Running Window Triangle plot within figure C.5 is, by construction, the same colour as the top row of the Integer Segementation Plot (i.e. the whole data set). The two blocks below the apex refer to the data set with the last point removed (first block) and first point removed (second block). The next three blocks on the third row show the result for a block of size $N-2$, the first block on this row includes all but the last two points, the second block excludes the first and last point and the third excludes the first two data points. And so on. Clearly the slope is significant for all windows even with the removal of 50 data points. The Running Window Stacked Square plot within figure C.5 shows a similar story where

[4]The phrase **'highly significant'** will be reserved for Slope Probabilities greater than or equal to 99% or less than or equal to 1%

the slope is seen to be significant at all illustrated window sizes.

Figure C.6 shows the result for example C7.1.2. The t-test again rejects zero slope. The top row of the Integer Segmentation Plot shows a significant slope and this is retained up to when the data is quartered. The running window triangle plot shows that the trend is present in all data windows down to $78 = 128 - 50$.

Figure C.7 shows the result for example C7.1.3. In this case the Additive Observational Noise term has a standard deviation of 1.0 leading to errors that are larger than the span of the trend itself. The t-test still finds the slope to be significant. The Integer Segmentation Plot shows that the full data has a significant slope - but the first half of the data does not have a significant slope. Thereafter no subdivision shows a significant slope. The Running Window Triangle shows that the slope is significant even when up to 22 data points are removed, for all windows; down to the removal of 37 data points the triangle graphic all slopes are either significant or highly significant. Below this, the early part of the time series is not significant.

Figure C.8 shows the results for example C7.2.1. The whole data set has a slope that is not significant as shown by the light blue bar at the top of the Integer Segmentation Plot. This plot shows some patches of significant slopes (for example the two bright red first segments of rows 4 and 5. These would be expected to arise randomly and it is notable that the segments adjacent to these are blue. Overall there is no strong pattern in the Integer Segmentation Plot consistent with the process generating the data. The Running Window Triangle plot has regions of significant slopes (dark blue) surrounded by larger regions of insignificant slope (light blue, grey and light red). The majority of the windows illustrated in the Running Window Stacked Square plot are not significant.

Figures for examples C7.2.2 and C7.2.3 are not shown. The features are similar to C7.2.1. The slope is not significant in either case using either the standard statistical definition or the new definition using the Slope Probability.

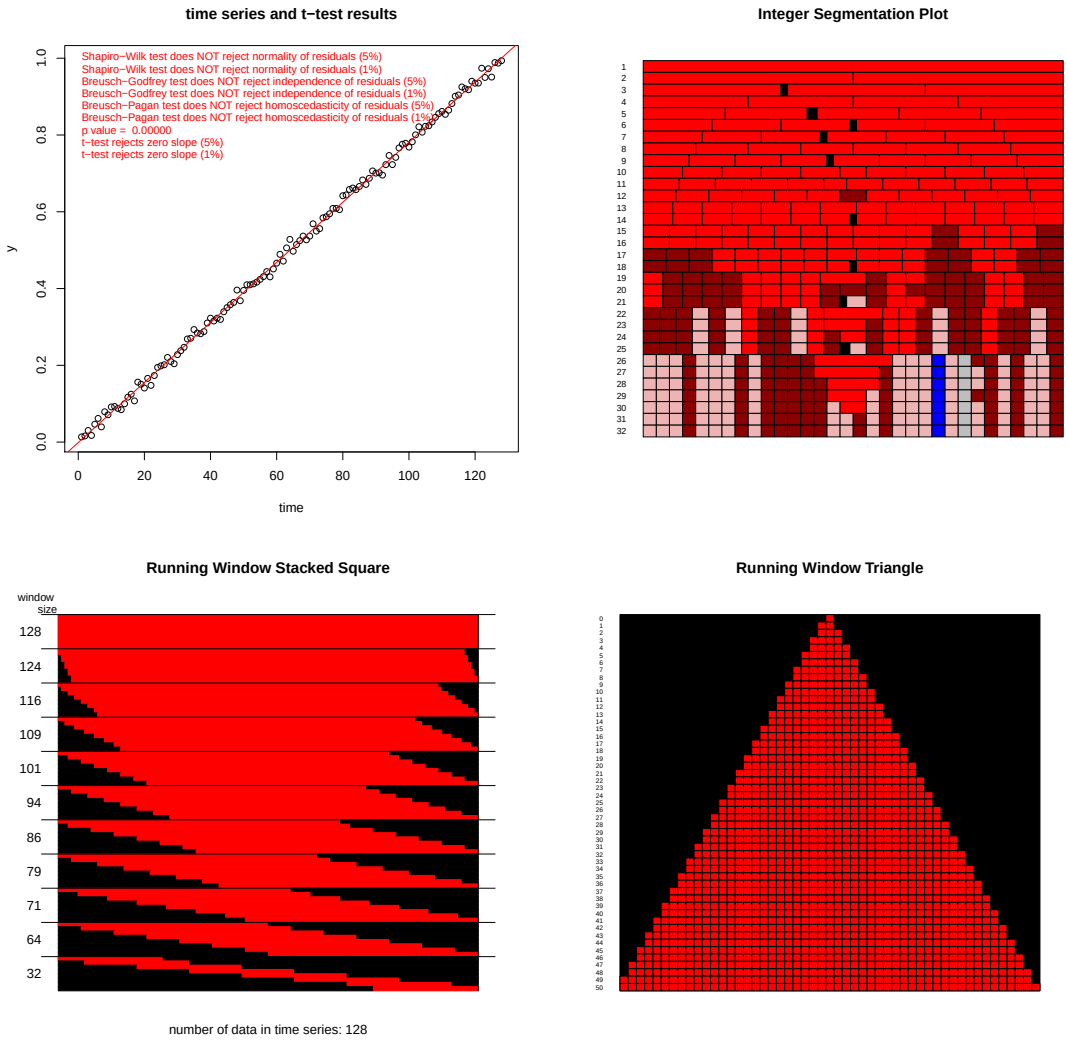


Figure C.5: Example C7.1.1: Perfect Gauss-Markov: $\sigma = 0.01$, slope $= \frac{1}{128}$. The Stacked Square and Triangle plots show highly significant slopes for all illustrated window sizes. The Integer Segmentation plot shows highly significant trends down to subdivision of the data into $\frac{1}{14}$ths.

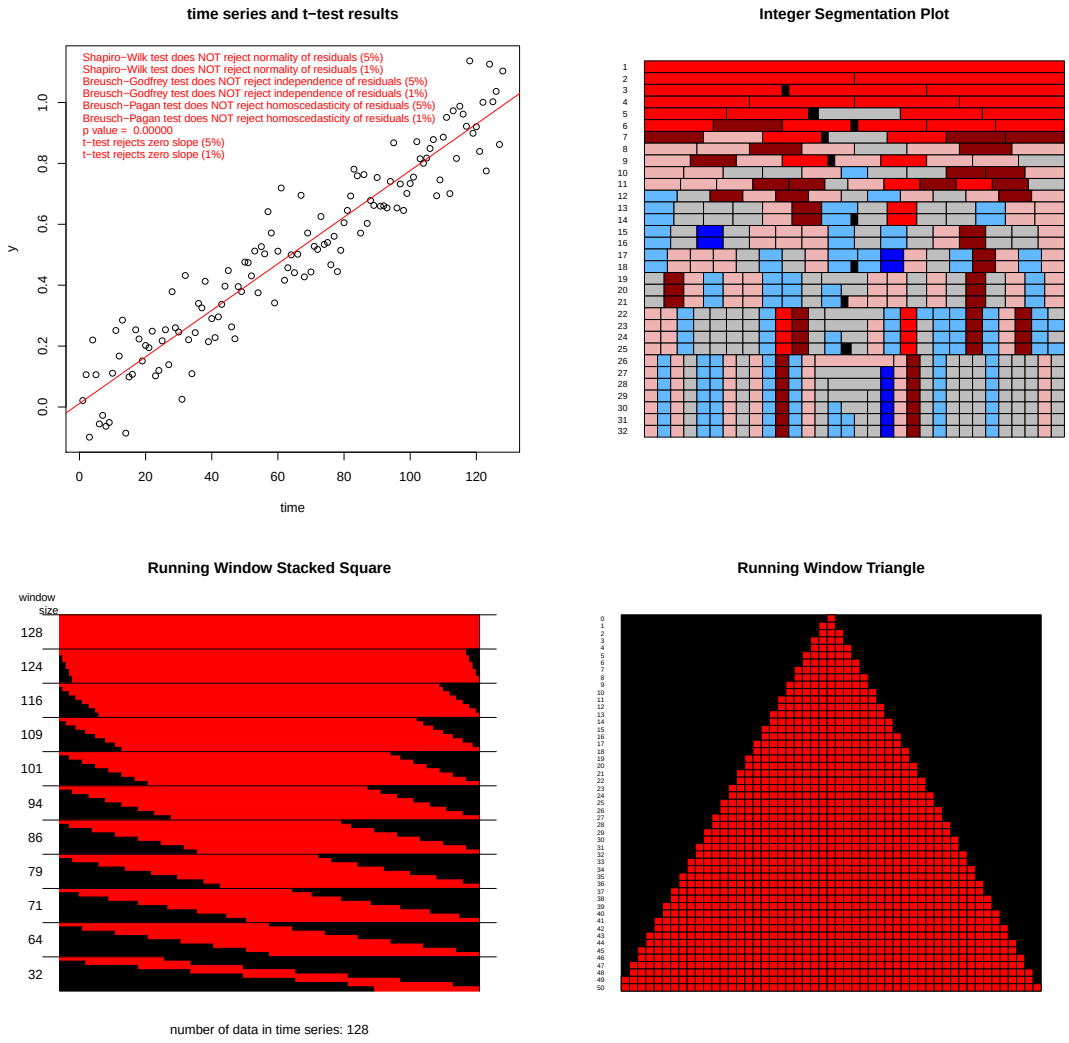


Figure C.6: Example C7.1.2: Perfect Gauss-Markov: $\sigma = 0.1$, slope $= \frac{1}{128}$. As with figure C.5 the Stacked Square and Triangle plots show highly significant slopes for all illustrated window sizes. The Integer Segmentation Plot shows highly significant slopes only up to when the data is divided into fifths (consistent with a higher variance ($\sigma = 0.1 > 0.01$) of the Gaussian Noise term), positive trends with mixed significance are evident until the data is subdivided by a factor of 9 after which there is no discernible pattern in the colours.

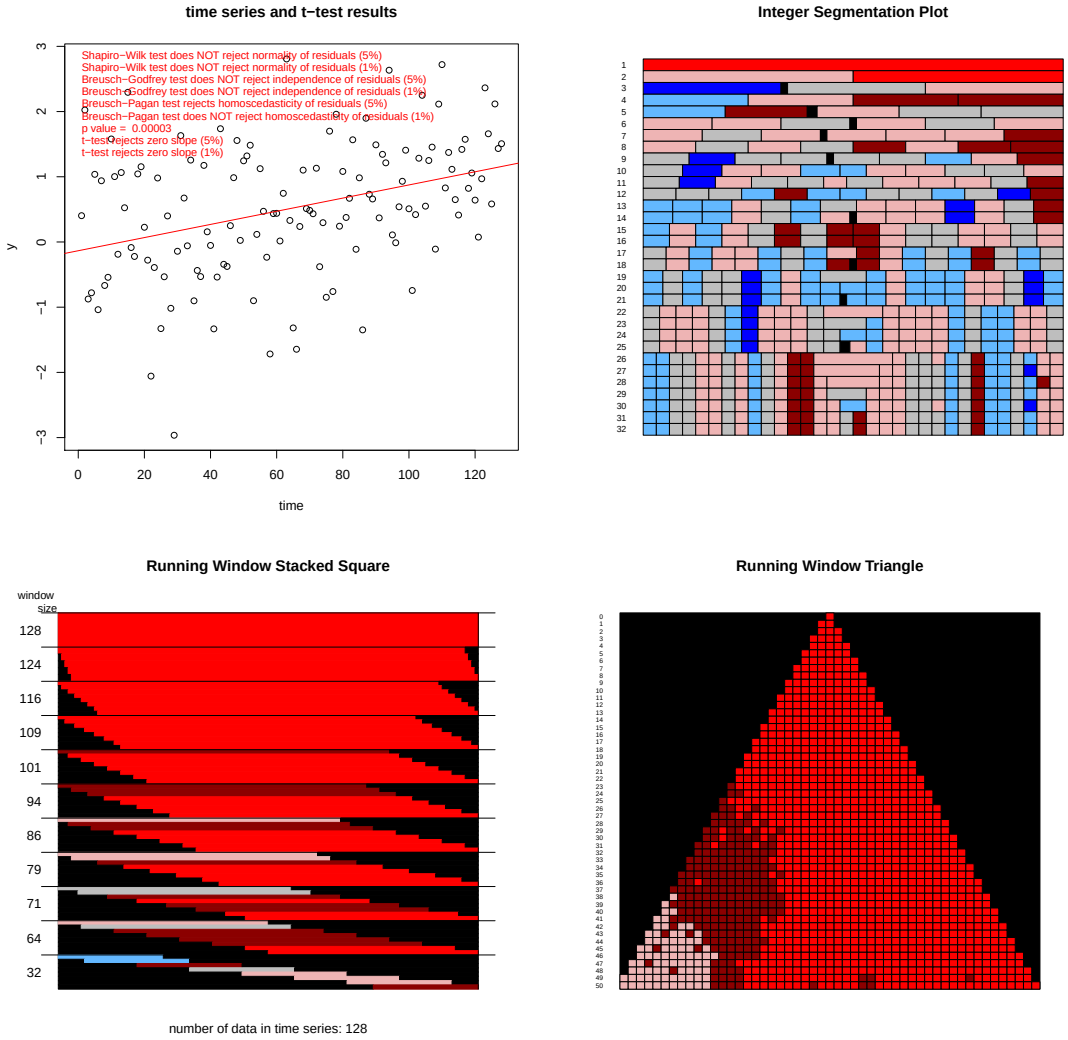


Figure C.7: Example C7.1.3: Perfect Gauss-Markov: $\sigma = 1.0$, slope $= \frac{1}{128}$. The variance term in time series underlying this series of graphics is much larger than in figure C.5 and C.6. Consistent with this, in the Integer Segmentation Plot the trend is only highly significant for the data set as a whole. The Triangle plot shows that this degree of high significance is retained for all windows with up to 22 points removed and the trend remains significant up to the removal of 37 points. The Integer Segmentation Plot shows that once the data is halved high significance is only evident in the second half of the data after which there is no discernible pattern in the colours.

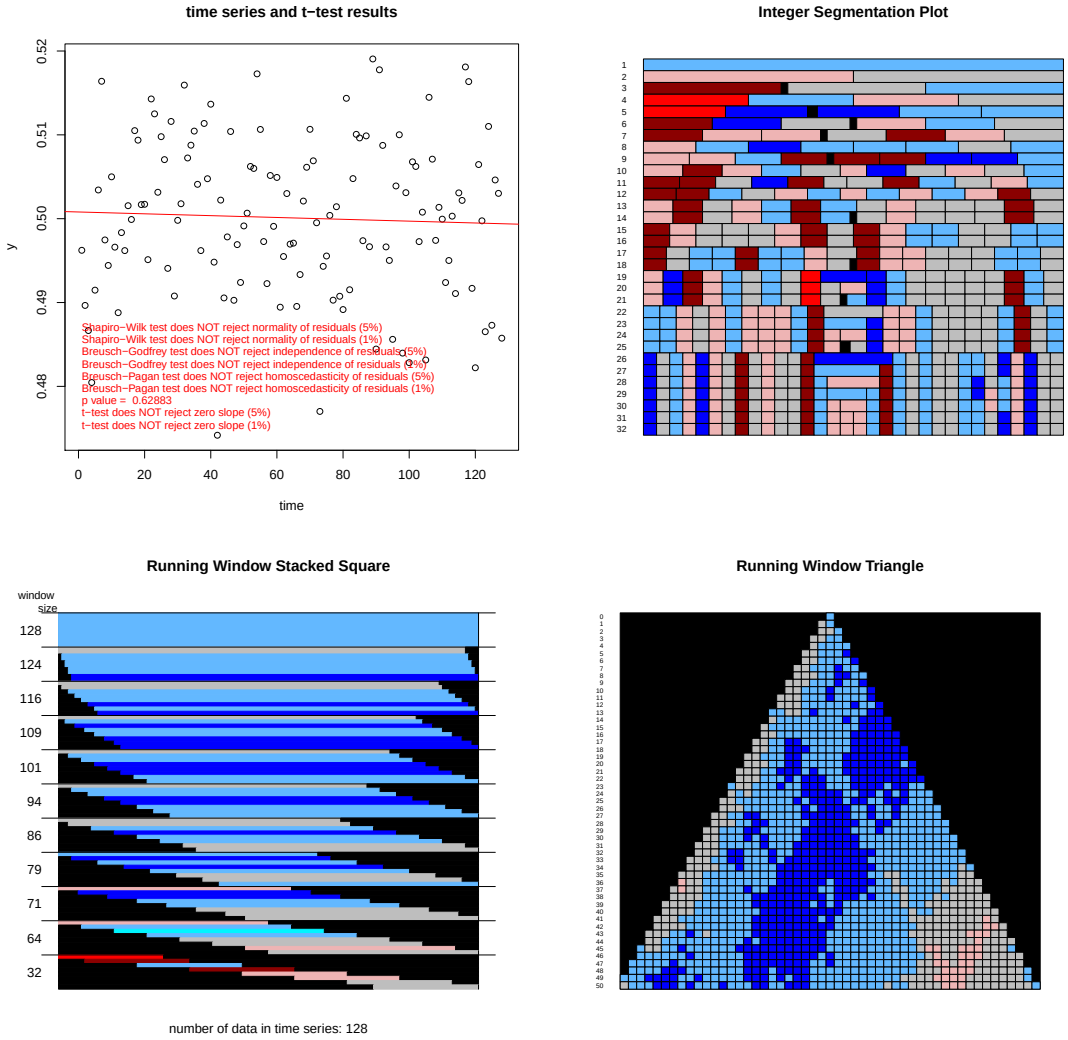


Figure C.8: Example C7.2.1: Perfect Gauss-Markov: $\sigma = 0.01$, slope $= 0$. The Integer Segmentation plot shows that whilst the slope of the ordinary least squares line is negative it is not significant. By construction the time series has no long term trend and any observed trend is an artefact of the sampled Gaussian Noise. The Triangle plot and Stacked Square plots show that that the sign of the slope (negative) is retained and even becomes significant for some smaller window sizes in some locations. The Integer Segmentation Plot, however, shows no discernible pattern in the colours.

C.5 Cosine with Gaussian Noise

Example C7.3 Cosine with Gaussian Additive Observational Noise	
Time series	$x_i = cos(t_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma)$, $t_i = -\pi(1 - \frac{i-1}{N-1})$
Parameters	$\sigma = 0.2$ and $N = 2^7$.

Figure C.9 (top plot) shows a sample with the cosine expectation shown in blue and a least squares regression line shown in red. The text in the plot shows the results of various standard statistical tests:

1. The Shapiro-Wilk tests for IID normality of the residuals and does not even reject this at the 5% level; this is despite the fact that the residuals (from the regression line) are not normal;

2. The Breusch-Godfrey test does reject the the independence of the residuals, however. This is sufficient to invalidate the use of the t-test as a method to determine the significance of any trend;

3. The Breush - Pagan test doesn't reject homoscedisticity because there is no linear relationship between the residuals and the time variables and that is what this test looks for. The bottom plot of figure C.9 shows the least squares regression line (red) for the residuals and this is clearly flat. The green line in that plot is a kernel smoother through the data which clearly follows the expected pattern in the residuals. The residuals clearly aren't independent by construction - hence the t-test hypotheses are not met.

Figure C.10 shows the results. The Integer Segmentation Plot shows that the Implied Linear Trend is significant for all segments up to division of the data by 7 and highly significant up to division by 4. The stacked square plot is completely red indicating high significance for all illustrated windows, as does the running window triangle. This illustrates a key utility of the new plots - they do not rely on the applicability of the t-test but can still clearly demonstrate that an Implied Linear Trend is present.

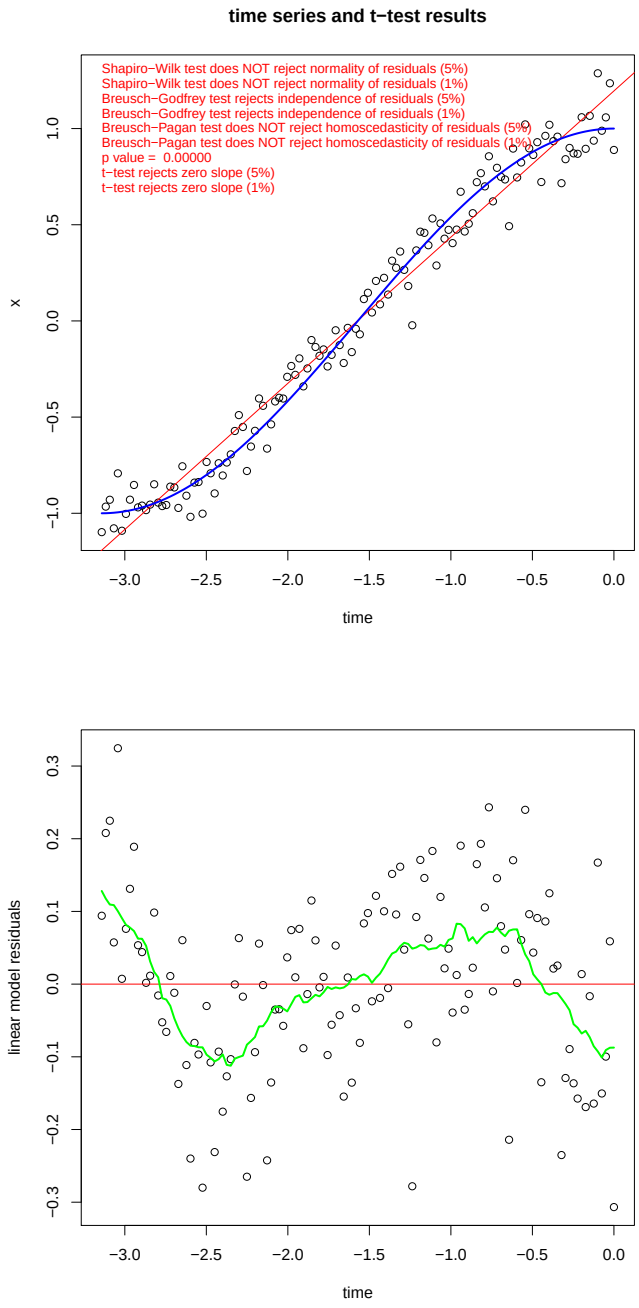


Figure C.9: Example C7.3, Cosine with Gaussian Additive Observational Noise : Sample from times series whose mean values vary with the cosine of time. Top plot shows a sample from the distribution, with the underlying mean values shown in blue and the fitted least-squares regression line in red. Lower plot shows the residuals from the linear model with a kernel smoother (green) through the data and also a linear trendline (red)

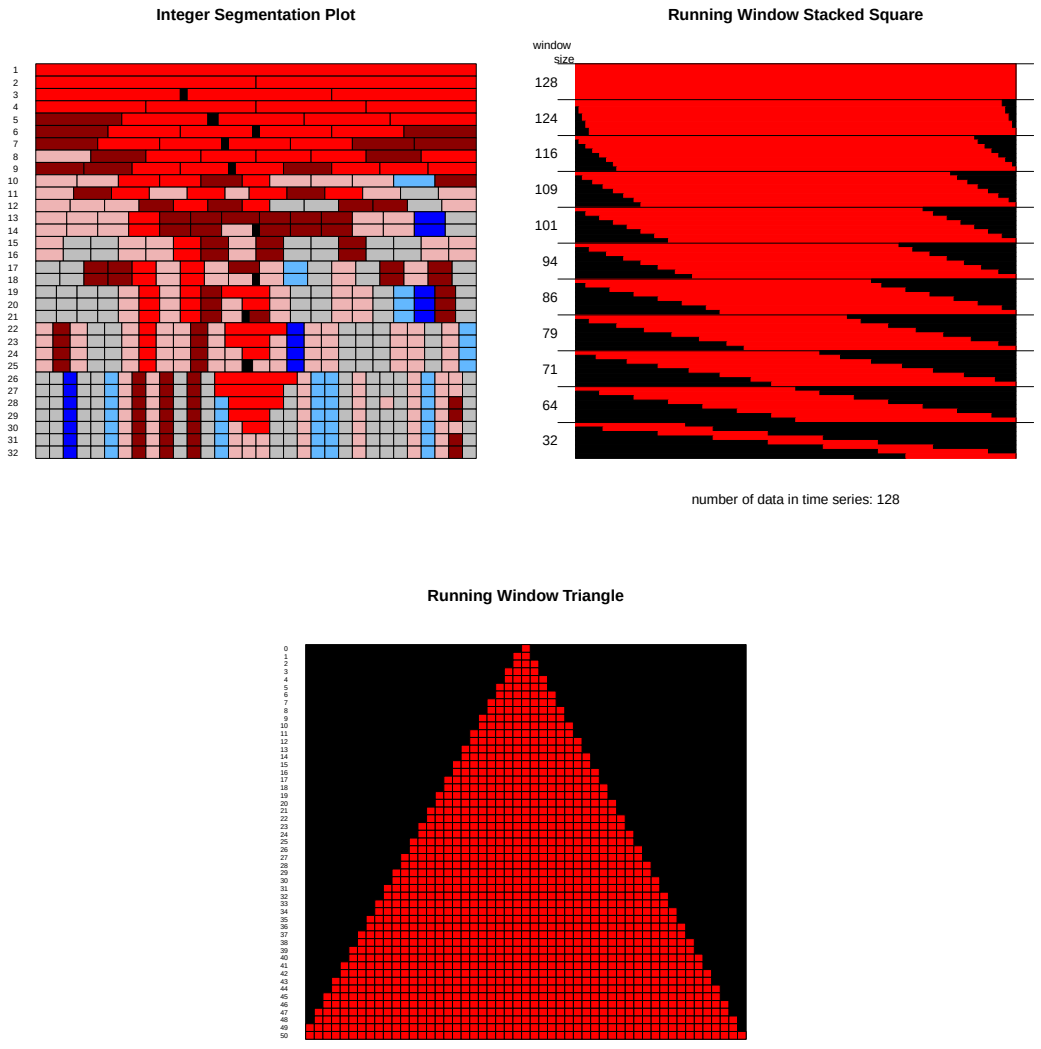


Figure C.10: Example C7.3, Cosine with Gaussian Additive Observational Noise: By construction in this time series the Gauss-Markov assumptions do not apply and so the t-test may not be used. The Integer Segmentation plot clearly shows a highly significant trend in the data down to division of the data by 4 - and a significant trend to division by 7. The Stacked Square and Triangle plots show highly significant trends for all window sizes considered.

C.6 $t^{1.5}$ with Gaussian noise

The following gives another example where the mean values of the data generating process follows a time dependent function but not a linear one. Consider the following time series:

Example C7.4.x $t^{1.5}$ with Gaussian Additive Observational Noise

---

Time series $x_i = t_i^{1.5} + \epsilon_i$, $\epsilon_i \sim N(0, \sigma)$, $t_i = 1.3(\frac{i-1}{N-1})$, where $i = 1, ...N$

Parameters

- C7.4.1: $\sigma = 0.1$, $N = 2^5$
- C7.4.2: $\sigma = 0.1$, $N = 2^6$
- C7.4.3: $\sigma = 0.1$, $N = 2^7$
- C7.4.4: $\sigma = 0.1$, $N = 2^8$
- C7.4.5: $\sigma = 0.3$, $N = 2^8$
- C7.4.6: $\sigma = 0.6$, $N = 2^8$
- C7.4.7: $\sigma = 0.9$, $N = 2^8$

As with the cosine example above the residuals (not shown) show a clear functional relationship with time and so the conditions of the t-test are not met.

Experiment A, cases C7.4.1, C7.4.2, C7.4.3, C7.4.4 Figure C.11 shows that the Breusch-Godfrey test does not reject independence of residuals when there are only 2^5 data points (C7.4.1), nor does it for 2^6 (C7.4.2). When $N = 2^7$ (C7.4.3) independence is rejected at the 5% level but not at the 1% level and when $N = 2^8$ (C7.4.4) independence is also rejected at the 1% level. Normality of residuals is not rejected by the Shapiro-Wilk test despite the fact that by construction they are not Gaussian. The Running Window plots (not shown) show significant trends for all data size. Figure C.12 shows the Integer Segmentation Plots for each case. Each plot shows that the Implied Linear Trend is significant for the whole data set. For case C7.4.1 the significance survives until the data is quartered; with case C7.4.2 until it

is divided into 5. With case C7.4.3 the significance survives for even more divisions which illustrates that there is sampling error. It seems reasonable to conclude, however, that the data show what would be expected: the more data points in the series the more significant the trend can be shown to be (if it is there). The true slope in the data increases with time so that the significance should be higher for later windows than earlier ones, this is evident in all the plots but particularly in case C7.4.4 where the bright red cells go deeper on the right than on the left. This is referred to as a **'wedge'**.

Experiment B: cases C7.4.4, C7.4.5, C7.4.6 and C7.4.7 Figure C.13 illustrates the four time series as the Additive Observational Noise increases. It is notable that the Breusch-Godfrey test only finds serial correlation when the standard deviation is 0.1 (C7.4.4) but this is hidden as the noise level increases (C7.4.5, C7.4.6 and C7.4.7), as such the use of the t-test may not be rejected in the noisier plots which might lead to erroneous conclusions. Figure C.14 shows that the slope is shown by the Integer Segmentation Plot to be highly significant for the full data set in all cases. When the data is divided into 2,3+ segments, however, the prevalence of highly significant (bright red) segments reduces in the noisier plots. Arguably the wedge shape described above is retained in the plots apart from high Additive Observational Noise case C7.4.7.

Likelihood of observing a wedge shape Figures C.15 and C.16 illustrate that a right handed wedge shape is likely to arise for a time series of type C7.4.5. In these figures 2^{10} time series are created from the same data generation process C7.4.5. For each time series the 4th and 8th rows of the Integer Segmentation Plots are created. Figure C.15 shows the log(density) of observing different slope probabilities in each segment. The colour equivalent of these values is shown in the bar at the top of the plot for easy reference. The figures are cropped at $y = -1$. Looking at the highly significant (bright red) end of the plot it is clear that it is more likely that the 4th quarter (and 8th eighths) will be more highly significant than the first. Figure C.16 tests the strength of this finding by bootstrap resampling the 1024 time series into 1024 groups of 512 (with replacement). The boxplots illustrate the variation of probability of observing high signficance in each segment. It is clear that the

probability does rise as the segment number increases from 1 to 8. Therefore it is concluded that a wedge shape is highly likely to arise for a time series of form C7.4.5. Such a wedge may arise in other convex situations though whether this is detectable is likely to depend on the degree of curvature and level of Additive Observational Noise.

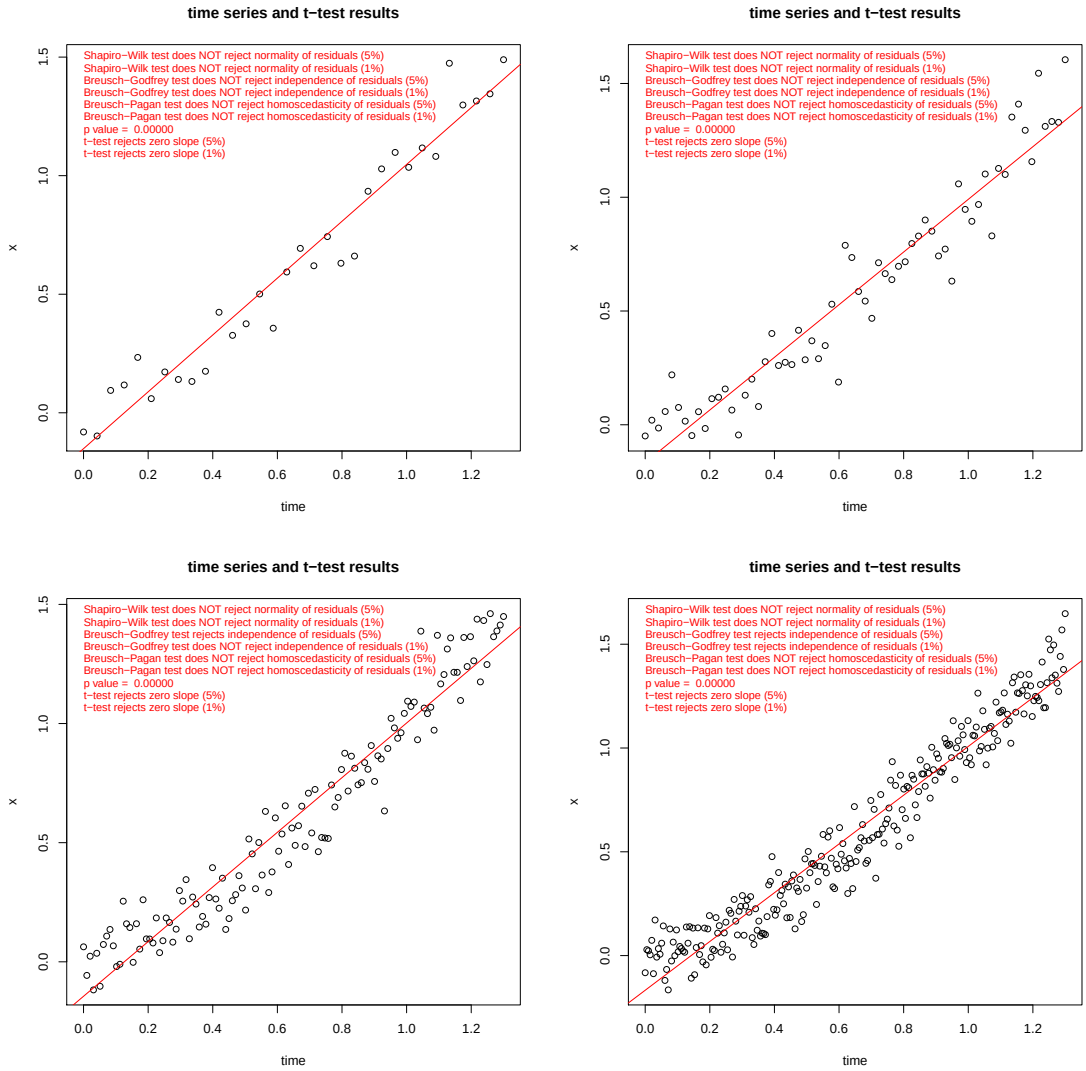


Figure C.11: Examples C7.4.1-4 Time Series plots: $x = t^{1.5}$ example, cases C7.4.1 (top left), C7.4.2 (top right), C7.4.3 (bottom left) and C7.4.4(bottom right). By construction the Gauss-Markov assumptions do not apply, it is therefore notable that the Shapiro-Wilk, Breusch-Godfrey and Breusch-Pagan tests do not reject Guassian, independent or homoskedastic residuals respectively in the top left and top right time series. In the bottom left and right plots the length of the time series is greater and the Breusch-Godfrey test rejects independence.

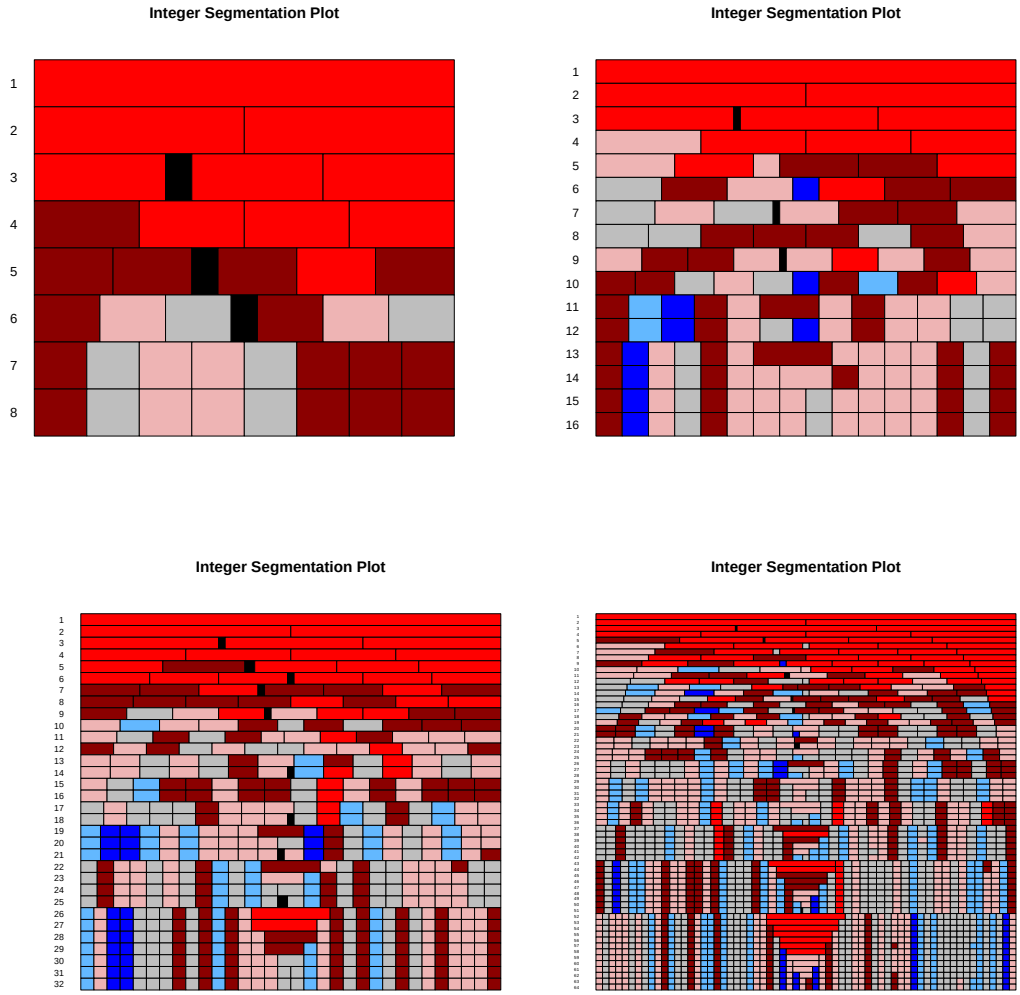


Figure C.12: Examples C7.4.1-4: Integer Segmentation Plots: $x = t^{1.5}$ example, cases C7.4.1 (top left), C7.4.2 (top right), C7.4.3 (bottom left) and C7.4.4(bottom right) Each of the Integer Segmentation plots shows that the trend is significant at least up to where the data is split into thirds. The bottom right plot shows a 'wedge' shape in the bright red coloured segments; consistent with the accelerating slope of a $t^{1.5}$ line.

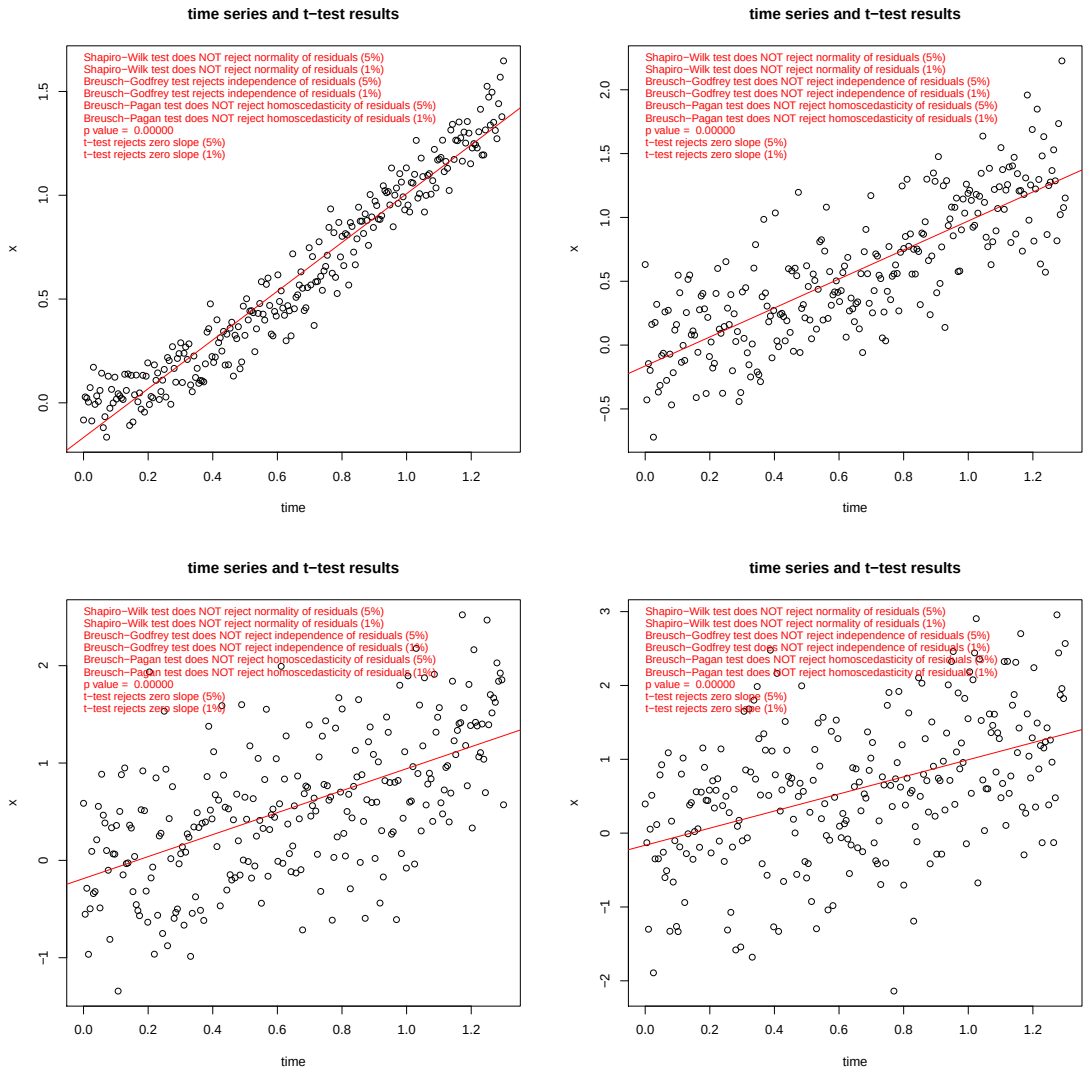


Figure C.13: Examples C7.4.4-7 Time Series plots: $x = t^{1.5}$ example, cases C7.4.4 (top left), C7.4.5 (top right), C7.4.6 (bottom left) and C7.4.7 (bottom right) In this series of plots the length of the time series is 128 but the variance of the Gaussian Noise term increases. It is notable that apart from the lowest variance plot (top left) the Breush-Godfrey test does not reject independence of residuals, as with figure C.11 it would appear that the t-test can be used despite this not being the case by construction.

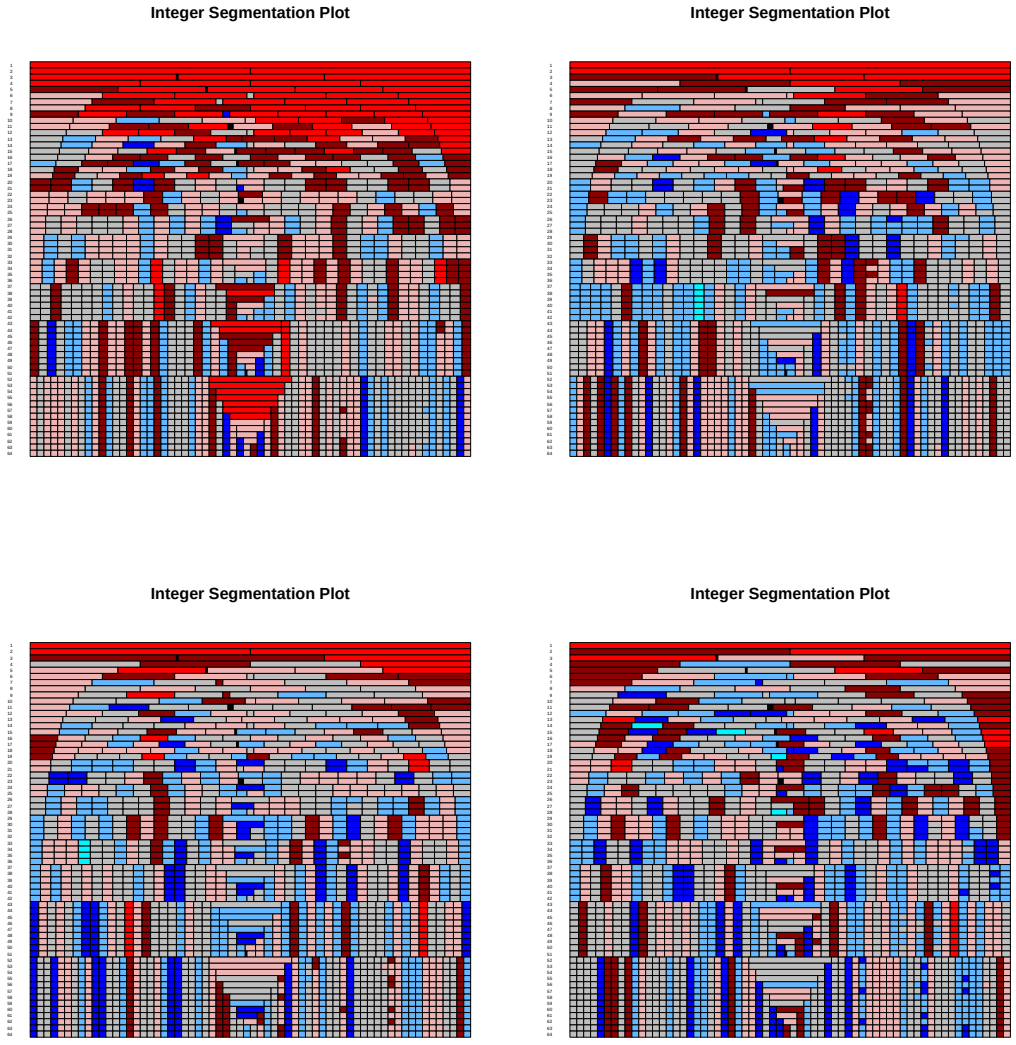


Figure C.14: Examples C7.4.4-7: Integer Segmentation Plots: $x = t^{1.5}$ example, cases C7.4.4 (top left), C7.4.5 (top right), C7.4.6 (bottom left) and C7.4.7 (bottom right). The row numbers are difficult to read at this scale but run form 1 to 64. Each plot shows a highly significant trend in the full data set and a significant trend when the data is halved. The wedge shape described in figure C.12 is arguably retained in the top right and bottom left figures here.

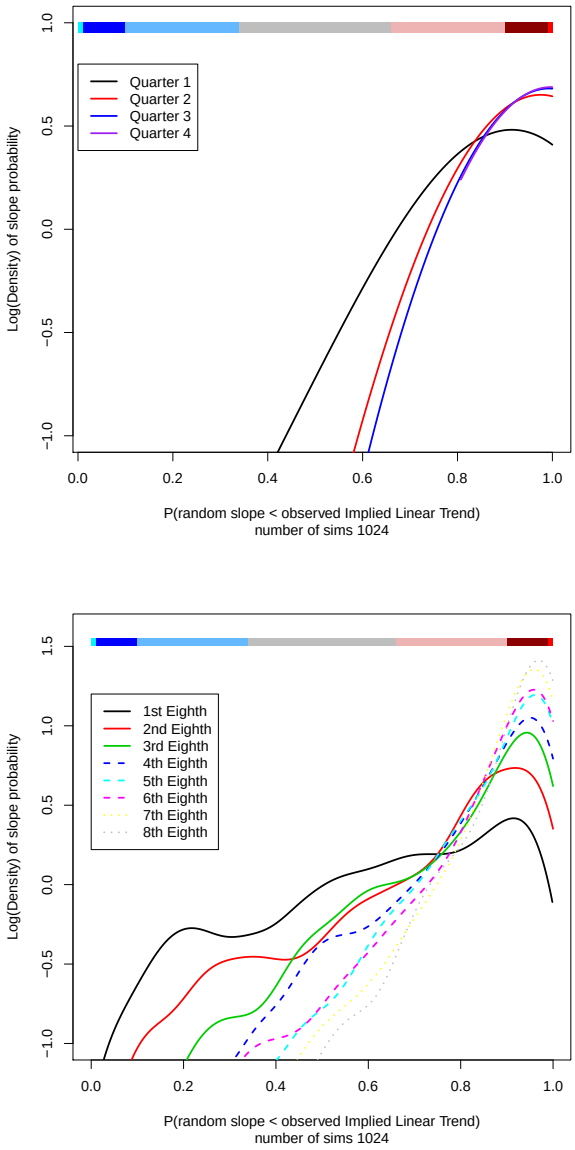


Figure C.15: Figure to illustrate that a time series of type C7.4.5 ($t^{1.5}$) is likely to lead to a right handed wedge in the Integer Segmentation Plot. Figure shows (y-axis) Log(Density) of simulations that have a given slope probability (x-axis), based on 1024 simulations. The colour key is shown as a strip at the top of the graphic for easy comparison. The top figure shows the results when the data set is quartered, the bottom figure shows the results when it is divided into 8.

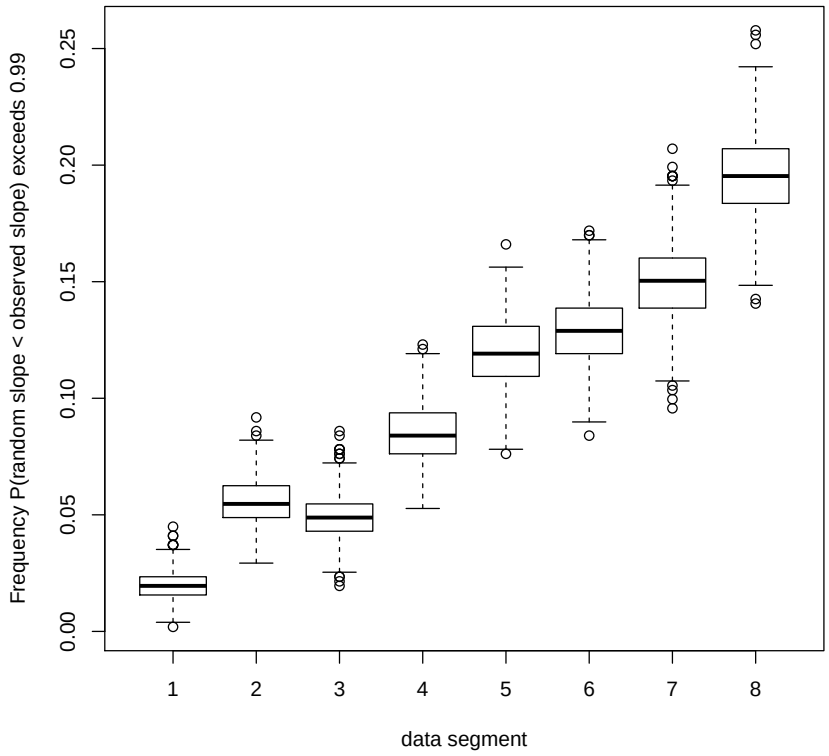


Figure C.16: These Box plots reflect the frequency with which high slope probability is detected in a sub segment, as a function of where that sub segment lies in the time series. The clear increase in frequency reflects the fact that detection in later segments is much more likely than in earlier segments; the trend supports the expectation that a right-handed wedge is very likely. The boxplots show results of bootstrap resampling of slope probabilities; 1024 resamples of 512 subsamples from 1024 time series. The frequency of occurrence of a slope probability that exceeds 0.99 is shown (the probability of observing a bright red segment). Conclusion: the 8th segment is more likely to be bright red than the 1st: a right handed wedge is very likely for time series of type Examples C7.4.5.

C.7 Atmospheric extreme events

The examples shown in the following three subsections use data taken from Neumayer and Barthel [186]. Using data from Munich Re, they explore the normalised (equation C.15) insurance losses (damages) arising from various atmospheric extreme events using the t-test, without discussion of its appropriateness. All financial values are in billions of US Dollars (USD bn). The purpose of this section is to compare conclusions from the slope probability method with those of the t-test, not to criticise the normalisation methods discussed in their paper. In the context of insurance, trends in atmospheric extreme events might indicate the pricing based on past claims is in danger of understating premiums.

Losses are normalised because reported losses at different times in the past may not be comparable. Past storms may have hit when population density was low so the damage caused is less than it would be today. Wealth levels were different in the past and so the likelihood of high value contents being damaged was lower. Also, general monetary values (proxied by the Gross Domestic Product[5], GDP) were lower - so a dollar 50 years ago was worth more than it is today (inflation). Neumeyer and Barthel [186] note there are a number of normalisation methods[6] used in the literature, the following being introduced by Pielke and Landsea in 1998 [200] for events that occur at time t, as follows:

$$ND_t^s = D_t \frac{G_s}{G_t} \frac{P_s}{P_t} \frac{W_s}{W_t} \tag{C.15}$$

where, ND_t^s is the Normalised damage (insurance losses) to time s from time t; D_t is the reported damages at time t; G_t is the GDP (inflation) deflator at time t; P_t represents population density at time t; W_t is the wealth per capita at time t.

[5]GDP - a measure of the value of total goods and services produced by a country in a given yaer

[6]Neumeyer and Barthel argue that if frequency of events (alone) increases or severity of the events (alone) increases then the loss levels in recent times would be certain to increase, but, they state, this is not the case with equation C.15. Part of their work is to propose an alternative approach which aims to improve the normalisation method. Their new method is **not** presented here since the purpose of this section is mostly to establish that the t-test is misused and then consider the Slope Probability approach.

The following three examples are considered:

Example C7.5.x Atmospheric extreme events
C7.5.1 Convective events USA
C7.5.2 Convective events Western EU
C7.5.3 Normalised hurricanes losses in USA

Table C.1 compares the slope significance results from Neumayer and Barthel with the new Slope Probability method. The table shows that the traditional and slope probability methods do not always agree. For example, the slope probability method finds a significant trend in US hurricanes where Neumayer and Barthel do not. The table also shows that the conditions required for the use of the t-test are not met in any of the data sets.

Table C.1: Comparison of slope significance results between Neumayer and Barthel and Slope Probability

Time series	t-test applicable?	Neumayer and Barthel	Slope Probability
Convective events USA	No	Significant	Highly significant
Convective events W.EU	No	Not significant	Not significant
US Hurricanes	No	Not significant	Significant

C.7.1 Example C7.5.1: Convective events USA

This subsection considers convective events in the USA. Data is that underlying Figure 7a of Neumayer and Barthel which is titled 'Disaster losses from convective events in the United States normalized with conventional approach'. The Normalised losses are based on 1,771 disasters including damages from flash floods, hail storms, tempest storms, tornados, and lightning. To determine whether the time series has a trend they use a t-test with a 10% significance level but they do not state whether the necessary conditions to use the t-test are met.

Whilst the residuals appear normally distributed according to the Shapiro-Wilk test they do not pass the Breusch-Pagan test for homoscedasticity - nor do they pass the Breusch-Godfrey test for serial correlation -therefore the conditions of the t-test do not apply.

The Integer Segmentation Plot in figure C.17 shows the slope is highly significant. When the data is halved only the first half of the data retains a highly significant slope. This agrees with intuition by looking at the time series graphic which contains material low values in the second half of the time series. When broken into thirds the strongest (and highly significant) slope is in the middle third; again with the weakest slope to the right of this.

The Running Window Triangle shows a similar result within figure C.17. The slope remains highly significant when up to 7 data points are removed. The slope is also significant when more points are removed for the left hand portions of the data - but less so for the right. This might call into question continuation of the trend in the future.

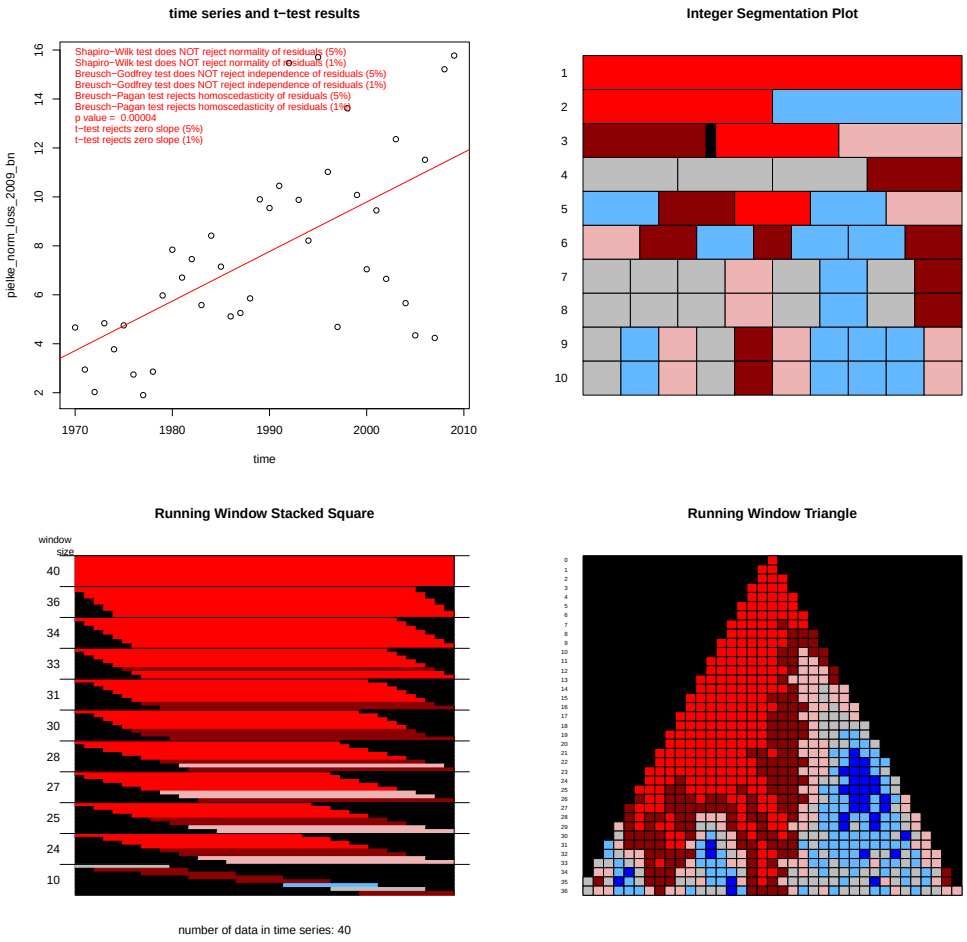


Figure C.17: Example C7.5.1: Convective events in the USA. The Integer Segmentation Plot and the Running Window Triangle plots clearly show that, whilst the full time series shows a highly significant trend, this level of significance is only retained in the first half of the data set when the data is subdivided. The Stacked Square plot, however, shows that the positive trend is significant for windows to size 30 and is retained in many windows to size 24.

C.7.2 Example C7.5.2: Convective events Western EU

This subsection considers convective events in Western Europe and uses the data underlying Neumayer and Barthel's Figure 7b which is described as 'Disaster losses from convective events in Western Europe normalized with conventional approach'. The losses are based on 1,296 disasters including damages from flash floods, hail storms, tempest storms, tornados, and lightning.

The Breusch-Godfrey test does not reject independence of residuals and the Breusch-Pagan tests do not reject homoscedasticity. The Shapiro-Wilk test, however, does reject normality of the residuals and so the conditions for use of the t-test

are not met.

The Integer Segmentation Plot shown in figure C.18 shows that random resamples of the data produce slopes that are more steep over 10% of the time. Hence this trend is not significant. It is significant in the second half of the data set since this is not affected by the large events in the first half of the data.

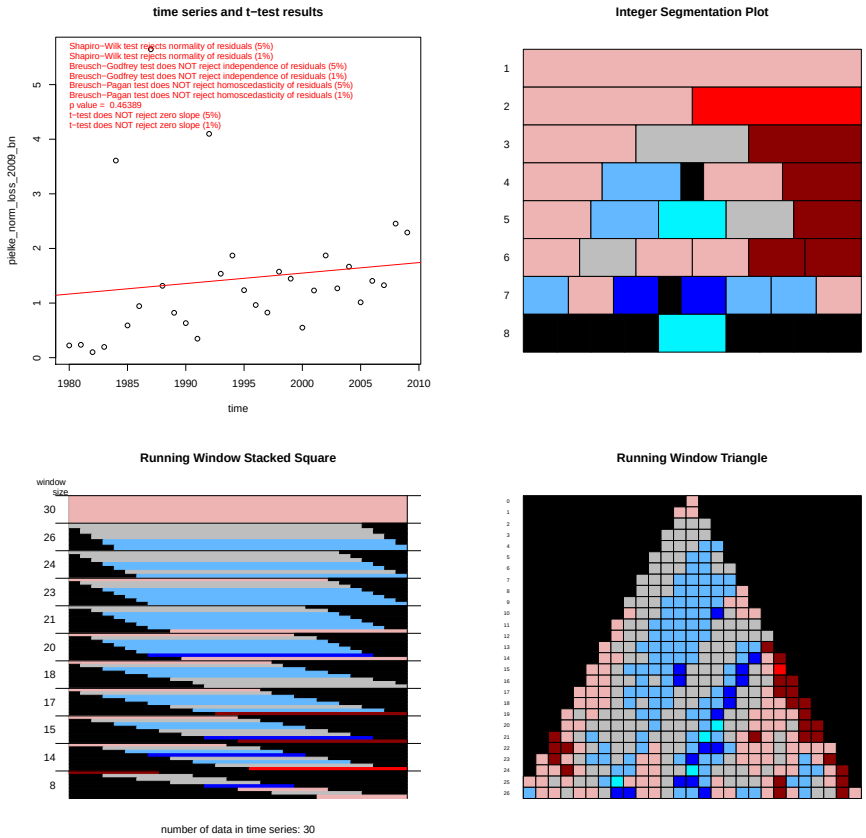


Figure C.18: Example C7.5.2: Convective events in the EU. The trend line in the top left plot is shown not to be significant in the Integer Segmentation Plot although there is a highly significant trend in the second half of the data set.

C.7.3 Example C7.5.3: Hurricane losses in USA

This subsection considers losses from hurricane events in the United States. Data is that underlying figure 7c of Neumayer and Barthel. Their plot is titled 'Disaster losses from hurricanes in the United States normalized with conventional approach'. The data is based on based on 118 disasters.

The Breusch-Godfrey test does not reject independence of residuals and the Breusch-Pagan test does not reject homoscedasticity of residuals. The Shapiro-wilk

test, however, does reject normality of the residuals and so the t-test should not be used. The Integer Segmentation plot in figure C.19 suggests a significant trend (1-10%) for the full data set, but this not significant for any other data subdivision. It appears that the large events in the second half of the time series are the main cause of the trend.

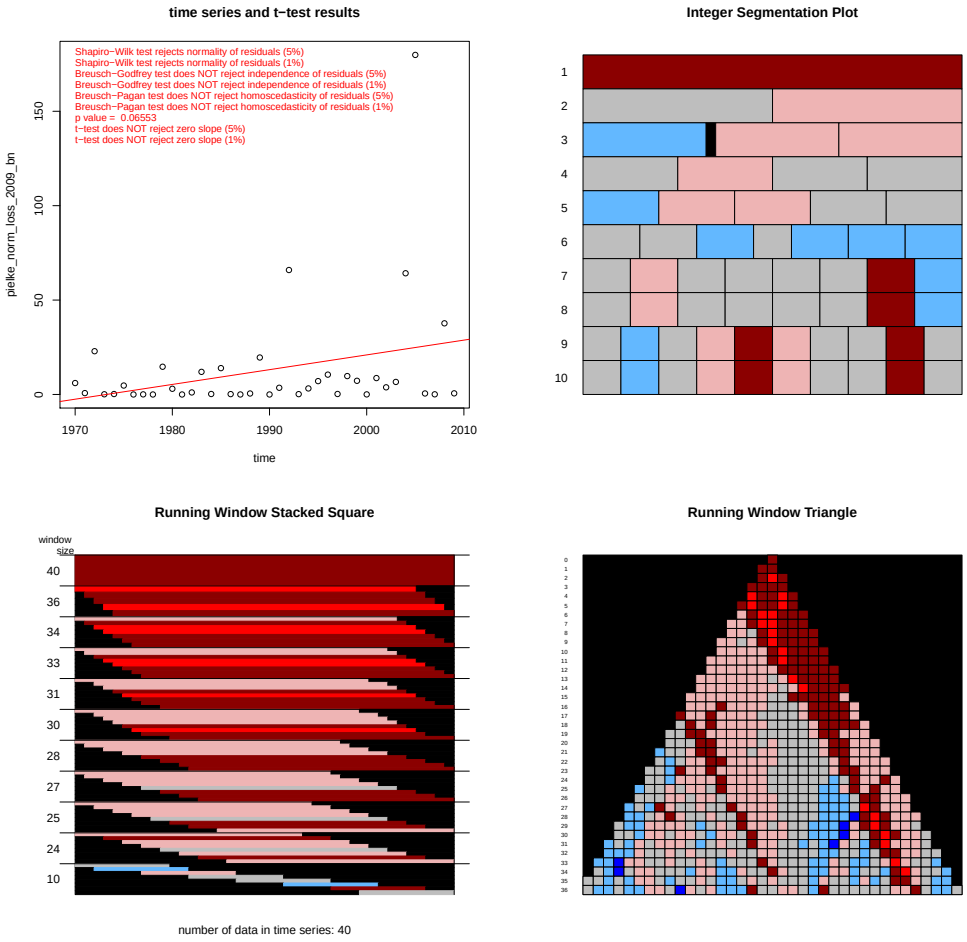


Figure C.19: Example C7.5.3: Normalised hurricane losses - USA. The Integer Segmentation plot shows a significant trend in normalised hurricane losses over the period. The Stacked Square and Triangle plots show that some windows of size 31-36 are highly significant.

C.8 Lorenz 63

The Lorenz 63 system was introduced in 1963 [152] to examine the feasibility of long range weather prediction. The equations are as follows:

$$\frac{dx}{dt} = \sigma(y - x)$$

$$\frac{dy}{dt} = x(\rho - z) - y \tag{C.16}$$

$$\frac{dz}{dt} = xy - \beta z$$

Example C7.6 Lorenz 63

Lorenz originally illustrated the case when $\sigma = 10$, $\rho = 28$ and $\beta = \frac{8}{3}$ - those values are also used here. A sample of $N_{obs} = 626$ values for the x variable are taken from the system - observations are made every at time values in steps of 0.04.

The data is suggestive of a two level process consistent with trajectories that alternate between the two 'lobes' of the attractor figure C.20(top) . This system exhibits deterministic chaos for some values of the parameters [59]. The Shapiro-Wilk test rejects normality and the Breusch-Pagan test rejects independence so the t-test does not apply. The Integer Segmentation Plot within figure C.21 shows that the slope of the regression line through the full data, whilst positive, does not have a significant Slope Probability. Despite this there are many sub-segments which are significant (and indeed highly so). This first appears when the data is quartered. When the data is divided into six segments there are three (second, third and fourth) which have highly significant alternating negative and positive trends. These coincide with trajectories that are in one lobe for half the segment and then jump to the other lobe for the other half. The last third of the data shows more frequent transitions between lobes and so no trends show up when the data is segmented into sixths in this region. When the data is divided into 125 segments alternating bands significant blue and red arise, these identify the trajectories around the approximately circular regions within a given lobe which appear as cycles when projected onto the x dimension.

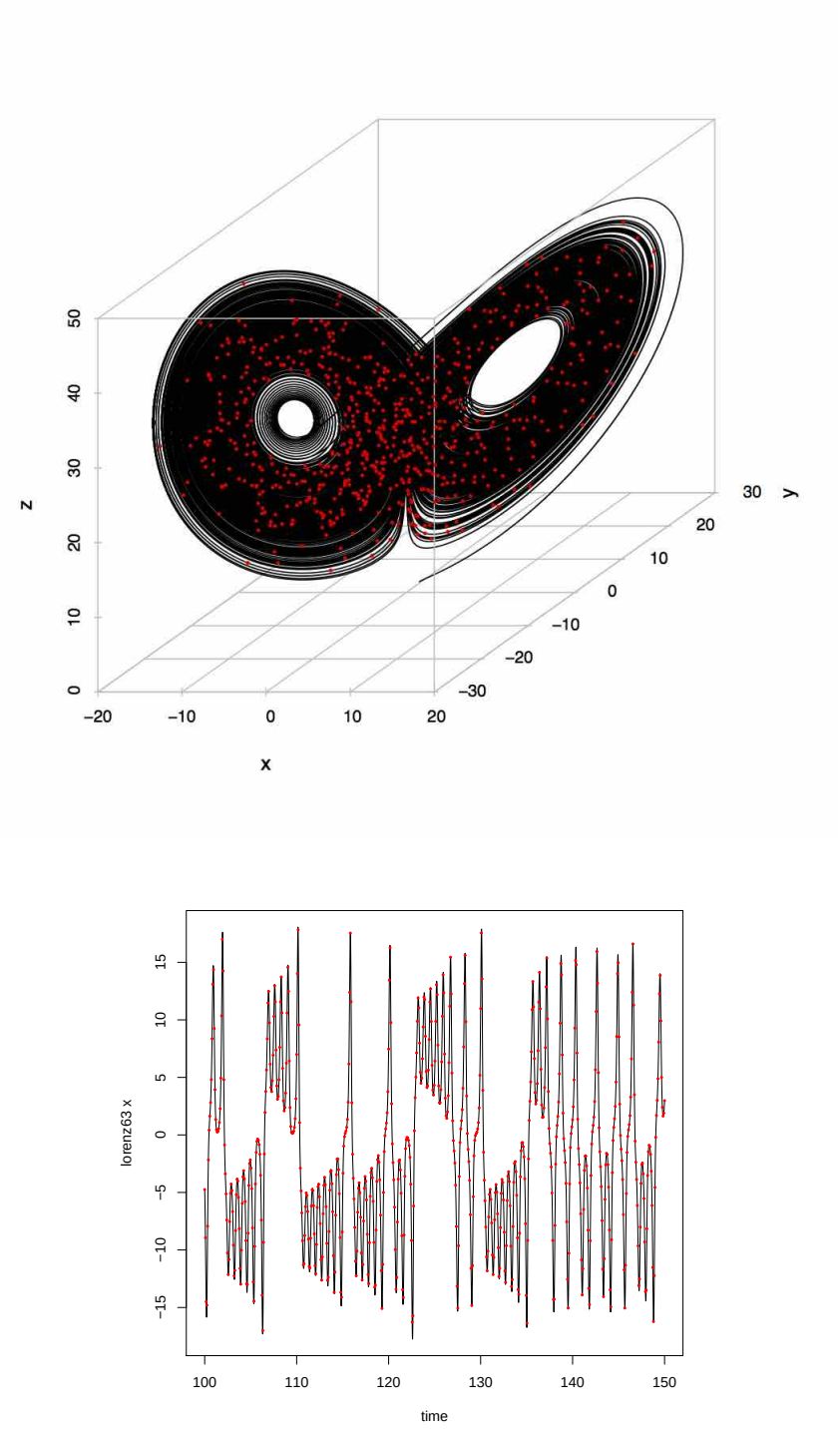


Figure C.20: Sample of values from Lorenz 63 system. The right hand plot shows the trajectory of x against time with the sampled values shown in red; the left hand plot shows the trajectory in xyz phase space - again with the sampled values highlighted in red.

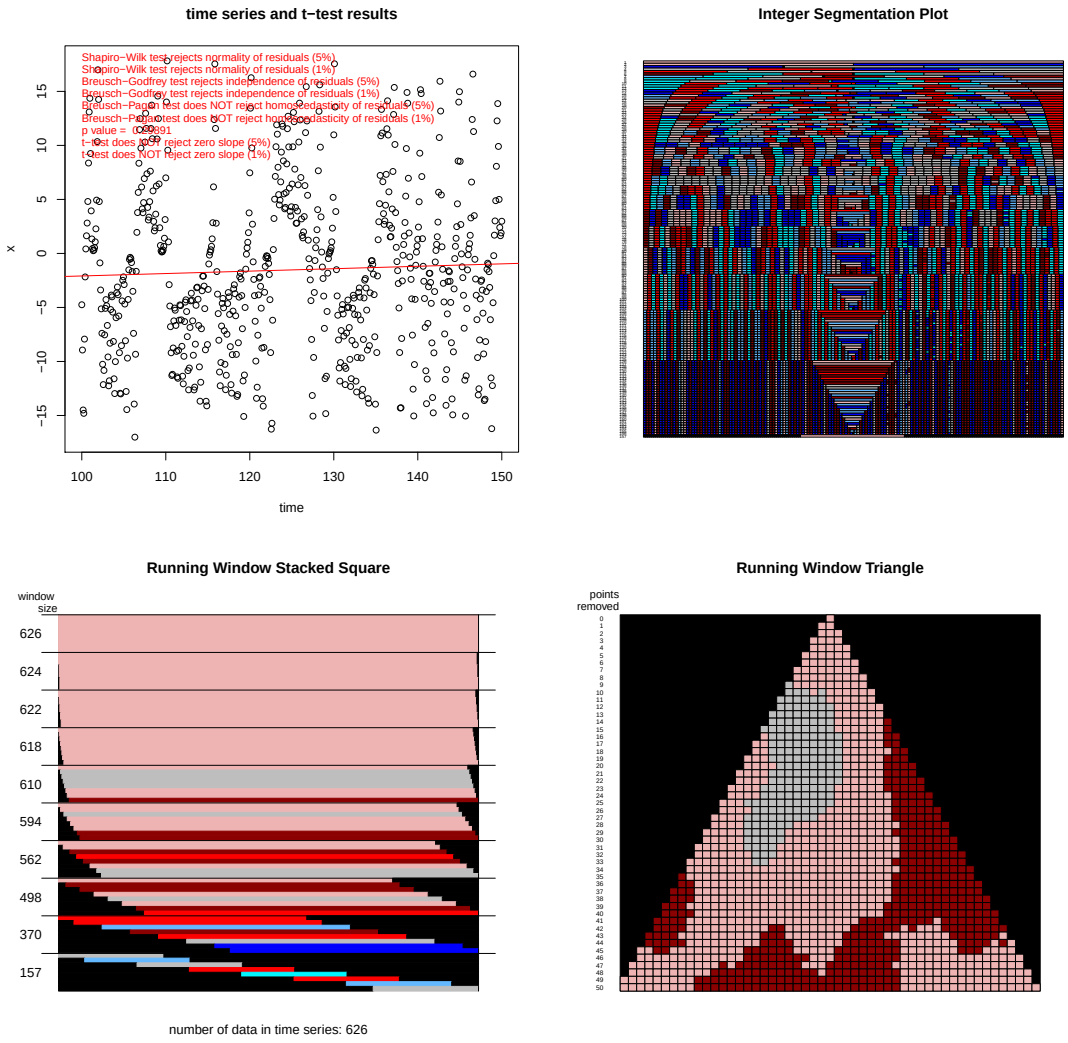


Figure C.21: Example C7.6: Lorenz 63 sample of x values. The Integer Segmentation, Triangle and Stacked Square plots all show that whilst there is a positive slope in the data this is not significant. The Integer Segmentation Plot has considerable structure, consistent with trajectories which alternate between the lobes of the Lorenz 63 attractor and also higher frequency cycles evident when the colours alternate between red and blue when the data is divided into more than 125 segments.

C.9 Tide Gauge Data - New York Battery

The Permanent Service for Mean Sea Level (PSMSL), based in Liverpool UK, was established in 1933 [197]. It collects sea level data from tide gauges around the world. Each gauge is given a unique GLOSS number (Global Sea Level Observing System). This section illustrates the sea level rise observed at the New York Battery.

Example C7.7 Tide Gauge Data - New York Battery

Data from Holgate et al., 2013 [197] extracted on 19 May 2015 from PSMSL website http://www.psmsl.org/data/obtaining/.

Figure C.22 shows the location of the tide gauge. Figure C.23 shows the time series itself, the Integer Segmentation Plot and the Running Window graphics. There is a gap in the data set just prior to 1900; this is not a problem since the set of time points as the original data is always held fixed, with the observations randomly allocated to those times. The Running Window Plots find a significant slope at all observed windows.

There is a suggestion of a right handed wedge (the concept described in section C.6) in the Integer Segmentation Plot of figure C.23. The slope is significant on the far right of the plot for all segments down to division of the data by 12, whereas on the left the significance on survives to subdivision into sixths. As suggested in section C.6 such wedges may arise in a convex (accelerating) situation though this is not proven in general. If sea level rise is accelerating this may explain the wedge shape in the Integer Segmentation Plot.

Testing whether there is a wedge in other gauge data There are 1414 gauges as at May 2015 in the PSMSL data set, these are illustrated in figure C.24. The slope probability is calculated for each of these data sets (full data) and a dot is plotted on the map at the location of the gauge; the colour of the dot is the corresponding slope probability on the usual scale. 1042 tide gauges show an increasing trend.

When testing for a wedge in the data it will be necessary to subdivide the

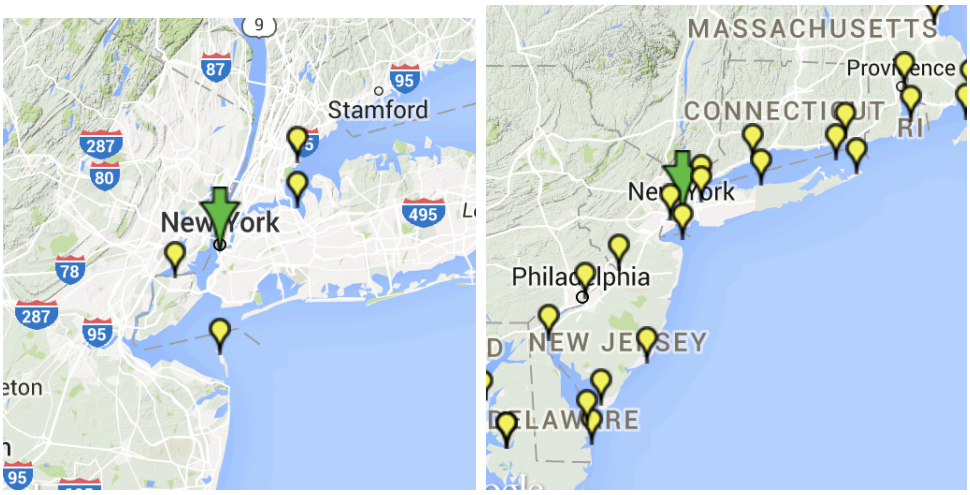


Figure C.22: Example C7.7: Tide Gauge Data: Location of NewYork gauge - number 12, denoted by green arrow.

data into equal sized segments. Of the 1414 gauges, 705 have less than 25 data points which is arguably not long enough to subdivide[7], these are removed from the analysis below and are highlighted with a black dot at the centre of the coloured plot character. This leaves 709 gauges with more than 25 data points. Some of the remaining gauges (165 of them) have negative slope; these are removed from analysis because the analysis below is conditional on the slope being positive. The negative slope cases are discussed further below. The remaining 544 gauges have more than 25 data points and a positive Implied Linear Trend and are referred to as the **'restricted set'** below.

The following analysis is carried out to asses whether the wedge shape suggested in the New York gauge data can be detected in the restricted set. Given time constraints (but also data length constraints) this was only assessed for row 4 of the Integer Segmentation Plot. The data was quartered for each of the gauges and the implied linear trend calculated for each (if the data series does not divide perfectly into 4 the residual data is discarded). The slope probability is also calculated in each case. Therefore, for each gauge (i), there are two derived series: (1) The slopes of the data quarters $s_{i,1}, ...s_{i,4}$ and (2) The slope probabilities for each quarter $p_{i,1}, ...p_{i,4}$.

The implied linear trend is calculated for each of the derived series (i.e. calculate

[7]Houston and Dean [109] suggest that any tide gauge less data set than 75 years should not be analysed due to natural cycles and other exogenous factors, I have chosen to retain shorter time series in this analysis.

the trend for the four data points $s_{i,1}, ..s_{i,4}$ and again for probabilities p), call this the **'4-block trend'** below. For the slopes s a positive trend is indicative of accelerating sea level rise (or at least relative rise) and for the probabilities p a positive trend is indicative of either (1) an accelerating rise, as with the $t^{1.5}$ example, or (2) decreasing size of error terms. If the 4-block trend is positive for both the slopes and slope probabilities it is arguable that acceleration is the most likely cause.

Of the 544 gauges in the restricted set some 67% have a positive trend in their sub-slopes $s_{.,j}$ and 62% have a positive trend in their slope probabilities $p_{.,j}$. The correlation between s and p is 67%. This is depressed due to several outliers amongst the slopes. If these are removed the correlation increases to 78%. The outliers are plotted red in figure C.25.

It is not immediately obvious whether the proportion of gauges with positive trends in their sub-slopes (or slope probabilities) is significant. It is possible that such proportions could occur randomly with high likelihood. To assess the significance of these proportions the following approach is taken:

- For each gauge (y_i) in the restricted set of gauges $(y_1, ...y_{544})$; determine a linear model through the data using least squares $\hat{y}_i(t) = a_i + b_i t$
- For gauge (y_i) create $j = 1, ...G$ pseudo gauge data sets $(\tilde{y}_{j,i})$ by adding random errors to the fitted values, i.e. $\tilde{y}_{j,i}(t) = \hat{y}_i(t) + \epsilon_i(t)$, where $\epsilon_i(t) \sim N(0, \sigma_i)$ and σ_i is the standard deviation of the residuals $(y_i - \hat{y}_i)$. If the linear model above describes the process underlying the generation of the gauge data then each of these pseudo data sets could have arisen in theory.
- For each pseudo gauge $\tilde{y}_{j,i}$ subdivide the data into four blocks (discarding remaining data as above) and calculate the implied linear trend for each block $s_{j,i,1}, ...s_{j,i,4}$ and also the slope probability $p_{j,i,1}, ...p_{j,i,4}$. As for the true gauge data calculate the 4-block trend for the $s_{j,i}$ and $p_{j,i}$. Determine whether these are positive, creating indicator variables $\hat{s}_{j,i}$ and $\hat{p}_{j,i}$ which is 1 if the 4-block trend is positive and 0 if not.
- For each gauge in the restricted set, randomly choose a number $(j(i))$ between 1 and G such that each choice is equally likely. Let $f_p = \frac{\sum_{i=1}^{544} \hat{p}_{j(i),i}}{544}$, the fraction of gauges that have a positive trend in their slope probability and define f_s similarly for positive trends in the slopes themselves.

- Repeat the above step M times with different random selections $j(i)$.
- In this analysis $G = 2^7$ and $M = 2^{15}$.

Based on the above method, confidence intervals for the proportion of gauges with a positive 4-block trend in slopes and slope probabilities can be estimated by taking quantiles of the above results. These suggest that **if** a linear trend with (gauge dependent) Gaussian noise is an appropriate model for the gauge data then the proportion of 4-block trends for the slopes that are positive has a symmetric 99.9% confidence interval of (43.0% 56.8%) with a median of 50.0%; and for the slope probability a 99.9% symmetric confidence interval of (42.3% 55.9%) with a median of 49.1%. As such, the calculated values of 62% and 67% are materially outside the confidence intervals and suggest that the hypothesis of a linear model with Gaussian noise should be rejected. The IPCC fifth assessment report series argues that sea level rise *is* faster in the 20th century than the 19th [114] and also that the rate is likely to increase further this century [115]. There is, however, an active debate on whether acceleration has been detected [263]: some authors find evidence for acceleration(for example [25, 206, 275]), whilst others find evidence for none or even deceleration (for example [108, 265]). The analysis in this appendix supports the evidence for accelerating sea level rise but does not claim to be definitive and uses only one technique which the results are dependent on.

The gauges with negative slope have also been analysed. In their case the question is whether a decelerating negative slope can be detected. This would be consistent with two opposing forces (1) the force causing the observed sea level to decrease (such as glacial rebound) and (2) Sea level rise due to warming water. This analysis has been carried out using the same method as above to estimate confidence intervals of the observed 4-block trends in slope and slope probabilities. In this case the proportion of 4-block trends for the slopes that are positive has a symmetric 99.9% confidence interval of (38.1% 63.6%) with a median of 50.1%; and for the slope probability a 99.9% symmetric confidence interval of (38.8% 63.6%) with a median of 50.1%. As such, the observed values of 46.1% and 48.5%, respectively are not significant and a there is no evidence for a decelerating negative slope.

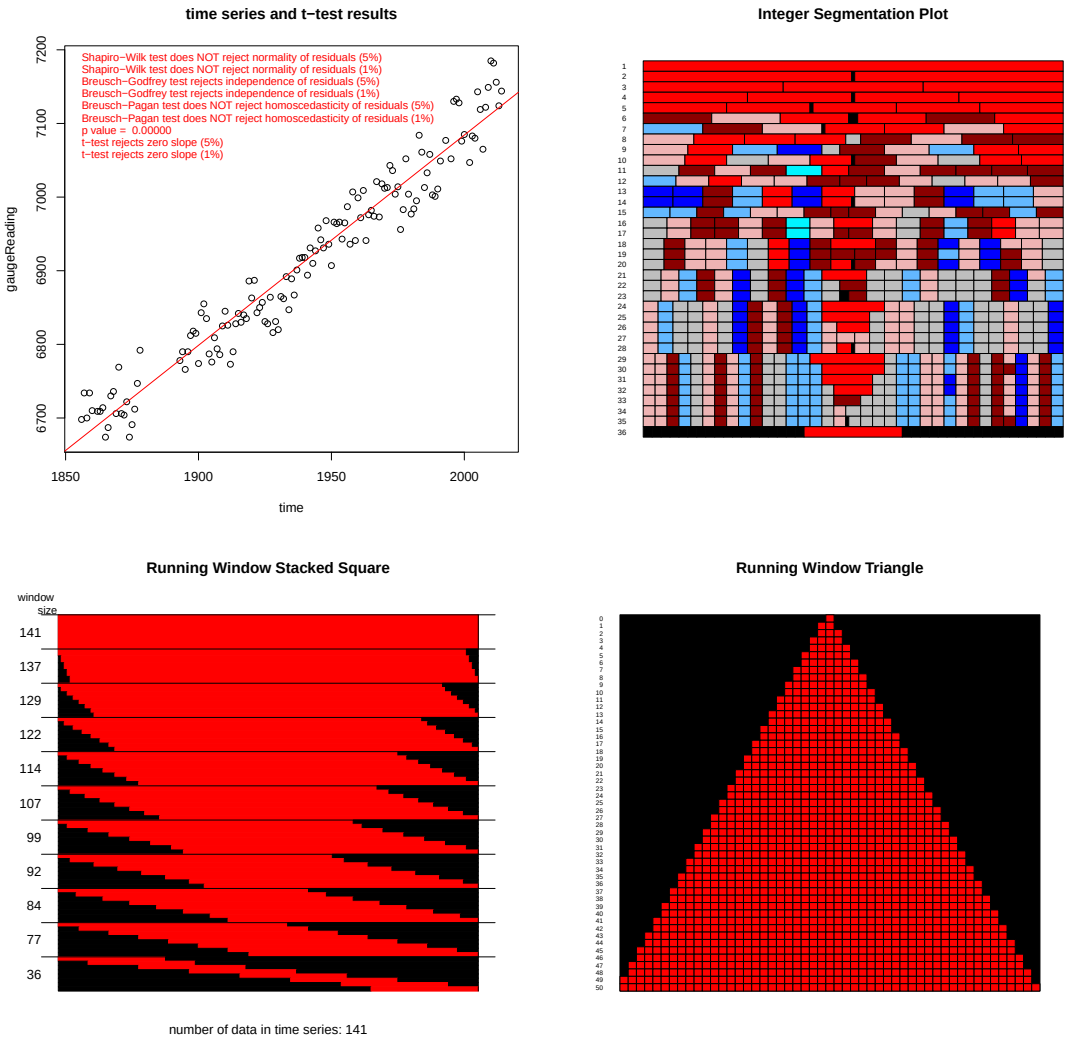


Figure C.23: Example C7.7: Tide Gauge Data: NewYork - number 12. The Triangle and Stacked Square plots show a highly significant trend for all window sizes considered. The Integer Segmentation plot shows that the trend remains significant up to when the data is divided into 5. There is some evidence of a 'wedge' shape (similar to that of figure C.12) where the significant segments appear more on the right of the plot than the left down to where the data is divided by 12.

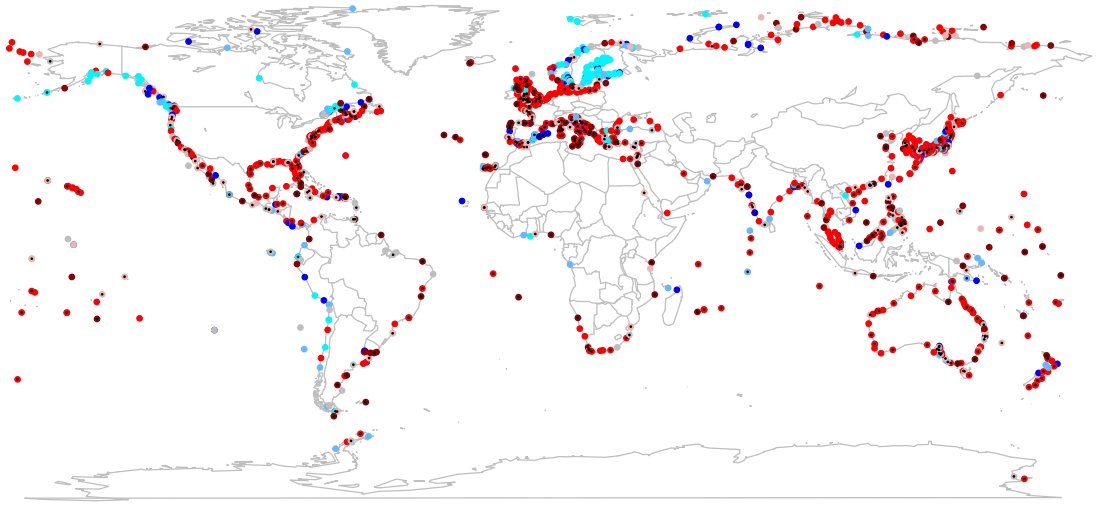


Figure C.24: All tide gauges plotted at their geographical location. Colour indicates slope probability as per colour key. Red shades arise when sea level is rising at the given location and blue for falling levels. A black dot within the plot character indicates a gauge with a short time series (less than 25 data points).

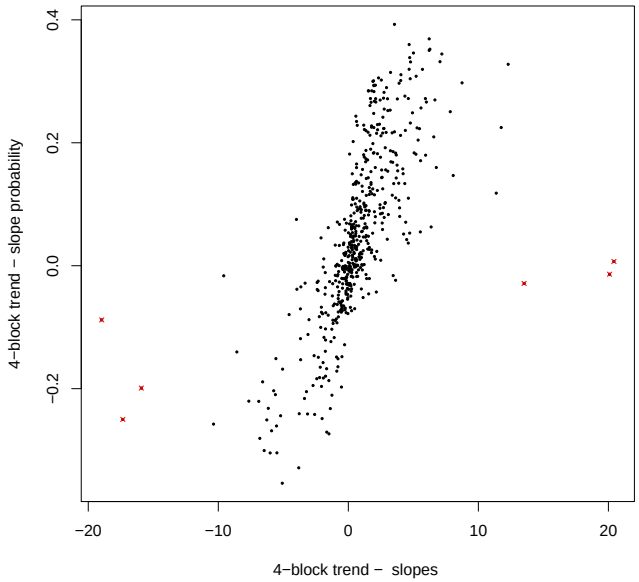


Figure C.25: Tide gauge data, restricted set. 4-block slopes versus, 4-block slope probabilities. Outliers highlighted with a red cross. Data indicates 67% correlation between the slopes and slope probabilities which rises to 78% with the outliers removed.

C.10 Sunspots (ISSN)

The data is described by SILSO Royal Observatory of Belgium, Brussels as 'Yearly mean total sunspot number obtained by taking a simple arithmetic mean of the daily total sunspot number over all days of each year.' [228] . They note that the yearly data is produced by averaging daily data which will give a different figure to averaging the monthly data that is also supplied. Furthermore they note that there are some differences in observation methods over the time range and some interpolation was used prior to 1849 when the counts were not available every day. Originally, the observations were collected by Rudolph Wolf and so the data is sometimes referred to as 'Wolf Sunspot Numbers'.

Example C7.8 Sunspot data

Data extracted on 19 May 2015 from SILSO website http://www.sidc.be/silso/datafiles. Data includes annual sunspot numbers from 1700 to 2014.

The Shapiro-Wilk test rejects normality of residuals, Breusch-Godfrey rejects independence of residuals and the Breusch-Pagan test rejects their homoscedasticity hence the t-test is not meaningful. The Running Window Triangle within figure C.26 shows a significant upward trend in the whole data set which is retained in the majority of windows with up to 50 points removed. The trend is not significant in the first half of the data but retained in the second half. When the data is subdivided into four or more groups the start to be significant and even highly significant negative trends. The Integer Segmentation Plot's bands of alternating bright red/cyan (or sometimes dark red/ dark blue) become evident when the data is divided into ∼40 segments and pronounced when the data is divided into 50+ groups. Similar to the Lorenz 63 example before this is suggestive of cycles in the data. Given there are around 300 data points this would lead to 6 observations per segment. Hence each red/blue couple spans around 11-12 years which ties in with the approximately 11 year solar cycle [95].

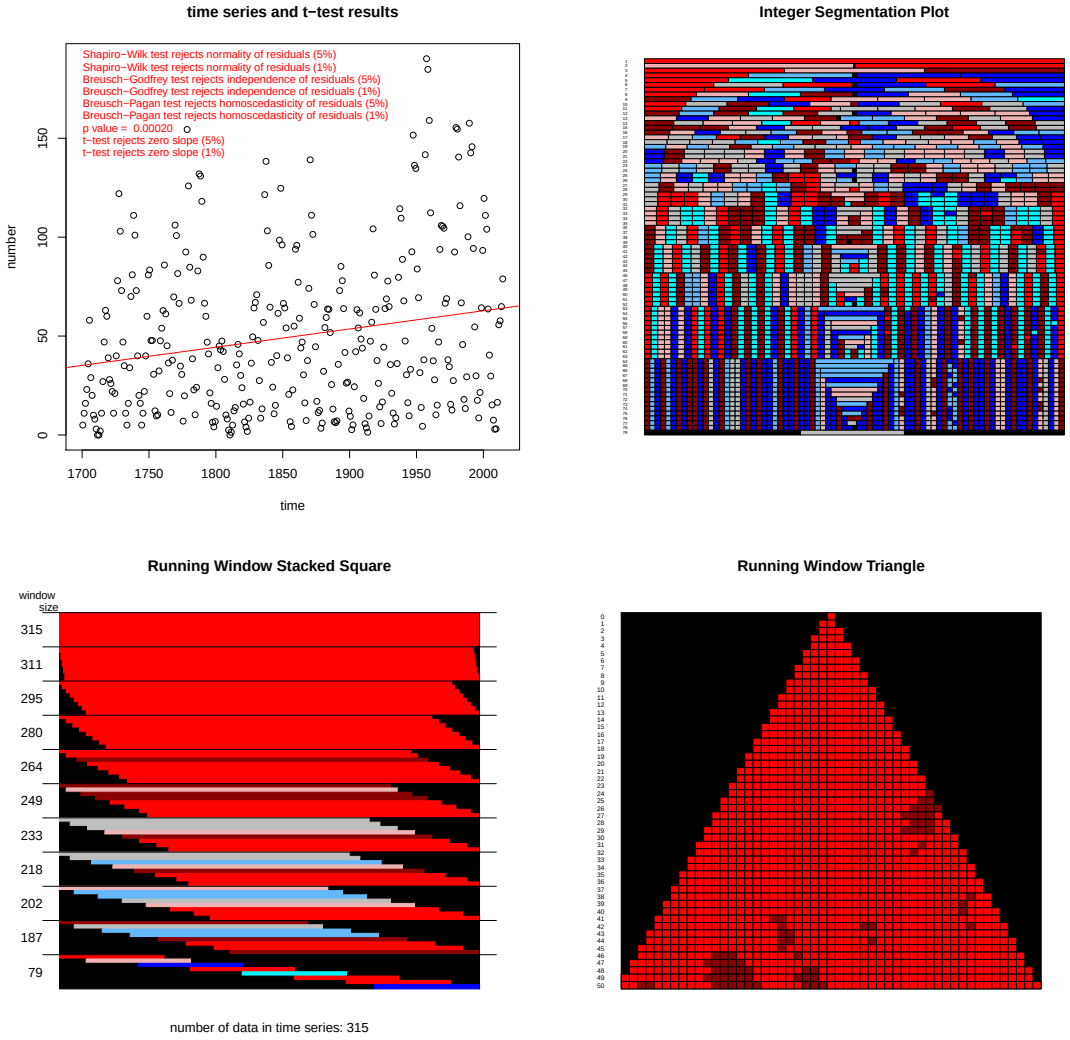


Figure C.26: Example C7.8: Sunspot numbers: SIDC. Annual mean. All plots show a highly significant trend in the full data set. The Integer Segmentation Plot has considerable structure and alternating bands of red and blue are consistent with the 11 year solar cycle when the data is subdivided into 50+ groups.

C.11 Conclusions

This appendix has discussed trend detection and proposes a permutation method to define the significance of calculated slope parameters. There are multiple definitions for trends [98, 217, 225, 256, 263] and trend detection [98, 207], many requiring strong assumptions on the behaviour and distribution of the residual[8] terms. Similar to other non-parametric approaches [225, 256] the Slope Probability method proposed here does not require these assumptions. As such, the Slope Probability method can be used in many situations where parametric methods are invalid. This appendix explores multiple synthetic and real data sets through the use of three novel graphical methods (Integer Segmentation, Running Window Stacked Square and Running Window Triangle) each using the Slope Probability to test significance of trends for different subdivisions of the data. These graphical methods are complementary illustrating different features of the time series analysed. For example the Integer Segmentation plot suggested a new way to characterise accelerating sea level rise; this was explored and the results support evidence that sea level rise is positive and accelerating at multiple locations around the world. The use of modern computing power here enables a return to the basic concept of a trend as a series of observations, with increasing or decreasing tendency, whose ordering would be rare based on a prior assumption of independence. The methods are therefore in line with Hendry's comments [102] made over three decades ago.

[8] i.e. the difference between the observed values and the fitted trend

Bibliography

[1] A M Best. Supplemental rating questionairre. AM Best publication http://www.ambest.com/ratings/PC_SRQ.PDF, 2014.

[2] ABI, Climate Risk Management, and Metroeconomica. Financial risks of climate change. ABI Research Paper, June 2005.

[3] O. Achieng. Actuarial modeling for insurance claim severity in motor comprehensive policy using industrial statistical distributions. International Congress of Actuaries, March 2010.

[4] AIR Worldwide. Climatological influences on hurricane activity: The air warm sst conditioned catalog. AIR Publication, 2008.

[5] Allianz. The Weather Business How companies can protect against increasing weather volatility. Allianz Global Corporate and Specialty, 2013.

[6] American International Group (AIG). Aig diversified cat bond fund (prospectus with integrated fund regulations). April 2006.

[7] AON. Aon global risk insight platform. http://www.aon.com/attachments/risk-services/grip_client_discussion_brochure.pdf, 2015.

[8] Aon Benfield. Rms u.s. hurricane medium term rate change and its impact on the renewal season. April 2013.

[9] Appleby. Guide to the Bermuda insurance market. ISBN: 1-894916-56-5, 2007.

[10] L. Arellano, P. Fernandez, R. Fonts, N. Rose, U. Nickus, H. Thies, E. Stuchlik, L. Camarero, J. Catalan, and J. Grimalt. Increasing and

decreasing trends of the atmospheric deposition of organochlorine compounds in european remote areas during the last decade. Atmos. Chem. Phys., 15:6069–6085, 2015.

[11] H. Arnold. Stochastic Parametrisation and Model Uncertainty. PhD thesis, Jesus College, Oxford, Trinity term 2013 2013.

[12] K. Arrow and G. Debreu. Existence of an equilibrium for a competitive economy. Econometrica, 22(3):265–290, 1954.

[13] Artemis. Latest catastrophe bonds and insurance-linked securities. Artemis deal directory, 2015.

[14] P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. Mathematical Finance, 9(3):203, 1999.

[15] Aviva. Home insurance policy document. aviva.com, 2015.

[16] A. Bailey. A generalised theory of credibility. Proceedings of the Casualty Actuarial Society, 32(62):13, 1945.

[17] A. Bailey. Credibility procedures laplace's generalization of bayes' rule and the combination of collateral knowledge with observed data. PCAS, 1950.

[18] M. Batty et al. Predictive modeling for life insurance. Deloitte Consulting LLP, April 2010.

[19] M. Baxter and A. Rennie. Financial Calculus. Cambridge University Press, 1996.

[20] R. Benedetti. Scoring rules for forecast verification. Mon. Wea. Rev, 138:203 – 211, 2010.

[21] S. Benjamin. Solvency and profitability in insurance. Transactions of 21st International Congress of Actuaries, 1:33–46, 1980.

[22] B. Berliner. Large risks and limits of insurability. The Geneva Papers on Risk and Insurance, 10(37):313–329, 1985.

[23] J. M. Bernardo. Expected information as expected utility. The Annals of Statistics, 7(3):686–690, 1979.

[24] C. C. Bissell. Historical perspectives - the moniac a hydromechanical analog computer of the 1950s. IEEE Control Systems Magazine, 27(1):59–64, 2007.

[25] J. Boon. Evidence of sea level acceleration at U.S. and Canadian tide stations, atlantic coast, north america. J. Coastal Res, 28(6):1437–1445, 2012.

[26] R. BORNHUETTER and R. FERGUSON. The actuary and ibnr. PCAS, 59:181–95, 1972.

[27] G. Box and N. Draper. Empirical Model Building and Response Surfaces. John Wiley and Sons Ltd, 1987.

[28] T. Breusch and A. Pagan. A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47(5):1287–1294, 1979.

[29] J. Brocker and L. Smith. Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting, 22:382, 2005.

[30] J. Brocker and L. Smith. From ensemble forecasts to predictive distribution functions. Tellus, 60A:663–678, 2008.

[31] M. J. Brockman and T. S. Wright. Statistical motor rating: making effective use of your data. Journal of the Institute of Actuaries, 119:457–543, 1992.

[32] P. Brockwell and R. Davis. Introduction to Time Series and Forecasting. Springer, 1996.

[33] P. Brockwell, R. Davis, and Y. Yang. Continuous time gaussian autoregression. Statistica Sinica, 17:63–80, 2007.

[34] T. Brown. Admissible scoring systems of continuous distributions. RAND Paper Series, August 1974.

[35] Business Insurance. Kalista Global launches El Niño cover. http://www.businessinsurance.com/article/20130131/NEWS09/130139957, January 2013.

[36] C Scullion Lloyd's of London. Personal communication with experienced Lloyd's reserving expert. March 2016.

[37] M. Carriquiry and D. Osgood. Index insurance, probabilistic climate forecasts, and production. Center for Agricultural and Rural Development Working Paper, 465(08), 2008.

[38] A. Cave. Fury over FSA role in Independent failure. Daily Telegraph, July 2001.

[39] CCRIF. Understanding CCRIF: A collection of questions and answers. CCRIF SPC, March 2015.

[40] R. E. Chandler and E. M. Scott. Statistical methods for trend detection and analyis in the environmental sciences. John Wiley and Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO198SQ, 2011.

[41] G. Chow. Analysis and control of dynamic economic processes. Princeton Research Memorandum, (126), January 1971.

[42] G. Chow. Usefulness of imperfect models for the formulation of stabilization policies. Economic Research Program (Princeton), 199, 1976.

[43] S. Christofides. Pricing of catastrophe linked securities. ASTIN Colloquium in Bergen, June 2004.

[44] S. Christofides and A. Smith. Dfa - the value of risk. Casualty Actuarial Society DFA Forum, 2001.

[45] J. Church and N. White. Sea-level rise from the late 19th to the early 21st century. Surv Geophys, 32:585–602, 2011.

[46] K. Clarke. Near term hurricane models. Karen Clarke and Company (publication), 2011.

[47] S. Cole, D. Stein, and J. Tobacman. What Is Rainfall Index Insurance Worth? A Comparison of Valuation Techniques. LSE Internal paper, 2011.

[48] K. Coughlin, E. Bellone, T. Laepple, SJewson, and K. Nzerema. A relationship between all Atlantic hurricanes and those that make landfall in the usa. Q. J. R. Meteorol. Soc., 135:371–379, 2009.

[49] S. Coutts and E. Devitt. The assessment of the financial strength of insurance companies - a generalised cash-flow model. Financial models of Insurance Solvency, Kluwer Academic Publishers, pages 1–37, 1986.

[50] S. Coutts and T. Thomas. Modelling the impact of reinsurance on financial strength. British Actuarial Journal, 3(3):583–653, 1997.

[51] S. M. Coutts and G. J. Clark. A stochastic approach to asset allocation within a general insurance company. AFIR Colloqium, 4:95–112, 1991.

[52] D. Cummins and P. Trainar. Securitization, insurance and reinsurance. SCOR Papers, 5, July 2009.

[53] D Simmons (Willis Re). Personal communication with insurance expert. 2015.

[54] P. Dailey, M. Huddlestone, S. Brown, and D. Fasking. The financial risks of climate change examining the financial implications of climate change using climate models and insurance catastrophe risk models. ABI Research Paper, 19, 2009.

[55] P. Dailey, G. Zuba, G. Ljung, I. Dima, and J. Guin. On the relationship between North Atlantic sea surface temperatures and u.s. hurricane landfall risk. American Meteorological Society, 2008.

[56] A. Damodaran. The promise and peril of real options. Stern School of Business, New York, 1999.

[57] J. Daron. Examining the Decision-Relevance of Climate Model Information for the Insurance Industry. PhD thesis, London School of Econmics and Political Science, 2011.

[58] J. Daron and D. Stainforth. Assessing pricing assumptions for weather index insurance in a changing climate. Climate Risk Management, 2014.

[59] J. Daron and D. Stainforth. On quantifying the climate of the nonautonomous Lorenz-63 model. Chaos An Interdisciplinary Journal of Nonlinear Science, April 2015.

[60] M. Davey, M. Huddelstone, A. Brookshaw, and Lighthill Risk Network. Global impact of El Niño and La Nina. Met office : Advisor, 2011.

[61] C. Daykin et al. The solvency of general insurance companies. Journal of the Institute of Actuaries, 111:279–336, 1984.

[62] C. Daykin et al. Assessing the solvency and financial strength of a general insurance company. Journal of the Institute of Actuaries, 114:227–310, 1987.

[63] C. Daykin et al. Managing uncertainty in a general insurance company. Journal of the Institute of Actuaries, 117:173–277, 1990.

[64] C. D. Daykin and G. D. Bernstein. A simulation model to examine questions of solvency in the light of asset and run off risks. ASTIN colloqium, October 1985.

[65] M. Desrosiers. How individuals purchase insurance: Going beyond expected utility theory. Casualty Actuarial Society E-Forum, 2, 2012.

[66] D. Diers. Stochastic modelling of catastrophe risks in DFA models. ASTIN Colloquium in Manchester, 5(2009):53–79, July 2008.

[67] A. Dlugolecki. The impact of changing weather patterns on property insurance. CII research report, 1994.

[68] A. Dobson. An introduction to generalized linear models, volume Texts in Statistical Science. Chapman and Hall, 2 edition, 2000.

[69] H. Du and L. Smith. Parameter estimation through ignorance. Physical Review E, 2012.

[70] G. Duffing. Erzwungene schwingungen bei veranderlicher eigenfrequenz und ihre technische bedeuting,. G Druck und Verlag von Fridr. Vieweg and Sohn, Braunschweig., 1918.

[71] Earthquake Commission. EQ Cover an insurer's guide. http://www.eqc.govt.nz, 2012.

[72] B. Efron. 1997 rietz lecture: Bootstrap methods:another look at the jackknife. The Annals of Statistics, 7(1):1–26, 1979.

[73] K. Emanuel. Increasing destructiveness of tropical cyclones over the past 30 years. Nature online, July 2005.

[74] K. Emmanuel. Increasing destructiveness of tropical cyclones over the past 30 years. Nature, 436—4, August 2005.

[75] K. Emmanuel et al. Potential economic value of seasonal hurricane forecasts. American Meteorological Society, 4:110–117, 2012.

[76] Eqecat. Rqe - eqecat's new platform. www.sbafla.com/method/portals/methodology/Meetings/2013/20130618_RQE_Platform.pdf, 2013.

[77] Ernst and Young. Risk-based global insurance capital standard. EY Perspective, 2015.

[78] Federal Emergency Management Agency. National Flood Insurance Program: Program description. Fema.gov, August 2002.

[79] Federal Insurance Office. The breadth and scope of the global reinsurance market and the critical role such market plays in supporting insurance in the United States. US Department of the Treasury publication, December 2014.

[80] G. Fellingham and A. Kottas. Parametric and nonparametric bayesian methods to model health insurance claims costs. University of California at Santa Cruz, Department of Applied Math and Statistics Technical Reports, 2007.

[81] C. Ferro. Fair scores for ensemble forecasts. Quarterly journal of the Royal Meteorological Society, 140:1917–1923, 2014.

[82] C. Ferro, D. Richardson, and A. Weigel. On the effect of ensemble size on the discrete and continuous ranked probability scores. RMetS, 15:19–24, 1998.

[83] Financial Services Authority. Icas one year on one year on. FSA Insurance Sector Briefing:, 2005.

[84] T. Fricker, C. Ferro, and D. Stephenson. Three recommendations for evaluating climate predictions. Meteorological Applications RMetS, 20:246–255, 2013.

[85] J. Friedland. Fundamentals of General Insurance Actuarial Analysis. Society of Actuaries, 2013.

[86] D. Friedman. Effective scoring rules for probabilistic forecasts. Management Science, 29(4), April 1983.

[87] M. Friedman and L. Savage. The utility analysis of choices involving risk. Journal of Political Economy, 56(4):279–304, August 1948.

[88] L. Fu and R. Moncher. Severity distributions fo GLMs: Gamma or lognormal? evidence from monte carlo simulations. Casualty Actuarial Society Discussion Paper Program, pages 149–230, 2004.

[89] K. Gabriel. Evaluation of the power of re-randomization tests, with application to weather modification experiments. Technical Report: University of Rochester USA, 81/11, June 1981.

[90] L. Gandin and A. Murphy. Equitable skill scores of categorical forecasts. Monthly Weather Review, AMS, 120:361–370, February 1992.

[91] N. Gatzert, H. Schmeiser, and D. Toplek. An analysis of pricing and basis risk for industry loss warranties. University of St Gallen (working paper), 43, June 2007.

[92] M. Gesmann. Personal communication with experienced Lloyd's data scientist. March 2016.

[93] M. Gesmann, Lloyd's of London, et al. Claims inflation - discussion document. Lloyds.com, November 2014.

[94] Gneiting and Raftery. Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association, 102(477):359–376, March 2007.

[95] M. Gnevyshev. Essential features of the 11-year solar cycle. Solar Physics, 51(1):175–183, 1977.

[96] L. Godfrey. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica, 46:1293–1302, 1978.

[97] I. J. Good. Rational decisions. Journal of the Royal Statistical Society. Series B, 14(1):107–114, 1952.

[98] K. Gray. Comparison of trend detection methods (dphil thesis). The University of Montana, May 2007.

[99] J. Greenwald. Hurricane Andrew changed the worldwide reinsurance market. Business Insurance (online media), August 2012.

[100] Guy Carpenter. Media resources (glossary). Guy Carpenter publication, 2015.

[101] T. Hall and K. Hereid. The frequency and duration of U.S hurricane droughts. Geophys. Res. Lett., 42:3482–3485, 2015.

[102] D. Hendry. Econometrics - alchemy or science. Economica, 47:387–406, 1980.

[103] S. Herrera, J. Fernandez, M. Rodriguez, and J. Gutierrez. Spatio-temporal error growth in the multi-scale Lorenz' 96 model. Nonlin. Processes Geophys, 17(329–337), 2010.

[104] Hibbert et al. A stochastic asset model and calibration for long-term financial planning purposes. Barrie and Hibbert publication, 2001.

[105] A. Hitchcox, I. Hinder, A. Kaufman, T. Maynard, A. Smith, and M. White. Assessment of target capital for general insurance firms. Institute of Actuaries Sessional meeting paper, November 2006.

[106] Hooker et al. Risk based capital in general insurance. British Actuarial Journal, 2(2):265–323, 1996.

[107] H. Hotelling and M. Pabst. Rank correlation and tests of significance involving no assumption of normality. Jstor: Annals of mathematical statistics, 1936.

[108] J. Houston and R. Dean. Sea-level accelerations based on U.S. tide gauges and extensions of previous global-gauge analyses. J. Coastal Res, 27(3):409–417, 2011.

[109] J. Houston and R. Dean. Effects of sea-level decadal variability on acceleration and trend difference. J. Coastal Res.,, 29(5):1062–1072, 2013.

[110] J. Hull. (Chapter 28) Options Future and Other Derivatives. Prentice Hall, 5 edition, 2003.

[111] Info Please. Coastline of the United States. Infoplease.com ipa, 2007.

[112] Insurance Networking News. Top 50 global reinsurers for 2012. http://www.insurancenetworking.com/news/core-systems/top-50-global-reinsurers-for-2012-32900-1.html, 2012.

[113] International Actuarial Association. Comments on catastrophe provisions. http://www.actuaries.org/LIBRARY/Submissions/IASC_Insurance_Issues/Catastrophe_Provisions.pdf, May 2000.

[114] IPCC. Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2013.

[115] IPCC. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. 151pp. IPCC,Geneva, Switzerland, 2014.

[116] A. Ishaq. Reinsuring for catastrophes through industry loss warranties – a practical approach. Casualty Actuarial Society Forum, Spring 2005.

[117] T. Jagger, J. Elsner, and M. Saunders. Forecasting us insured hurricane losses. Climate Extremes and Society - CUP, pages 189–208, 2008.

[118] S. Jarvis, F. Southall, and E. Varnell. Modern valuation techniques. The Staple Inn Actuarial Society, February 2001.

[119] S. Jewson, E. Bellone, T. Laepple, K. Nzerem, S. Khare, M. L. A. O'Shay, J. Penzer, and K. Coughlin. 5 year prediction of the number of hurricanes which make u.s. landfall. 2009.

[120] S. Jewson, T. Laepple, K. Nzerem, and J. Penzer. Predicting landfalling hurricane numbers from sea surface temperature: theoretical comparisons of direct and indirect approaches. arXiv:physics/0701176, 2008.

[121] J. Johnston. Econometric methods. McGraw Hill, 2 edition, 1972.

[122] I. Jolliffe and D. Stephenson. Proper scores for probability forecasts can never be equitable. Monthly Weather Review, AMS, 136:1505–1510, April 2008.

[123] K. Judd and L. Smith. Indistinguishable states ii the imperfect model scenario. Physica D, 196:224–242, 2004.

[124] R. Katz. Stochastic modeling of hurricane damage. Journal of Applied Meteorology, 41(7):754–762, 2002.

[125] P. Klotzbach. Atlantic basin seasonal hurricane forecasts. Risk Frontiers, 12(4), March 2013.

[126] P. Klotzbach and W. Gray. Twenty-five years of atlantic basin seasonal hurricane forecasts. Geophysical research letters, 36, May 2009.

[127] S. Klugman, T. Rhodes, M. Purushotham, and S. Gill. Credibility theory practices. Society of Actuaries, December 2009.

[128] R. Kreps. Reinsurer risk loads from marginal surplus requirements. PCAS, LXXVII:196, 1990.

[129] F. Kwasniok. Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling. Phil. Trans. R. Soc. A, 370:1061–1086, February 2012.

[130] D. Laidler. Phillips in retrospect (a review essay). http://economics.uwo.ca/people/laidler_docs/phillips.pdf, 2000.

[131] H. Larson. Introduction to Probability Theory and Statistical Inference. Wiley International. John Wiley and Sons Ltd, 3 edition, 1982.

[132] Lloyd's Market Association. Catastrophe modelling guidance for non-catastrophe modellers. www.lmalloyds.com, June 2013.

[133] Lloyd's of London. Glossary https://www.lloyds.com/common/help/glossary.

[134] Lloyd's of London. New central fund byelaw (no. 23). 1996.

[135] Lloyd's of London. Syndicate accounting byelaw (no. 8). 2005.

[136] Lloyd's of London. Adapt or bust. Lloyds 360 Insight Report, 2006.

[137] Lloyd's of London. Ica minimum standards and guidance. 2010.

[138] Lloyd's of London. Annual report and accounts. 2011.

[139] Lloyd's of London. Forecasting risk: The value of long-range forecasting for the insurance industry. Lloyds.com, One Lime Street, London, 2011.

[140] Lloyd's of London. Solvency II - Model Validation Guidance. Lloyds.com, 2012.

[141] Lloyd's of London. Statistics relating to lloyd's. https://www.lloyds.com/the-market/tools-and-resources/resources/statistics-relating-to-lloyds/abou 2013.

[142] Lloyd's of London. Catastrophe modelling and climate change. Lloyds.com, 2015.

[143] Lloyd's of London. Glossary https://www.lloyds.com/common/help/glossary. Lloyds.com, 2015.

[144] Lloyd's of London. Lloyd's Minimum Standards MS1.1 - Underwriting strategy and planning. www.lloyds.com, October 2015.

[145] Lloyd's of London. Lloyds risk code scheme. Market Bulletin Y4886, April 2015.

[146] Lloyds of London. Realistic disaster scenarios. Scenario Specification - Lloyd's publication, 2015.

[147] Lloyd's of London. What is lloyd's: The lloyd's market. https://www.lloyds.com/lloyds/about-us/what-is-lloyds/the-lloyds-market, 2015.

[148] Lloyd's of London and Met Office. Hurricanes and long term variability. Lloyd's Emerging Risks Report, May 2012.

[149] Lloyds of London and RAND. Litigation and business: Transatlantic trends. Lloyds 360 Insight Report, 2008.

[150] London Market Group and Boston Consulting Group. London matters: The competitive position of the london insurance market. norisco.com, 2014.

[151] M. Lonfat. Rms medium term perspective on hurricane activity. Florida Commission on Hurricane Loss Projection Methodology Workshop, July 2006.

[152] E. Lorenz. Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20:130–141, March 1963.

[153] E. Lorenz. Irregularity: a fundamental property of the atmosphere. Tellus, 36(A):98–110, 1984.

[154] E. Lorenz. Atmospheric models as dynamical systems. Perspectives in Nonlinear Dynamics, World Scientific Publishing Co., 1-17, 1986.

[155] E. Lorenz. Predictability - a problem partly solved. Shinfield Park, Reading, United Kingdom, 1996. ECMWF.

[156] S. Lowe and J. Stanard. An integrated dynamic financial analysis and decision support system for a property catastrophe reinsurer. Astin Bulletin, 27(2):339–371, 1997.

[157] H. Luo, J. Skees, and M. Marchant. Weather information and the potential for inter-temporal adverse selection in crop insurance. Review of Agricultural Economics, 16:441–451, 1994.

[158] A. Luoma and A. Puustelli. Hedging equity-linked life insurance contracts with american-style options in bayesian framework. AFIR Colloquium, Munich, 2009.

[159] Marsh. Catastrophe modeling: Why all the fuss? https://www.marsh.com/us/insights/catastrophe-modeling.html, 2015.

[160] A. Mayerson. A bayesian view of credibility. PCAS, 51:85, 1964.

[161] T. Maynard. Hurricanes in the North Atlantic, should insurance pricing be based on long-term averages? Astin Colloquia, July 2008.

[162] T. Maynard. Modelling insurance markets: value of seasonal weather forecasts. Presentation at CASS business school: LSE.ac.uk: CATS talks, March 2011.

[163] T. Maynard. Potential uses of decadal forecasts: an insurance perspective www.rmets.org/sites/default/files/pdf/presentation/20110216-maynard.pdf, 2011.

[164] T. Maynard and N. Ranger. What role for 'long-term insurance' in adaptation? Insurance Economics, January 2013.

[165] J. McCarthy. Measures of the value of information. Proceedings of the National Academy of Science, 42:654 – 655, September 1956.

[166] P. McSharry and L. Smith. Consistent nonlinear dynamics: identifying model inadequacy. Physica D: Nonlinear Phenomena, 192(1-2):1–22, 2004.

[167] M. Merriman. Method of Least Squares. John Wiley and Sons Ltd, 15 Astor Place, 1884.

[168] R. Merton. Applications of option-pricing theory: Twenty-five years later. The American Economic Review, 88(3):323–349, 1998.

[169] Met Office (UK). Met Office climate prediction model: HadCM3. http://www.metoffice.gov.uk/research/modelling-systems/unified-model/climate-models September 2013.

[170] N. Michaelides et al. The premium rating of commerical risks (working party paper). Institute and Faculty of Actuaries: General Insurance Convention, October 1997.

[171] E. Michel-Kerjan, S. Hochrainer-Stigler, H. Kunreuther, J. Linnerooth-Bayer, R. Mechler, R. Muir-Wood, N. Ranger, P. Vaziri, and M. Young. Catastrophe risk models for evaluating disaster risk reduction investments in developing countries. Wharton School - working paper, March 2012.

[172] S. Mikkonen, M. Laine, H. Makela, H. Gregow, H. Tuomenvirta, M. Lahtinen, and A. Laaksonen. Trends in the average temperature in Finland 1847-2013. Stch Environ Res Risk Assess, 29:1521–1529, 2015.

[173] E. Mills. Insurance in a climate of change. Science, 309:1040–1044, 2005.

[174] J. Milnor. On the concept of attractor. Communications in Mathematical Physics, 99:177–195, 1985.

[175] Munich Re. How much risk can the world take? Group Annual Report, 2001.

[176] Munich Re. http://www.munichre-foundation.org/home/Microinsurance.html. Munich Re Foundation, 2015.

[177] Munich Re America. A basic guide to Facultative and Treaty reinsurance. Munichre.com, 2010.

[178] A. Murphy. The Finley Affair: A signal event in the history of forecast verification. Forecasting, AMS, 11(1), March 1992.

[179] N Ralph Lloyd's of London. Personal communication with experienced Lloyd's underwriter. April 2015.

[180] NAIC. http://www.naic.org/index_about.htm. National Association of Insurance Commissioners - Web site, 2015.

[181] National Centre for Atmospheric Research (NCAR. About ncar. http://ncar.ucar.edu/about-ncar, 2015.

[182] National Hurricane Centre. Atlantic marine forecasts. http://www.nhc.noaa.gov, 2015.

[183] National Hurricane Centre. Saffir-simpson hurricane wind scale http://www.nhc.noaa.gov/aboutsshws.php, 2015.

[184] E. Nelson and K. Nikolov. Uk inflation in the 1970s and 1980s: the role of output gap mismeasurement. Bank of England's working paper series http://www.bankofengland.co.uk/archive/Documents/historicpubs/workingpapers/2001/wp148.pdf 2001.

[185] Nephila. Description of weather risk indices http://nephila.com/industry/weather-risk.aspx, June 2015.

[186] E. Neumayer and F. Barthel. Normalizing economic loss from natural disasters: a global analysis. Munich Re Programme, 2010.

[187] Y. B. W. Nornadiah Mohd Razali. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1):21–33, 2011.

[188] OECD. Global insurance market trends. OECD Publication, 2014.

[189] D. Orrell and L. A. Smith. Visualizing bifurcations in high dimensional systems: The spectral bifurcation diagram. International Journal of Bifurcation and Chaos, 13:3015–3027, 2003.

[190] D. Osgood, M. McLaurin, M. Carriquiry, A. Mishra, F. Fiondella, J. Hansen, N. Peterson, and N. Ward. Designing weather insurance contracts for

farmers in malawi, tanzania, and kenya, final report to the commodity risk management group ard. World Bank. International Research Institute for Climate and Society (IRI), 2007.

[191] D. Osgood, P. Suarez, J. Hansen, M. Carriquiry, and A. Mishra. Integrating seasonal forecasts and insurance for adaptation among subsistence farmers: The case of malawi. World Bank Policy Research Working Paper 4651., June 2008.

[192] N. Parikh, M. Pencina, T. Wang, K. Lanier, C. Fox, R. D'Agostino, and R. Vasan. Increasing trends in incidence of overweight and obesity over 5 decades. The American Journal of Medicine, 120:242–250, 2007.

[193] G. Patrik. Estimating casualty insurance loss amount distributions. Proceedings of the Casualty Actuarial Society, 68:57–109, 1980.

[194] T. Pentikainen and J. Rantala. Solvency of insurers and equalisation reserves. Insurance Publishing Company, Helsinki, 1982.

[195] T. Pentikainen and J. Rantala. Run-off risk as part of claims fluctuations. ASTIN Bulletin, 16(2):113–47, 1986.

[196] F. Pereira. Practical "modern" bayesian statistics in actuarial science. GIRO (UK Actuarial Profession), 1999.

[197] Permanent Service for Mean Sea Level. Recieved from psmsl.org in June. 2015.

[198] D. P. Pesonen. Practical Risk Theory for Actuaries. Monographs on Statistics and Applied Probability. Chapman and Hall, 53 edition, (1994).

[199] S. Philbrick. A practical guide to the single parameter pareto distribution. Proceedings of the Casualty Actuarial Society, LXXII(137 and 138):44–84, 1985.

[200] R. Pielke and C. Landsea. Normalized hurricane damages in the United States 1925-1995. Weather and Forecasting, pages 621–631, September 1998.

[201] Practical Law Company. Returning capital to shareholders, the rise and rise of the share buyback? Website, May 1999.

[202] Price Waterhouse Coopers. Success through excess: How property and casualty insurers are boosting proftis by entering the excess and surplus market. PWC publication; FS Viewpoint, September 2012.

[203] Prudential Regulation Authority. The Prudential Regulation Authority's approach to insurance supervision. Bank of England publication, April 2013.

[204] Prudential Regulation Authority. The impact of climate change on the uk insurance sector. Bank of England publication, Sep 2015.

[205] A. PUUSTELLI. Bayesian modelling of financial guarantee insurance. Insurance: Mathematics and Economics, 43:245–254, 2008.

[206] S. Rahmstorf and M. Vermeer. Discussion of: Houston, J.R. and Dean, R.G., 2011. sea-level acceleration based on U.S. tide gauges and extensions of previous global-gauge analyses, journal of coastal research, 27(3), 409–417. J. Coastal Res, 27(4):784–787, 2011.

[207] R. Ramanathan. Introductory econometrics with applications. Dryden Press, 4 edition, 1998.

[208] N. Ranger and F. Niehorster. Deep uncertainty in long-term hurricane risk: scenario generation and implications for future climate experiments. Munich Re Programme Technical Papers, (8), July 2011.

[209] R. REISS and M. TOMAS. A new class of bayesian estimator in paretian excess-of-loss reinsurance. ASTIN Bulletin, 29:339–349.

[210] Reserve Bank of New Zealand. A.w.h.(bill) phillips mbe and the moniac. http://www.rbnz.govt.nz/research_and_publications/fact_sheets_and_guides/3121411.pdf.

[211] Risk Management Solutions. A guide to catastrophe modelling. The Review (Informa), 2008.

[212] Risk Management Solutions and R. Muir-Wood. Updates in rms modeled hurricane risk costs. Evidence given to NAIC committee, March 2011.

[213] R. Roberts. Did anyone learn anything from the Equitable Life? Institute of Contemporary British History (ICBH) Kings College London, September 2012.

[214] A. Robertson, A. Barnston, J. Hansen, U. Lall, V. Moron, D. Osgood, and L. Sun. On the potential value of seasonal climate forecasts for index insurance. International Research Institute for Climate and Society (IRI): Climate and Society, 2, 2010.

[215] M. Ross. Credit derivatives and insurance - a world apart? Norton Rose Publication, 2001.

[216] M. Roulston and L. Smith. Evaluating probabilistic forecasts using information theory. Notes and Correspondence: American Meteorological Society, 130:1653–1660, 2002.

[217] D. Ruppert et al. Semiparametric regression. Cambridge Series in Statistical and Probabilistic Mathematics, 1 edition, July 2003.

[218] Ryan et al. Financial condition assessment. British Actuarial Journal, 7(4):519–584, 2001.

[219] J. Ryan. An application of model office techniques to the solvency question. Transactions of 21st International Congress of Actuaries, 1:403–410, 1980.

[220] J. Ryan. Application of simulation techniques to solvency testing for a non-life office. Transactions of 22nd International Congress of Actuaries, 3:269–277, 1984.

[221] M. Saunders and A. Lea. Tropical storm risk seasonal hurricane forecasts. http://www.tropicalstormrisk.com/about, September 2015.

[222] D. SCOLLNIK. An introduction to markov chain monte carlo methods and their actuarial applications. Casualty Actuarial Society, 1996.

[223] SCOR Global. Seasonal hurricane forecast skill and relevance to the (re)insurance industry. SCOR Technical Newletter, 2014.

[224] R. Selten. Axiomatic characterization of the quadratic scoring rule. Experimental Economics, 1:43–62, 1998.

[225] P. Sen. Estimates of regression coefficient based on kendall's tau. J. Am. Stat. Ass., 63(324):1379–1389, December 1968.

[226] J. Seo and O. Mahul. The impact of climate change on catastrophe risk model: Implications for catastrophe risk markets in developing countries. World Bank Policy Research Working Paper 4959., 2009.

[227] S. S. Shapiro and M. B. Wilk. Analysis of variance test for normality (complete samples). Biometrika, 52:591–611, 1965.

[228] SILSO World Data Center. The international sunspot number. International Sunspot Number Monthly Bulletin and online catalogue, 2015.

[229] B. Silverman. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall, 1986.

[230] D. Simmons and M. Saunders. Using hurricane forecasts to adjust peril model loss probabilities. ReMetrics Review (Benfield), October 2005.

[231] T. Sinha. Strategic expansion for insurance companies: Quantitative methods using real options. ARCH, Society of Actuaries, 1, 2003.

[232] J. Skees, P. Hazell, and M. Miranda. New approaches to crop yield insurance in developing countries. International Food Policy Research Institute EPTD DISCUSSION PAPER, (55), 1999.

[233] J. Skees, P. Hazell, and M. Miranda. New approaches to public / private crop yield insurance.
EPTD Discussion Paper No. 55 International Food Policy Research Institute, 1999.

[234] A. Smith. How actuaries can use financial economics. British Actuarial Journal, 2:1057–1174, 1996.

[235] A. Smith and F. Southall. A stochastic asset model for fair values in pensions and insurance. GIRO and Casualty Actuarial Society, October 2001.

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

[237] L. Smith. Accountability and error in ensemble forecasting. ECMWF Seminar on Predictability, Vol 1:351–368, 1995.

[238] L. A. Smith. The maintenance of uncertainty. Proc International School of Physics "Enrico Fermi", CXXXIII:177–246, 1997.

[239] L. A. Smith. Disentangling uncertainty and error: On the predictability of nonlinear systems. In Nonlinear Dynamics and Statistics, pages 31 – 64. Birkhauser, 2000.

[240] L. A. Smith, E. B. Suckling, E. L. Thompson, T. Maynard, and H. Du. Towards improving the framework for probabilistic forecast evaluation. Springer Climatic Change, 2015.

[241] K. Soar, S. Cooper, J. Heuvels, J. Ewen, and L. Fernandes. A legal guide to industry loss warranty contracts. Ince and Co publication: Insurance and Reinsurance, 2013.

[242] A. Solow and L. Moore. Testing for a trend in a partially incomplete hurricane record. Journal of Climate, 13, 2000.

[243] H. Song et al. Evaluation of precipitation simulated by seven SCMs against the ARM observations at the SGP site. American Meteorological Society, 26:5467–5492, August 2013.

[244] Spiers, Maynard, et al. Paper on with-profits investment strategy. Life convention working paper, 2004.

[245] E. Suckling and L. Smith. An evaluation of decadal probability forecasts from state-of-the-art climate models. Journal of Climate American Meteorological Society, 26, 2013.

[246] R. Sullivan. Financial models useful but limited. Financial Times, April 2011.

[247] Swiss Re. Securitization—new opportunities for insurers and investors. Sigma, July 2006.

[248] Swiss Re. Natural catastrophes and man-made disasters in 2014: convective and winter storms generate most losses, 2015.

[249] N. N. Taleb. Fooled by randomness. Random House, third edition edition, 2005.

[250] G. Taylor. A simple model of insurance market dynamics. North American Actuarial Journal, 12(3), 2012.

[251] The economist. Independent Insurance: Not my fault. nobody wants to take the blame for the demise of Independent Insurance. Economist print edition, June 2001.

[252] The European Parliament. Clause 64. DIRECTIVE 2009/138/EC, November 2009.

[253] The Insurance Insider. It's never really payback time. Insurance Insider (trade press), January 2014.

[254] The Prince's Rainforests Project. An emergency package for tropical foreasts. The Prince's Charities, March 2009.

[255] The Telegraph. How asbestos brought Lloyd's of London to its knees in the 90s. The Telegraph, April 2011.

[256] L. Thomas and K. Martin. The importance of analysis method for breeding bird survey population trend estimates. Conservation Biology, 10:479–490, 1996.

[257] Tropical Storm Risk. Beta products: Probabilistic forecast u.s. market insured loss. TSR Website http://www.tropicalstormrisk.com/business/, 2015.

[258] UK Government. Life assurance (gambling) act. 14 Geo 3, Chapter 48, 1774.

[259] UK Government. Marine insurance act. 6 Edw 7, Chapter 41, 1906.

[260] UK Met Office. North Atlantic tropical storm seasonal forecast 2015. http://www.metoffice.gov.uk/weather/tropicalcyclone/seasonal/northatlantic2015, 2015.

[261] UK Met Office. Storm tracker. http://www.metoffice.gov.uk/public/weather/storm-tracker/, 2015.

[262] R. Verrall. A bayesian generalized linear model for the bornhuetter-ferguson method of claims reserving. Actuarial Research Paper No 139, 2001.

[263] H. S. Visser, S. Dangendorf, and A. C. Petersen. A review of trend models applied to sea level data with reference to the 'acceleration-deceleration' debate. J Geophys Res Oceans, 120(3873-3895), 2015.

[264] J. Waters and T. Sabbatelli. We're still all wondering – where have all the hurricanes gone? RMS Blog http://www.rms.com/blog/tag/medium-term-rate/, 2015.

[265] P. Watson. Is there evidence yet of acceleration in mean sea level rise around mainland Australia? J. Coastal Res, 27(2):368–377, 2011.

[266] P. Webster, G. Holland, J. Curry, and H. Chang. Changes in tropical cyclone number, duration, and intensity in a warming environment. Science, 309:1844–1846, 2005.

[267] M. Wedel. Hedging catastrophe risk using industry loss warranties. AIR Worldwide - Website, 2012.

[268] E. Wheatcroft. Improving predictability of the future by grasping probability less tightly. Draft Thesis, LSE, 2015.

[269] R. Wilcox. African Risk Capacity (ARC) Briefing Book. DFID and the Rockefeller Foundation, 2015.

[270] A. D. Wilkie. A stochastic investment model for actuarial use. Transactions of the Faculty of Actuaries, 39:341–381, 1986.

[271] D. Wilks. Effects of stochastic parametrizations in the Lorenz' 96 system. Q.J.R Meteorol Soc, 131(606):389–407, January 2005.

[272] D. Wilks. Effects of stochastic parametrization on conceptual climate models. Phil Trans R. Soc, 366:2477–2490, April 2008.

[273] D. Wilks. 'superparametrization' and statistical emulation in the lorenz '96 system. Q.J.R Meteorol Soc, 138:1379–1387, 2012.

[274] R. Winkler and A. Murphy. Good probability assessors. Journal of applied meteorology, 7:751 – 758, October 1968.

[275] P. Woodworth, M. Menendez, and W. Gehrels. Evidence for century-timescale acceleration in mean sea levels and for recent changes in extreme sea levels. Surv Geophys, 32:603–618, 2010.

[276] WRF Committees. The weather research and forecasting model. http://www.wrf-model.org/index.php, 2015.

[277] D. Wright, I. Owadally, and F. Zhou. Insurance pricing cycles: An artificial agent-based analysis. Cass Business School, 2012.

[278] XL. The ups and downs of the insurance market cycle. http://resources.xlgroup.com/segment/xli/INSight_Am/docs /Insurance Market Cycle.pdf.

[279] S. Yazici. The Turkish catastrophe insurance pool (TCIP) and the compulsory earthquake insurance scheme. Catastrophic Risks and Insurance;Policy Issues in Insurance, (8), 2005.

[280] C. Zhou. Path-dependent option valuation when the underling path is discontinuous. Federal Reserve Board, March 1997.