Report Cover Page

ACERA Project
2006 Round 1, Project 02
Title
Assessment of strategies for evaluating extreme risks
Author(s) / Address (es)
James Franklin, School of Mathematics and Statistics, University of New South Wales
Material Type and Status (Internal draft, Final Technical or Project report, Manuscript, Manual, Software)
Project final report
Summary
The aim of this project was to review literature on the analysis of extreme risks, focusing primarily on ideas and developments in a range of disciplines that may be useful in the data poor environments typical of biosecurity risk assessments. The report begins by outlining several case studies with varying levels of data, examining the role for extreme event risk analysis. The case studies include BA's analysis of fire blight and New Zealand apples, bank operational risk and several technical failures. The report then surveys recent developments in methods relevant to evaluating extreme risks and evaluates their properties. These include methods for fraud detection in banks, formal extreme value theory, Bayesian approaches, qualitative reasoning, and adversary and advocacy models. The document includes a supplementary report as an appendix, providing an overview of the quantification of bank operational risks. The report concludes that evaluations of extreme risks should be supported by quantitative analysis, even in data poor environments, but that the most effective strategies give the final word to qualitative reasoning. Evaluations are most effective when stakeholders and experts are able to examine and test data, models and reasoning, and when disagreements are heard by an independent arbitrator or panel.

ACERA Use only	Received By:	Date:
ACERA Use only	ACERA / AMSI SAC Approval:	Date:
ACERA Use only	ACERA / AMSI SAC Approval:	Date:

Assessment of Strategies for Evaluating Extreme Risks

ACERA Project No 0602

James Franklin & Scott Sisson

School of Mathematics and Statistics,

University of New South Wales

*(*with a report on Quantifying Bank Operational Risk *by Gareth Peters and Venta Terauds)*

Final Report; March 2007

Acknowledgements

This report is a product of the Australian Centre of Excellence for Risk Analysis (ACERA). In preparing this report, the authors acknowledge the financial and other support provided by the Department of Agriculture, Fisheries and Forestry (DAFF), the University of Melbourne, Australian Mathematical Sciences Institute (AMSI) and Australian Research Centre for Urban Ecology (ARCUE).

Disclaimer

This report has been prepared by consultants for the Australian Centre of Excellence for Risk Analysis (ACERA) and the views expressed do not necessarily reflect those of ACERA. ACERA cannot guarantee the accuracy of the report, and does not accept liability for any loss or damage incurred as a result of relying on its accuracy.

Table of contents

1. Executive Summary

It is in the nature of risk for extreme events that there is no or very little directly relevant data, so expert opinion must be relied on heavily. But expert opinion must be as fully informed as possible – by the data that is available, by other experts, by reasoned opinions of stakeholders, and by the use of commonsense reasoning applied to the diverse reasons put "on the table". We survey a variety of case studies and a number of quantitative and non-quantitative methods that show promise for improving extreme risk analysis. We argue that an "advocacy model" similar to that used in the Basel II compliance regime for bank operational risks and Biosecurity Australia's Import Risk Assessments is ideal for permitting the diversity of relevant evidence to be presented and soundly evaluated. We recommend that the process be enhanced in four ways – by better education of the risk evaluators in certain statistical methods (extreme value theory, Bayesian methods of combining expert opinion with data, and robustness methods such as InfoGap Theory); by better education of statisticians in non-numerical methods including legal-style advocacy and causal modeling; by education of all parties in the psychological findings on expert judgement; and by the use of independent facilitators such as consultants to mediate between the regulator/evaluator and the client/stakeholder.

2. Introduction: Extreme risks and the poverty of data

A risk is called "extreme" when it concerns an outcome that has happened very rarely or never. Normally "extreme" is used of events that are of high (negative) consequence as well as low probability, but since the methods to be discussed in this report deal with probabilities rather than consequences, the emphasis is on events of very low probability. Such an event is at the edge of or outside the range of what has occurred, possibly far outside. There is therefore very little or no directly relevant data, and any data set there may be is too small to be reliably representative.

The probability of an extreme event must therefore be evaluated by putting together disparate sources of relevant evidence, none of which are reliable in isolation. The sources of evidence include what data there is, how far the event of interest is from the data, the opinion of experts (possibly in diverse disciplines), arguments from analogy (that is, from events whose similarity to the event in question is debatable), specialist scientific causal knowledge relevant to the case, and commonsense knowledge of "how the world works". There is no established methodology either for eliciting the probabilities arising from these sources of knowledge or for combining them once elicited. But the reasons for the difficulty of reaching a correct answer are the same as the reasons why it is important to succeed – because of the paucity of data, neglecting any source of evidence or any method of interpreting it will lead to the misevaluation of extreme risks and hence to avoidable disasters.

In this report, we survey first a number of cases and a variety of methods applicable to extreme risk analysis. Although they can be read separately, we believe that taken together they suggest an overall approach to extreme risk analysis that we call the "advocacy model". In brief, the model envisages a tribunal that reaches a final decision after submission of evidence by stakeholders with different interests, evidence which may in principle be of any sort (quantitative, qualitative, or informed opinion). The model is inspired by the well-know adversary model of Anglo-American law where opposing counsel argue before a neutral judge and jury, but is more co-operative and more amenable to technical evidence. Our case studies include some in which a method recognisably like this has been applied. After our survey of methods, we set out the advocacy model in more detail and explain how the methods fit into it. We conclude by making recommendations on how to improve current practice and on some further research directions.

3. Case Studies

We next give an overview of several cases of extreme risk assessment in order to ground our conclusions on the relevance of the methods to be described later.

3.1 Biosecurity Australia analyses of fire blight risk from import of NZ apples

In response to requests from New Zealand to permit the import of apples to Australia, the Biosecurity Australia (or its predecessor) produced major analyses of the risk that importing apples from New Zealand would introduce disease to Australia. (AQIS, 1998; Biosecurity Australia, 2006). (Although the reports deal with many other pests we will confine attention to fire blight). Both reports were developed over considerable time periods and were informed by submissions from stakeholders. There was a formal process similar to the "advocacy model" to be described in section 6 below: BA compiles a draft report using its own and contracted expertise to evaluate information from public and scientific documents and from stakeholders, then after a period for stakeholder comment on the draft, a final report is issued that must show how the comments have been addressed.

The main stakeholders were strongly motivated by opposite concerns – the New Zealanders were concerned that the likelihood of disease had been over-estimated and representatives of the Australian apple industry were concerned that it had been under-estimated. Both sides presented detailed scientific analyses and the final report responded to many of the detailed arguments raised. The conclusions of the two reports in terms of risk were substantially similar, though the recommendations were different: the 1998 report recommended against import while the 2006 recommended for it, but the 1998 report concerned a New Zealand proposal to import fruit without special measures to guard against pests, while the recommendation in 2006 would permit import only after adequate onerous and expensive inspection and disinfection measures.

The analyses are considerably more complex, especially in evaluating quantitative probabilities, than most of Biosecurity Australia's Import Risk Analyses. That makes them more robust in the politically charged atmosphere of apple import controversies, which has included grilling of AQIS's Executive Director by a Senate committee on the possible motives of New Zealand scientists in looking for fire blight in Australian botanic gardens (Senate Hansard, 1997), direct recommendations by the Senate committee on how AQIS should conduct its risk assessment (Senate, 2005) including an allegation of a "methodological leaning towards qualitative rather than quantitative analysis", and comment by the New Zealand Minister for Agriculture that "the concept of honest science has no meaning [in Australia]". (Knight, 2005) In addition Australia needs to comply with the guidelines of the International Plant Protection Convention, and BA's scientists naturally desire to show to themselves and to the international scientific community that their results are not swayed by political pressures. Such political pressures are stressful for all concerned, in much the same way as it is stressful to be cross-examined in court by an experienced QC. From the point of view of the advocacy model, however, that is not necessarily a bad thing. Pressures from different directions are integral to the advocacy situation and (at least if the pressures are reasonably balanced) can encourage care and transparency in the risk evaluation process.

The analyses looked at the possible chains of causes by which fire blight from New Zealand might become established in Australia through the commercial import of apples (as opposed to illegal import such as by tourists). A particularly difficult point in the analysis, and the one most relevant to the study of extreme risks, came in trying to evaluate the probability of what was believed to be the most unlikely event in the most likely chain, the transfer of the pest from a discarded apple to

an Australian plant. The most likely scenario for transfer, it was believed, would be something like an apple core discarded near one of the few plants which can be affected by the pest, such as a cotoneaster, and being transferred to the plant by direct contact or by insects. There are a great number of imponderables in such a scenario – including specifying the scenario with any exactitude, knowing the possible mechanisms of transfer, dealing with the different possible levels of fire blight infection of the apple core, and noting the seasonal differences in the probability of transfer. Since the probability of transfer (that is, the probability of a potential host becoming infected, given that an apple core with fire blight is discarded near it) is believed to be of the order of one in a million, experimentation is not feasible – it would take several million experiments to achieve any moderately reliable estimate of the probability. The analyses therefore relied on an expert review of marginally relevant evidence (Roberts et al, 1998), a paper which emphasises the lack of any experimental confirmation of any possible modes of spread of the fire blight bacterium. The paper said merely "Trials where a contaminated fruit was suspended adjacent to open flowers produced no infection in those flowers. With much uncertainty, P(5) [the probability that the bacterium present in a fruit near a host is transferred to the host] is estimated as between 0.001 and 0.00001 with a median value of .0001." (Roberts et al, 1998, p. 25) That is a very wide range of probabilities. Biosecurity Australia's conclusion was that the probability in such a case of the bacterium being transferred to the host and then establishing on the host was "in the range of Uniform (0, 10-6)", that is, somewhere between zero and one in a million. (Biosecurity Australia, 1996, p. 90). That is also a wide range of probabilities and one poorly based on data – but inevitably so, given the low probabilities involved.

The analysis of chains of causes is a relevant topic, both in biosecurity work and in, for example, air accident investigations, where a chain of errors is typical in the causation of disasters. There are low-probability events in the chain, but the chain is repeated many times. Breaking the chain into many units for analysis of probabilities is obviously desirable, but choosing the correct unit, especially for the "choke point" of lowest probability, is difficult, and there are difficulties with determining correlations between errors at different points in the chain. A more detailed survey has not been attempted in this report, but is certainly desirable.

3.2 Bank operational risk in the Basel II compliance regime

Bank operational risk ("oprisk") is more than a single case study. It is a rapidly developing area in which massive resources have recently been committed to the study of, in part, the quantification of extreme risks. We therefore append a more complete report on the field of bank oprisk, especially as it applies to extreme risk evaluation. Here we provide a very short introduction to the area and what can be learned from it.

In banking, a powerful international body, the Committee on Banking Supervision of the Bank for International Settlements in Basel, enforces the Basel II standards. (Bank for International Settlements, 2004; Marrison, 2002, ch. 23). Banks are regulated in various ways, but from the point of view of risk the most important target of regulation is banks' reserves against risk. The nature of a bank is to take in funds, then lend most of them out for profit while reserving some against risks. The risks are varied: of default by creditors, of movements in exchange rates, of the disappearance or devaluation of assets, and "operational risk", a grab-bag of unusual and extreme events ranging from massive internal fraud to tsunamis, typing errors in crucial places, incompetent CEOs and major technological change. We concentrate here on operational risk, since credit and market risk are rich in data and statistically tractable, whereas operational risk includes the extreme risks of the sort that are the focus of this report.

Basel II permits larger banks to evaluate their risks using any internal models and sophisticated statistical technology they wish, provided they disclose them to the (national) regulator (in Australia, the Australian Prudential Regulation Authority, APRA) and the regulator approves. That naturally allows free rein for statistical expertise, both on the side of banks and on the side of the regulator. It promises to improve risk evaluation greatly, by enforcing best practice in data collection and statistical methodology.

Operational risk ("the risk of direct or indirect loss resulting from inadequate or failed internal processes, people and systems or from external events") is a classification covering a great variety of risks, mostly of a rare and/or extreme nature. (Bank for International Settlements, 2002; King, 2001) They include the risks that may cause complete collapse of a bank. Merely classifying the kinds of operational risk and establishing who has expertise in those various areas is a substantial intellectual exercise. A table of some of the kinds of operational and related risks is given at the end of Appendix 1, at the end of the supplementary report on Quantifying Bank Operational Risk.

It is widely agreed that there are unusual difficulties in the way of a bank's quantifying its operational risks adequately, or even of getting a "ballpark" figure for many of them. Availability of data is a major challenge. Internal frauds, for example, are rarely reported publicly by individual banks unless they are catastrophic. Therefore an individual bank has very little data on past events of the sort it fears may impact on it severely in the future. It is not usual for individual banks to hold data on public events like tsunamis; banks are not in the business of environmental modelling. (Rosen and Corregia, 2004)

It is also generally agreed that the diversity of operational risks creates methodological difficulties both in quantifying the individual risks and in estimating their interactions. Given that the (downside) tails of the distribution of events are crucial and that there is little data on tail events, it is necessary to avoid assuming that the events follow a standard distribution (such as the normal distribution) even if that fits the middle range of events well. Extreme value theory is the study of the extrapolation of the tails of distributions beyond the range of existing data, and is a specialised topic in statistics that still needs further study and wider dissemination of what is already known. (Embrechts, Klüppelberg and Mikosch, 1997; Embrechts, 2000) The paucity of data on operational risks also means that it is essential to combine what data there is with expert opinion.

The calibration of expert opinion by small data sets is itself a difficult theoretical area. (Bedford and Cooke, 2001; Clemen and Winkler, 1999). These issues and how they are dealt with in bank practice are taken up more fully in the supplementary report.

3.3 Ernst & Young Structured Asset Portfolio Case

EY kindly gave our team briefings on a case in which they acted as consultants. The details remain confidential, but the important structure of the case involved EY's client marketing an innovative financial product which required a reliable stream of income over a considerable number of years. Unusually, the stream was to be covered by the income from a structured asset portfolio, including infrastructure, property and fixed interest. EY were asked to quantify the economic risk behind using structured assets to meet a defined cashflow stream and quantify the risk profile over time and how it could be offset by management actions. They did this by performing a detailed analysis of the risks for the company and assessing the requirements to hold a required reserve to reduce the probability of risk within the risk appetite of key stakeholders including the board, management and regulators. The bulk of the task was the analysis of the risks involved in meeting the liability cashflows from the asset portfolio, developing models to project risks and management behaviour in response to those risks, and the attendant process of embedding risk management systems in the company's activities.

It is in the nature of many non-financial assets such as property or an infrastructure asset that one must rely in part on judgments about the prospects of an individual asset, such as a pipeline, rather than on market averages. There will be relevant data, but it must be combined with opinion on the unique characteristics of the individual asset. In the long time scale required in this case, one must also consider the probability of major downturns in the asset value. Taking property as an example, there has not been a significant downturn since the one that followed the stock market crash of 1987, so data is again scarce. Then there is the possibility of natural disasters, terrorist attacks and the like (which could affect several properties at once, for example if they are all located in the CBD).

EY approached this problem in part by developing a suite of risk scenarios, preparatory to evaluating the risk of the product. Scenarios to which the product must be robust are not purely fanciful – they should be ones of low but non-negligible probability – of the order of one-in-a-thousand chance of occurring in a year. To obtain ballpark figures for the probability of risks such as natural disasters, EY was able to draw on their experience and data. In what could be termed a consensus method, such information was used to make a reasonable judgment of, say, the probability that all assets in a common location would be weather-damaged and unusable for a year. This risk figure comprises the estimated financial impact of such an event (again, drawing on market figures), along with the probability. In evaluating such scenarios, there is relevant data, but there is an essential step in intelligently applying it to the particular case and in then convincing the stakeholders that the application to the case is reasonable. An important component of the analysis was to disaggregate the risks into the component parts each on a timeline and identify actions that could be taken at each point in time should an adverse event occur that threatened cashflow.

At the same time, management and mitigation strategies – how can we reduce the risk? how could we deal with such an eventuality? – must be developed. Some risks, for example physical damage, are insurable, so insurance is one potential mitigation strategy. But if insurance is taken out, the potential gaps in the insurance cover (that is, the risk remaining after insurance) must be evaluated. It is a requirement of the regulator – and essential for the company – that risk management systems be in place. However, such systems necessarily alter the sensitivity to risk, so there is a feedback process between quantifying risks and controlling exposure to them.

3.4 The Vargas Flood Tragedy, Venezuela

In December 1999, a daily precipitation event of magnitude 410mm was recorded at Maiquetia International Airport, in Vargas, Venezuela – almost three times the magnitude of the previously recorded maximum (Figure 1). It was the main cause of the worst environmentally related tragedy in Venezuelan history and one of the largest historical rainfall-induced debris flows documented in the world. Massive flooding and landslides washed away an entire state, producing a number of deaths that has been estimated at between 15,000 and 50,000. Around 8,000 residences and 700 apartment buildings were destroyed or badly affected with damages being estimated at around US$2 billion (Coles et al 2003).

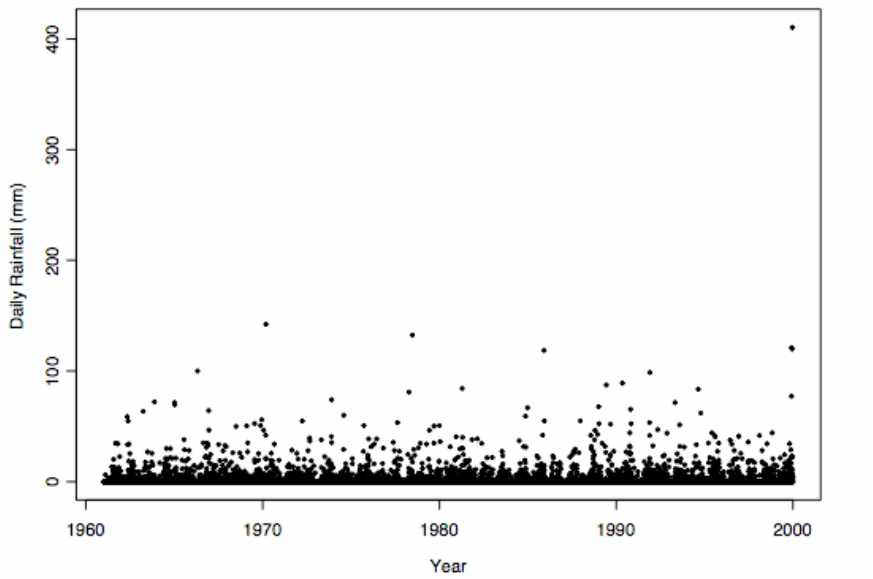

Figure 1: Daily rainfall values recorded at Maiquetia International Airport, Venezuela. (Source: Sisson *et al*, 2006).

A natural question to ask is what was the likelihood of the observed disaster occurring, before it was witnessed, based on historical rainfall records? Recorded rainfall data exists from 1951 for annual maximum daily rainfall, and from 1961 for daily observations. Standard models fitted to the pre-1999 data attach a probability of virtually zero to the actual 1999 event. It was therefore argued that the event itself was impossible to forsee. Given that this event was observed in only 50 years of recorded data, an objective observer may have reservations regarding this conclusion. Implicitly, this suggests a model failure, either because of a violation of the assumptions on which the model was built, or because of a sudden change in the meteorological climate (Coles et al 2003 found no evidence of climate change).

We will show how extreme value theory can be applied in this setting in order to produce expected "return periods," measured in years, for the 1999 event. We will demonstrate that careful statistical modelling is required. Naïve application of extreme value theory can result in clearly erroneous estimates that the 1999 event will occur once every 18 million years on average (Coles and Pericchi, 2003). More considered approaches, including the utilisation of Bayesian inference, can provide more realistic inference. We will return to this study shortly.

3.5 The Sinking of the *M. V. Derbyshire*

On 9th September 1980 the bulk carrier M. V. Derbyshire sank in the Pacific Ocean when she was caught in Typhoon Orchid while transporting iron ore from Canada to Japan. All 44 people on board died. The Derbyshire remains the largest UK ship to have been lost at sea, and in spite of being in good operating condition, appears to have sunk suddenly and without warning. Evidence that the Derbyshire may have suffered a structural design weakness came from incidents involving two of her sister ships, both of which suffered a particular form of structural failure. Following an earlier investigation that tentatively blamed crew error for the sinking, a high-court enquiry was heard during April—July 2000 (Coleman, 2000).

Underwater photography suggested that the hatch cover on the front-most hold failed during the typhoon. The safety standards at the time of the ship's construction required that the hold covers were capable of withstanding an impact pressure of 42 kilopascals (kPa). Interest was therefore in the probability that the pressure of wave impacts on the hold cover exceeded its collapse pressure

of 42 kPa at any point during Typhoon Orchid. Uncertainty about the precise state of the vessel, her speed during the typhoon and wave conditions meant that risk analyses were performed on a range of scenarios, in order to identify the conditions most likely to lead to the sinking. However, previous studies showed greater sensitivity to the choice of statistical model rather than from vessel or wave conditions. These analyses fitted a Weibull distribution (widely used in modelling wave impacts) and Gumbel distribution (poorly justified under an extreme value theory argument) to wave impacts generated from a wave test tank for each of 60 different scenarios, comprising 15 sets of wave and initial damage conditions, and four different ship speeds. However, it was soon found that differences were more dependent on the statistical model than the differing scenarios, thereby masking evidence for the cause of the sinking. For example, under one set of wave and vessel conditions, estimates of the probability of a hull breach were in the range [0.73, 1.00] for the Weibull model and [0,0.3] for the Gumbel model.

Two extreme value theory specialists (Heffernan and Tawn, 2003, 2004) were brought before the enquiry to resolve these problems and perform a modern statistical analysis. We shall return to this study in the section on Extreme Value Theory.

3.6 The Challenger Disaster

On January 27, 1986, the night before the disastrous launch of the Space Shuttle Challenger, a long teleconference with faxed data sheets was held between a number of engineers and managers. Discussion focussed on the issue of the performance of the solid rocket motor's O-rings at low temperatures, the problem that in the event caused the disaster. The team of engineers in Utah from Morton Thiokol, the manufacturers of the rockets, expressed various doubts on the reliability of the O-rings at the predicted launch temperature of 31°F. All previous launches had taken place at temperatures between 53°F and 81°F, so it was rightly understood that the main issue was whether there was any relationship between O-ring performance and temperature.

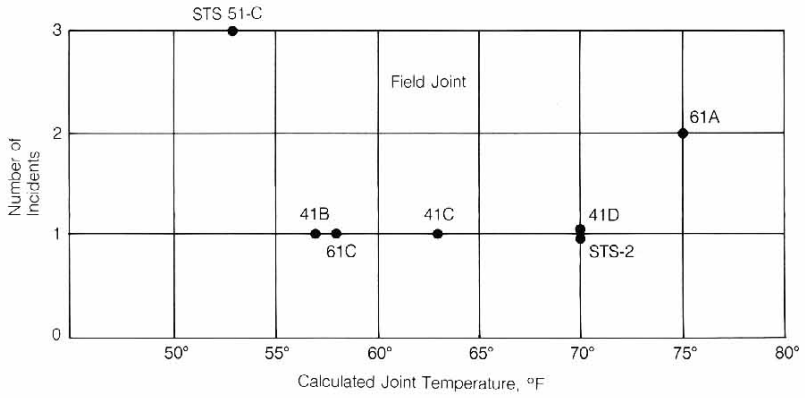

Figure 6
Plot of flights with incidents of O-ring thermal distress as function of temperature

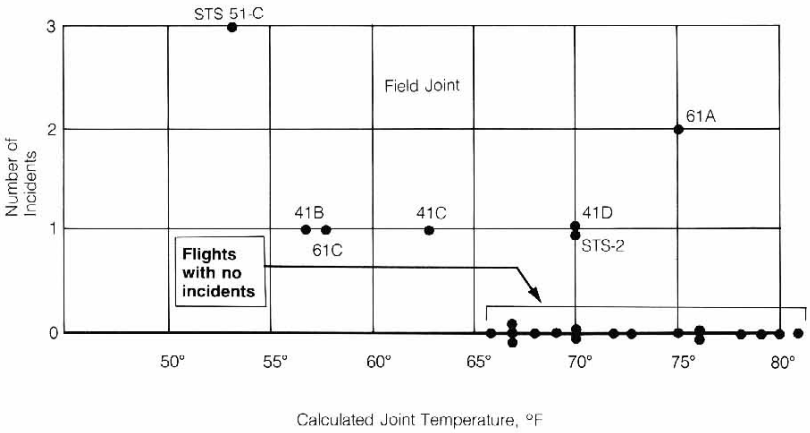

Figure 7
Plot of flights with and without incidents of O-ring thermal distress

NOTE: Thermal distress defined as O-ring erosion, blow-by, or excessive heating

Figure 2: Incidents of Thermal Distress to O-rings prior to the Challenger launch (Source: *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, Volume 1 p. l46, http://history.nasa.gov/rogersrep/v1p146.htm)

The engineering and management backgrounds of the participants led them to generally concentrate on causal considerations concerning individual cases rather than graphing the various observed faults against temperature. (Lighthall, 1991) Even then, the data presented had a major flaw. Above, from the report of the Presidential Commission of inquiry, is a graph of the information actually presented to the teleconference (top), and what should have been presented (bottom). (Even the top graph is very much clearer than the tables in fact presented (printed in Vaughan, 1996, pp. 293-9 and analysed in Tufte, 1997, pp. 39-45)). The top graph has only the data points (a minority) in which damage to the O-rings occurred, the data from flights where there was no damage having been omitted in the erroneous belief that it was uninformative.

The information presented to the teleconference makes it appear that there is probably no relationship between temperature and O-ring damage, whereas the graph with all the data makes it obvious that all the flights with the coldest temperatures suffered damage, whereas very few of the high-temperature launches were damaged. The fact that the worst damage occurred on the coldest previous launch, which at 53°F was still very much warmer than the 31°F predicted for the

forthcoming launch, should have raised alarm. Though that fact is in theory visible in the information actually presented, the absence of the full data set hid the trend. That allowed the managers in Florida keen to launch on schedule to concentrate attention on the other unusual data point, the damage to two O-rings at the warm temperature of 75°F. The Morton Thiokol engineers broke off the teleconference for half an hour to discuss among themselves. They then accepted the reasoning of the managers that if there were a problem with cold temperature then damage at 75°F would be most unlikely. The launch was approved.

Later elaborate analyses by statisticians using logistic regressions and other possible models (Dalal et al, 1989; Lavine, 1991; Tappin, 1994), add little to the basic point. The trend obvious by eye in the second (full) graph is not made more convincing by the fitted models – on the contrary, the models are appropriate choices mainly because they agree with the intuitive fit. What was lacking in the teleconference's analysis was not any sophisticated statistical model or formula, but a basic statistical perspective that would have understood the importance of graphing all the data points on a chart with the relevant variables on the axes, and an appreciation that extrapolation far beyond a data set is dangerous and is not compensated for by causal narratives, especially ones that concentrate on one or two extreme data points.

3.7 The Browns Ferry nuclear reactor fire

This remarkable event, one of the worst nuclear reactor accidents in a Western country, is a classic reminder of what philosophers of science call the "threat of the unknown hypothesis": the probability that a disaster is a result of a concatenation of causes entirely outside the range of what one has considered in one's risk analysis. (Earman, 1992, pp. 168, 228-9) As one account describes it:

On March 22, 1975, a fire at the Browns Ferry Nuclear Power Plant fundamentally changed the concept of fire protection and associated regulatory requirements for U.S. nuclear power plants. Plant workers were fixing leaks in the cable spreading room outside the reactor building. The workers used a candle to test seals for air leaks into the reactor building. The polyurethane foam seal, however, was not fire-rated. The flame from the candle ignited both the seal and the electrical cables that passed through it.

By the time firefighters extinguished the fire, it had burned for almost 7 hours. More than 1600 electrical cables were affected, 628 of which were important to plant safety. The fire damaged electrical power, control systems, and instrumentation cables and impaired cooling systems for the reactor. Operators could not monitor the plant normally and had to perform emergency repairs on systems needed to shut the reactor down safely. (USNRC, 2006)

There were also major inadequacies in the human response to the unfolding disaster.

Plainly, statistical methods are not capable of finding the chance of something entirely unexpected happening, since statistics is based on counting events in some pre-defined space of possibilities, a space which is by definition not available for unknown hypotheses.

4. Relevance of individual data points

Before moving to a survey of methods, there is one lesson for extreme risk evaluation that is suggested by the case studies. A characteristic of extreme risk assessments is the existence of individual data points whose relevance is itself a matter of dispute. Certain disasters or near-disasters that have actually happened will be well-known, but for the very reason that they are known, steps will have been taken to prevent them happening again, so the relevance of the incident to present evaluations is unclear and needs to be assessed – certainly, before the data point is included in a data set to which statistical methods will be applied (methods which will typically make much of an extreme value).

An example is NAB's loss of about $330m from rogue trading in 2004. It is one of the few large (known) operational risk losses in recent times by an Australian bank. Its relevance to present operational risk evaluations, whether for NAB or other banks, is unclear, since NAB has had to submit to APRA detailed evidence of the steps it has taken to tighten procedures to prevent a recurrence (and APRA believes other institutions have been encouraged by the case to improve their risk management as well.) It is typical of bank operational risk data, especially in the stable Australian environment, that the few comparatively large losses are from some years ago and hence are of doubtful relevance (the lack of recent losses of the same order itself suggesting that procedures are now improved). Yet the operational risk analysis, if conducted rightly with heavy-tailed distributions, will be very sensitive to these few data points, and in fact most of the reserved capital for operational risk may be due to them.

Another instance of a data point needing debate comes from Biosecurity Australia's 1998 Import Risk Assessment for New Zealand apples. The Assessment mainly dealt with the risk of imported apples causing an outbreak of fire blight, a disease that is not present in Australia. In 1997, while the report was being prepared, fire blight was discovered on two shrubs in the Royal Botanic Gardens, Melbourne, and the assessment was suspended pending study and eradication. It never became clear how the disease entered the Gardens or how long it had been there. On the one hand, the outbreak had been there for an unknown time without spreading, while on the other hand it was rather unlikely that the outbreak had been caused by the import of commercial fruit. The IRA concluded that nothing of relevance to the IRA could be learned from the episode. (AQIS, 1998, pp. 9, 21)

Argument about such individual cases is essential to the analysis of extreme risks for several reasons. The cases are probably well studied, so that there may be much that can be learned from them, and reasonable judgments can be made as to whether the same could happen again (or something different but in some way from a similar cause). It is unsatisfactory to either simply delete the data point as no longer relevant or simply leave it in to drive some quantitative statistical method.

Methods for predicting extreme values are very sensitive to the few most extreme values in the data, so great care needs to be taken in determining if they are "from the same distribution", that is, fully relevant to the prediction problem at hand. Mechanical methods (though sometimes applicable for identifying outliers) are not suitable for examining data points where there is extra knowledge of the particular case. An advocacy model will allow that knowledge to be brought to bear before the data point contaminates any later analysis.

Equally important in principle is the problem that extremes may be missing from the data. For example extremes of rainfall in hurricanes may be missing because the worst floods were large

enough to wash away the recording instruments. (Hellin *et al.*, 1999) Plainly the data needs to be queried for such possibilities before any quantitative analysis is applied.

5. Methodology

Next we survey a number of methods and bodies of knowledge, both quantitative and qualitative, that are at present not standardly used in risk analysis but which show promise of application in that area.

5.1 Outlier detection and fraud detection

Before any statistical method for extreme risks is applied, it is crucial to distinguish extremes from outliers. Extremes are the data points at the edge of the distribution, such as years of particularly high rainfall. They are the important data points on which the assessment of extreme risks will most depend. Outliers, by contrast, are points that are not part of the distribution at all. Typically an outlier is a mistake (e.g. drunks have urinated in the rain gauge, data entry typists have put the decimal point in the wrong place) and it should be deleted from the data set. But an outlier may also indicate some kind of contamination of the data by "another distribution", which could indicate an event deserving further investigation. For example, an outlier reading of some measure of an environmental pollutant may indicate an illegal discharge – the normal range of data comes from natural processes, but the illegal discharge is a completely different cause or distribution. (On this way of understanding the concepts, outliers and extremes are different by definition, and there is a statistical problem, sometimes difficult, of knowing whether a given data point is one of the other.)

So a basic knowledge of statistical methods of outlier detection is essential for the analyst of extreme risk.

Hodge and Austin (2004) summarise its applications to monitoring:

> Outlier detection is a critical task in many safety critical environments as the outlier indicates abnormal running conditions from which significant performance degradation may well result, such as an aircraft engine rotation defect or a flow problem in a pipeline. An outlier can denote an anomalous object in an image such as a land mine. An outlier may pinpoint an intruder inside a system with malicious intentions so rapid detection is essential. Outlier detection can detect a fault on a factory production line by constantly monitoring specific features of the products and comparing the real-time data with either the features of normal products or those for faults. It is imperative in tasks such as credit card usage monitoring or mobile phone monitoring to detect a sudden change in the usage pattern which may indicate fraudulent usage such as stolen card or stolen phone airtime. Outlier detection accomplishes this by analyzing and comparing the time series of usage statistics. For application processing, such as loan application processing or social security benefit payments, an outlier detection system can detect any anomalies in the application before approval or payment. Outlier Detection can additionally monitor the circumstances of a benefit claimant over time to ensure the payment has not slipped into fraud.

There are various methods, both visual and formula-based, for testing the discordancy of a suspected outlier with the rest of the data. (We do not survey them here but refer to Barnett, 2004, ch. 3; further in Hand and Bolton, 2004; for established "control-charting" graphical methods suitable for univariate time series data not too far from normal see Fox, 2007) They all involve modeling the data in some way, that is, finding some distribution that fits (most of) the data well, and which if true implies that the outlier is very unlikely to have occurred. An outlier, as the *Encyclopedia of Statistical Sciences* defines it, is "some observation whose discordancy with the majority of the sample is excessive in relation to the assumed distributional model for the sample, thereby leading to the suspicion that it is not generated by this model." That points up the problem of knowing the distribution of the data, especially in cases where the data itself, including the

outlier(s) (rather than some causal knowledge), is the only source of knowledge of the distribution. There is no good solution to this fundamental problem; Barnett (2004, p. 55) writes:

> In practice, this dilemma is well recognized and it is usually resolved by a judicious mix of historical precedent (`such and such a model has always been used in this problem') broad principle (`randomness in time and space implies a Poisson form'), association (linking the problem to apparently similar types of situation), some data-fitting (e.g. probability plotting), and an element of wishful thinking (as in all areas of model-based statistics).

The best-developed theories of outlier detection (also called in various contexts "novelty detection", "anomaly detection" and "exception mining") deal with univariate (one-dimensional) data, where the data are points on a line and an outlier is a point well beyond the limits of the rest of the data. In a very simple example,

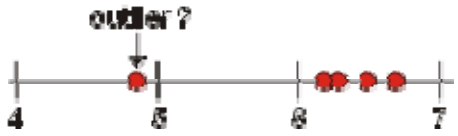

Figure 3: An outlier in a 1-dimensional data set
(Source: http://www.chem.uoa.gr/Applets/AppletQtest/Text_Qtest2.htm)

The outlier detection problem for multivariate data is essentially harder, since the data are not ordered so it is not so clear what "beyond the rest of the data set" means. Nevertheless it is sometimes clear why a data point is far from the natural model of the rest of the data, as in this example, where a straight line fit seems intuitively appropriate for the main body of the data:

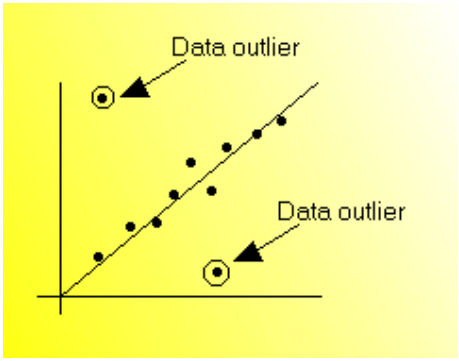

Figure 4: Outliers in a 2-dimensional data set
(Source: http://www.ecfc.u-net.com/cost/compare.htm)

Visual methods and some of the formula-based methods are not so easily applicable to data with many dimensions – and the impressions easily gained from one- or two-dimensional data sets do not easily carry over to high-dimensional space.

But here there have been many recent advances driven by the needs of fraud detection. The problem in fraud detection typically involves a large data set with each data point having many attributes, and the aim is to identify by automatic methods "unusual" or "fringe" data points that warrant further investigation. A challenging application area is the detection of fraud in credit card transactions, where millions of requests per day need to be scanned for possible frauds of unpredictable kinds, and action taken in real time to refuse suspicious transactions and suspend the

card (while not annoying legitimate customers with false alarms). Many of the details of how it is done are not publicly available, for obvious reasons, but the methods used include a combination of supervised and unsupervised approaches. Supervised methods learn from existing tagged data, so can take advantage of persistent known patterns in fraud, such as sudden multiple purchases of jewellery by cardholders who have not purchased jewellery before. Unsupervised methods, which look for new patterns far from existing data, are added to deal with the ever-developing ingenuity of fraudsters. (Bolton and Hand, 2002, section 3) There have been uses of these methods in such problems as AUSTRAC's money laundering detection and the detection of intrusion in computer systems, but so far less application to such potential areas as environmental monitoring, epidemic alerting or quarantine inspection.

5.2 Extreme Value Theory: Scope and Limits

We now introduce the basic ideas involving the statistical modelling of extremes. Further details can be found in e.g. Coles (2001), Kotz and Nadarajah (2000), Beirlant *et al* (2004).

Suppose we have a sequence of independent random variables $X_1, X_2, \ldots, X_n$ drawn from a common distribution function F. Classical extreme value theory models focus on the statistical behaviour of

$$M_n = max\{ X_1, X_2, \ldots X_n \}.$$

The X_i usually represent (continuous) values of a process observed on a regular timescale, such as daily rainfall amounts or log daily returns of some stock. Biosecurity applications have not been much studied but could include daily levels of a contaminant or the distance travelled by a spore. M_n therefore represents the maximum of the process over "blocks" of n units of observation – e.g. if n is the number of observations in one year, then M_n corresponds to the annual maximum. In theory, the distribution of M_n is known exactly, as:

$$\begin{aligned}\Pr(M_n \le z) &= \Pr(X_1 \le z, X_2 \le z, \ldots, X_n \le z) \\ &= \Pr(X_1 \le z)\Pr(X_2 \le z)\ldots\Pr(X_n \le z) \\ &= [F(z)]^n\end{aligned}$$

In practice however, the distribution $[F(z)]^n \to 0$ as $n \to \infty$, so that the distribution of M_n degenerates to a point mass on the smallest value such that $F(z)=1$. To avoid this, and in analogy with the central limit theorem, a linear re-normalisation is performed

$$M_n^* = (M_n - b_n)/a_n$$

for sequences of constants $\{a_n > 0\}$ and $\{b_n\}$. Appropriate choices of the constants stabilise the location and scale of M_n^* as n increases. It can be shown that for suitable choices of constants

$$Pr((M_n - b_n)/a_n \le z) \to G(z) \text{ as } n \to \infty$$

where $G(z)$ is the Generalised Extreme Value (GEV) distribution, with distribution function

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \quad (1)$$

where $[a]_+ = \max(0,a)$. The parameters μ, σ and ξ correspond to location, scale and tail shape parameters respectively. The GEV distribution incorporates into a single formulation, three families of extreme value distributions: the Weibull, Gumbel and Fréchet distributions. The limit of M_n^* converges to one of these types – for example, if F is the normal distribution then the limit distribution of the block maxima is the Gumbel. Limit discrimination arises through the value of the tail shape parameter. For $\xi<0$ the GEV yields Weibull tails with a finite upper endpoint of $\mu - \sigma/\xi$. Realising a Weibull distribution is often of high practical importance, as there is a clear maximum bound for the process under study. For example, Heffernan and Tawn (2003, 2004) established Weibull tails for their study of maximum wave impacts in the Sinking of the M. V. Derbyshire case study (see later). For $\xi>0$ the GEV gives Fréchet tails with no upper endpoint. This is a polynomial decay, where the larger the value of ξ, the heavier the tail. Fréchet tails are common in environmental applications, such as precipitation (Smith 1989, Sisson *et al* 2006), and in financial applications, such as insurance (Bottolo et al 2003). In the limit as $\xi \to 0$, the GEV reduces to the Gumbel distribution, with exponential decay tails. While certain distributions F do lead to Gumbel block maxima, it is uncommon to estimate $\xi = 0$ from data, as the uncertainty in the parameter estimate would also place the model in both Weibull and Fréchet domains. Insisting on the Gumbel in these situations can have severe implications (Coles and Pericchi, 2003).

For example, if the common distribution function F is known, the distribution of M_n^* may then be calculated explicitly. For example, if $X_1, X_2, \ldots, X_n$ is a sequence of independent exponential Exp(1) variables, $F(x) = 1 - exp(-x)$ for $x>0$. In this case, letting $a_n=1$ and $b_n=log(n)$,

$$\begin{aligned}
\Pr\left(\frac{M_n - b_n}{a_n} \le z\right) &= F^n\left(z + \log(n)\right) \\
&= \left[1 - \exp\left(-\left(z + \log(n)\right)\right)\right]^n \\
&= \left[1 - \exp(-z)/n\right]^n \\
&\to \exp(-\exp(-z))
\end{aligned}$$

as $n \to \infty$ for each fixed z. Hence with the chosen $\{a_n\}$ and $\{b_n\}$, the limit distribution of M_n as $n \to \infty$ is the Gumbel distribution, corresponding to $\xi=0$ in the GEV family.

In many applications, the distribution F is unknown. One possibility for inference is to use standard statistical techniques to estimate F from observed data, and then to use this to estimate F^n. Unfortunately, very small discrepancies in the estimate of F can lead to substantial discrepancies for F^n. In practice then, a better approach is to take advantage of the single GEV formulation of the Weibull, Gumbel and Fréchet distributions, by statistically fitting this distribution to the observed sample maxima. The estimated value of the tail shape parameter will determine which family the sample maximum belongs to. Of course, in basing estimation on observed, finite data, one is making an explicit assumption that the distribution of M_n for the observed finite n, has converged sufficiently closely to the distribution of M_n as $n \to \infty$, that the differences in finite and asymptotic maxima distributions are negligible. This assumption can sometimes break down. For example, Koutsoyiannis (2004) shows that even when the asymptotic distribution is Gumbel, the convergence of the sample maximum to this limit can be very slow and that fitting the full GEV distribution to observed data rather than any particular sub-family (Weibull, Gumbel, Fréchet) should be recommended in practice. This point is also made by Coles and Pericchi (2003) and Sisson *et al* (2006).

Similarly, the practitioner must make a conscious decision when partitioning observed data into blocks of size n, amounting to the trade-off between bias and variance: blocks that are too small mean that approximation by the limit model is likely to be poor, leading to bias in estimation; large blocks generate few block maxima leading to large estimation variance. In practice, pragmatic considerations often lead to the adoption of blocks of length one year. For example, daily temperatures are likely to vary according to season, violating the assumption that the X_i have a common distribution F. (Daily temperatures are also not independent, an issue discussed below.) If the data were blocked into block lengths of around 3 months, the maximum of the summer block is likely to be much greater than that of the winter block, and an inference that failed to take this non-homogeneity into account would be likely to give inaccurate results. Taking instead blocks of length one year means the assumption that individual block maxima have a common distribution is plausible, though formal justification for this is invalid (Coles, 2001).

Common approaches of fitting extreme value distributions include maximum likelihood, the method of moments, L-moments and graphical and quantile-based methods. Kotz and Nadarajah (2000) examine numerous methods. Maximum likelihood is common given the attractive properties of these estimators. Smith (1985) showed that for maximum likelihood to give reliable results, $\xi > -1/2$ is required. This is not usually a practical problem as $\xi \leq -1/2$ corresponds to a very short bounded upper tail, where $\xi > 0$ for most environmental problems, and $2 < \xi < 4$ for many financial applications.

Once estimates of model parameters are obtained, the GEV distribution function (1) may be inverted:

$$z_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right] & \xi \neq 0 \\ \mu - \sigma \log[-\log(1-p)] & \xi = 0 \end{cases}$$

where $G(z_p) = 1 - p$. In common terminology, z_p is the return level associated with the return period $1/p$, since to a reasonable degree of accuracy, the level z_p is expected to be exceeded on average once every $1/p$ years. More precisely, z_p is exceeded by the annual maximum in any particular year with probability p.

Returning to the Vargas Tragedy case study, Figure 5 illustrates a return level plot of the Gumbel and GEV models fitted to block maxima data where each block corresponds to one year. The data used for fitting excluded the 1999 datum. Visually, the GEV model appears to fit the upper values of the data better than the Gumbel. Interpreting the plot more precisely, the return-period estimate of the 1999 event of 410mm under the GEV model is approximately 4,280 years. Under the Gumbel it is around 17.6 million years. Both estimates attach a virtually zero probability to the 1999 event, implying that an event of this magnitude should be considered impossible, had it not been observed. It also illustrates the dangers of not fitting the full GEV model. Coles and Pericchi (2003) demonstrated that in a hypothesis test of GEV versus Gumbel (i.e. a test that the tail parameter $\xi = 0$) could not reject the Gumbel in favour of the GEV. However reducing the extreme value model to the Gumbel has multiple orders of magnitude of difference in the return level estimates of the observed event. This arises as the Gumbel and Fréchet distributions have substantially differing tail decay rates. The Gumbel decays only exponentially, whereas the Fréchet decays polynomially thereby giving greater weight to extremely large events. Model simplifications when parameters must be estimated from the data should therefore be treated with extreme caution.

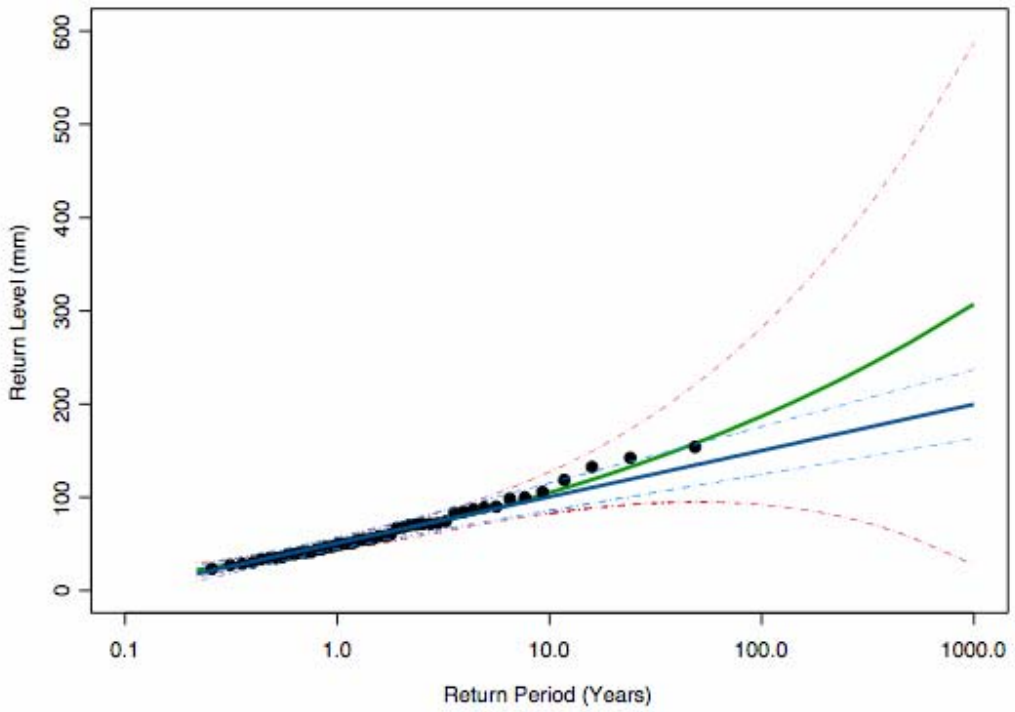

Figure 5: Return level plots for Gumbel (lower solid) and GEV (upper solid) models of Venezuelan annual maximum daily rainfall. Dotted lines correspond to 95% confidence intervals. Points correspond to empirical estimates $[i/(n+1), M_n^{(i)}]$, i=1,…,50, where $M_n^{(i)}$ is the *i*-th largest block maxima. (Source: Sisson *et al*, 2006).

It may be argued that block maxima analyses are wasteful of data, which is by definition already naturally scarce. Only the largest value in each block is used to fit the model. If other data on extremes are available they should also be used. An alternative formulation to extreme value modelling is based on threshold exceedance models (Smith, 1989). It is natural to regard as extreme events those of the X_i that exceed some high threshold u. For instance, rather than fitting (say) a Gaussian distribution to observed market daily returns, and then extrapolating into the tails of this distribution to estimate extreme quantiles, one may model the quantiles directly. In this manner, modelling assumptions regarding the form of tail decay for the body of the data, which may be too light (e.g. Gaussian distributions modelling daily returns), are not enforced on extreme levels of the process. It can be shown that for large enough u, the distribution function of $Y = (X - u)$, conditional upon $Y > 0$, is approximately distributed as a generalised Pareto distribution, with distribution function

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)_+^{-1/\xi} \qquad (2)$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ is a function of the threshold and respective GEV block maxima parameters. That is, the parameters of the generalised Pareto distribution of threshold excesses are uniquely determined by those of the associated GEV distribution of block maxima. In particular, the tail shape parameter ξ is identical. However, choosing different block sizes n would affect values of the GEV parameters, but not those of the corresponding generalised Pareto distribution of threshold excesses: ξ is invariant to block size, while calculation of $\tilde{\sigma}$ is unperturbed by the changes in μ and σ which are self-compensating.

The duality between the GEV and generalised Pareto families means that the shape parameter is dominant in determining the qualitative behaviour of the generalised Pareto distribution. If $\xi < 0$ the distribution of excesses has an upper bound of $u - \tilde{\sigma}/\xi$, if $\xi > 0$ the distribution has no upper limit. The distribution is also unbounded if $\xi = 0$, which should again be interpreted by taking the limit $\xi \to 0$ in (2), leading to an exponential distribution with parameter $1/\tilde{\sigma}$.

Inverting the Pareto distribution function gives

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right]$$

the level (x_m) that is exceeded on average once every m observations, provided that m is sufficiently large to ensure that $x_m > u$. Here $\zeta_u = \Pr(X > u)$ can be estimated from the data. For presentational convenience, if there are n_y observations per year, setting $m = N \times n_y$ will yield the N-year return level.

Choice of threshold is similarly akin to the variance/bias trade-off of block maxima. Threshold choice also remains somewhat ad-hoc, based on mean residual life plots or by fitting models to a range of thresholds, and identifying thresholds with parameter stability (Coles, 2001). It may also potentially be estimated by specifying models for non-extreme data, although arguably this cannot be justified as F is unknown.

Returning again to the Vargas Tragedy case study, fitting the threshold excess model to the *daily* rainfall observations, followed by inversion of the distribution function, yields a return period of 752 years for the observed event of 410mm (Coles *et al* 2003). This compares favourably with the estimates of 4,280 and 17.6 million years obtained via using annual maximum data only, as it now seems at least possible that we could have observed such an event within the 50 years of observed data.

However, in modelling the data on a daily threshold exceedance level, we are now probably in violation of the assumption that these data are drawn from the same distribution – rainfall patterns are unlikely to be constant over the entire year. In fact, meteorology in the Caribbean region can be broadly classified into "dry" and "wet" seasons (González and Córdova, 2000). One common approach in practice is to permit the model parameters σ and ξ to vary with time: $\sigma(t)$, $\xi(t)$. In the Venezuelan case, we can specify two seasons, and permit different (constant) parameter values within each season. Knowledge of where these seasonal "change points" occur is problematic though. We will return to this issue in the Bayesian setting shortly.

Returning to the M. V. Derbyshire case study, Heffernan and Tawn (2003, 2004) fitted the generalised Pareto distribution to the wave tank datasets under 60 different vessel and wave condition scenarios, previously analysed by Weibull and Gumbel distributions. Visual diagnostics, such as QQ-plots for the fitted models, revealed that the Weibull distribution substantially overestimated the probability of large events, and the Gumbel distribution did not fit the middle of the distribution. In contrast, the fit was remarkably good for generalised Pareto exceedances. Information about vessel and sea conditions was incorporated into the analysis by permitting the model parameters σ and ξ to be a function of these conditions. It was discovered that the estimates of the tail shape parameter were not dependent on the scenario conditions, and so the data could be pooled to estimate this parameter.

Since many of the estimated risks were small, Heffernan and Tawn (2003) were asked to provide estimates of the worst possible impact that the ship could have received in each set of conditions. This was only possible since their estimated tail shape parameters ξ were negative, and so the estimated maximum impact distributions had finite upper end points.

Paragraph 6.13 of the Judge's report (Coleman, 2000) states that the analyses were of "absolutely fundamental importance to the outcome of this Investigation." The report also raised questions about the adequacy of current regulations governing the strength of hatch covers.

At this point it's worth discussing two related issues. The first concerns whether the observations X_i of the process being modelled, come from the same distribution F. For example, in the Vargas Tragedy, there is a case for arguing that the December 1999 event was generated by a different meteorological process (with distribution F_2 say) than the pre-1999 event data (with distribution F_1). Obviously, before observing the December 1999 event, one may only make statistical inference based on the data drawn from F_1. Once we have observed the outlier event, we effectively have a random sample drawn from the mixture distribution

$$F = \omega F_1 + (1-\omega)F_2$$

where ω denotes the (large) proportion of observations from the first component (Coles *et al* 2003). Statistically estimating parameters based on all available data then effectively models and makes prediction according to the process F. That is, inference is based on the observed process without knowing anything about the causal process (this is the power of the mathematical results behind extreme value theory). However, there is still the assumption that the observations are drawn independently from F. That is, the December 1999 F_2 event could have occurred at any time in the observation period, with probability $(1-\omega)$. If it were the case that the occurrence of the December 1999 event had changed (or signalled the change of) the underlying generating process, then beyond this point all observations would be drawn from some new process distribution F_3. The data generated from distributions F_1 and F_2 would then be uninformative in making inference on future occurrences drawn from F_3. The exception to this is if there is some knowledge about how the effects of the processes (in terms of generating extreme events) may be linked (e.g. the likelihood of a similarly extreme event is reduced by some factor), in which case modelling may proceed under this assumption using all data.

The second issue then is that under both GEV and threshold modelling frameworks, there is an underlying assumption that the X_i are independent. This is clearly unlikely to hold in practice – maximum daily temperatures or rainfall amounts on consecutive days are highly correlated. The block maxima approach avoids this problem unless the dependence extends onto the (say) annual scale. In fact, for asymptotic convergence to hold, some form of weak dependence is permitted (Leadbetter *et al.* 1983), although convergence to the GEV or Pareto limit is accordingly slower. The permitted dependence is sufficiently weak that it cannot be justified for the modelling of dependent observations.

In practice there are two options for modelling dependent observations, depending on the form of the dependence. The first is to permit the model parameters to depend on some quantity (e.g. time, space or other predictors), which will account for the relationship between observations. This approach would be appropriate for the modelling of e.g. daily temperatures, which are expected to be higher in summer than in winter months. If this approach is not sufficient, then one must model the dependence explicitly. This requires moving to multivariate extreme value methods.

Multivariate extremes can be modelled analogously to their univariate counterparts. Consider a bivariate setting where (X_1,Y_1), (X_2,Y_2), … is a sequence of independent vectors with common

distribution function F_2. The extension to the multivariate setting is immediate. Similarly to the univariate setting, define:

$$M_{x,n} = max\{ X_1, X_2, \ldots X_n \}, \qquad M_{y,n} = max\{ Y_1, Y_2, \ldots Y_n \},$$

and then specify

$$\boldsymbol{M}_n = (M_{x,n}, M_{y,n})$$

as the vector of component-wise maxima. Note that the index i, for which the maximum of the X_i sequence occurs, need not be the same as that of the Y_i sequence, so $\boldsymbol{M}_n$ does not necessarily correspond to an observed vector in the original series. Asymptotic theory then states that as $n \to \infty$, not only does the distribution of each marginal distribution converge to a GEV limit, but the joint distribution also converges to a class of bi-variate (multi-variate in general) extreme value distributions. This is a large class of distributions incorporating the full range of dependence (from complete independence to total dependence).

A similar multivariate representation exists for threshold exceedance models (Haan and Resnick 1977, Pickands 1981). As in the univariate setting, all multivariate representations may be derived as special cases of the point process representation.

When modelling multivariate extremes, in addition to the greater scarcity of data, a particular difficulty can arise. Consider, for illustration, a vector (X,Y) which is distributed as bivariate Gaussian with correlation $\rho=0.999999$. The pair is very strongly dependent, and so when X is large, Y also tends to be large. Define

$$\chi = \lim_{z \to \infty} \Pr(Y > z \mid X > z)$$

to be a limiting measure of the tendency for one variable to be large conditional on the other variable being large (Coles *et al* 1999). For the bivariate Gaussian pair, no matter how strong the correlation between X and Y, subject to $\rho<1$, the limit $\chi=0$ states that they are asymptotically independent from each other (Sibuya 1960). The opposite effect can be conceived in financial markets: everyday market shifts in two unrelated stocks mean that at low—large levels, increases or decreases in their stock prices are likely to be unrelated to each other. However, consider a market crash situation. A precipitous drop in one stock is quite likely to occur concurrently with one in the other. That is, in this setting, the two processes are independent until "asymptotically" large levels are reached, when dependence is achieved. Evaluating return levels can lead to differing results at very large levels depending on the type of asymptotic dependence the model adopts. Exploration of models that admit both asymptotic independence and asymptotic dependence is one area of current research (e.g. Ledford and Tawn 1996, 1997, Coles and Pauli 2002).

In summary, extreme value theory is a very powerful, mathematically justified mechanism for making inference on extreme levels of a process. It is capable of evaluating the likelihood of future extreme values occurring beyond both the levels and the time-span of the observed data. To have some belief in the outcome, one is required to have belief in the underlying assumptions: that there is sufficient data for the limiting asymptotic models to be valid; that any form of dependence in the data has been adequately modelled; that for predictive purposes the future state of the model (e.g.

incorporating estimated trends, or explicitly modelled system change-points) is known or estimated. Violation of any of these principles may produce unreliable inference.

An additional assumption is that the statistical framework adopted for estimation and prediction is adequate. Coles and Pericchi (2003) and Sisson *et al* (2006) argue that the statistical framework of choice should be Bayesian. We now examine the Bayesian approach to inference, and evaluate the benefits it brings to the statistical analysis of extremes.

5.3 The Bayesian perspective

Bayesian inference offers a competitive alternative to "classical" statistical methods, and its use has been growing rapidly over many disciplines in the last 20—25 years. The essential difference between the two inferential frameworks is that Bayesian inference regards the parameters as uncertain and hence they have a probability distribution in the parameter space. Contrast this to the classical view that there is a single true value for each parameter.

Accordingly, joint information about the parameters, θ, and data, x, is encoded in distributional form

$$\begin{aligned}\pi(\theta, x) &= \pi(\theta \mid x)\pi(x) \\ &= \pi(x \mid \theta)\pi(\theta).\end{aligned}$$

Rearranging this gives

$$\pi(\theta \mid x) = \pi(x \mid \theta)\pi(\theta) / \pi(x).$$

That is, the distribution of the parameters having observed the data $\pi(\theta|x)$, is proportional to the distribution of the parameter before observing the data $\pi(\theta)$, multiplied by the model $\pi(x|\theta)$, often the likelihood. For observed data, the term $\pi(x)$ is just a constant and may often be ignored. Put another way, given one's initial belief about the model parameters through the prior distribution $\pi(\theta)$, one is able to update these beliefs to obtain the posterior distribution having observed the data $\pi(\theta|x)$ by considering the product of model and prior.

Once the posterior distribution is known, all information about the model can be derived. For example, a predictive distribution $h(y|x)$ of some future value of the process having observed the data x, may be derived through the integration

$$h(y \mid x) = \int h(y \mid \theta)\pi(\theta \mid x)\partial\theta.$$

Thus the predictive distribution averages out the uncertainty inherent in the value of the model parameters. Accordingly Bayesian prediction is less sensitive to the estimated parameter value in comparison to classical (e.g. maximum likelihood) inference, where only a single point estimate of θ is obtained and used. In certain situations this may have huge implications – see the Section on Bayesian extreme value theory below.

The prior distribution is then an opportunity to incorporate expert opinion on the state of the model parameters into the analysis. Deriving the prior distributional form from such experts, known as prior elicitation, can require careful thought, however (O'Hagan et al 2006). Priors incorporating

expert opinion are known as subjective priors. Often an objective prior is adopted so that a benchmark analysis may be performed, free of subjective bias. An objective prior expresses vague or general information about a parameter. Objective priors have been developed for broad classes of models. In data poor situations, the prior will have a strong influence on the form of the posterior, and so the form of the prior is highly important. When there is plenty of data, the data will dominate and the influence of the prior will be minimal, so there is less need for precision. The exception to this is where the prior is overwhelming (e.g. a point mass on some parameter value).

In practice, one is required to perform integrals of the above form in order to evaluate posterior predictive quantities of interest. Aside from trivial cases, analytical integration is untenable and so numerical approximation is needed. Suppose we are able to draw samples $\theta_1, \theta_2, \ldots, \theta_N$ directly from the posterior distribution $\pi(\theta|x)$. We may then use the Monte Carlo approximation

$$\int h(y \mid \theta)\pi(\theta \mid x)\partial\theta = E\left[h(y \mid \theta)\right] \approx \frac{1}{N}\sum_{i=1}^{N} h(y \mid \theta_i).$$

to estimate any integral of interest. The summation and integral are equivalent as $N \to \infty$.

Drawing samples directly from the posterior is not necessarily trivial, and generic algorithmic procedures are the subject of much current research. A number of popular methods include the Gibbs and Metropolis-Hastings samplers (e.g. Gamerman and Lopes, 2006).

5.4 Bayesian Extreme Value Theory

Bayesian inference is particularly suited to extreme value theory (Coles and Powell, 1996). The requirement of prior specification means that the natural scarcity of extreme data may be supplemented through an informative prior formulation from a subject matter expert (e.g. Coles and Tawn, 1996). The Bayesian approach naturally incorporates parameter uncertainty in a probabilistic framework, which is particularly useful for predictive inference as parameter uncertainty can be integrated out. As an extension of this concept, Bayesian inference may also incorporate *model* uncertainty into the analysis. This may generate posterior model probabilities on competing models, if this is of interest, or can integrate over each model (and the parameters within each model) in making predictive inference.

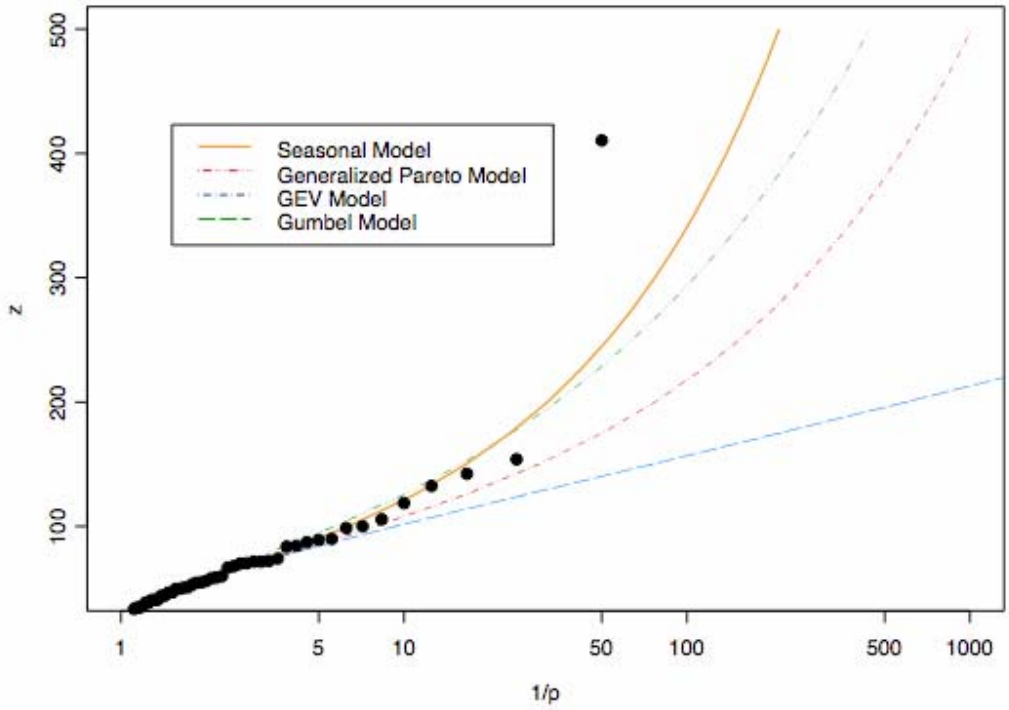

Figure 6: Bayesian predictive distributions for various models fitted to the Venezuelan rainfall data (without the 1999 event). Points correspond to empirical estimates *[i/(n+1), $M_n^{(i)}$]*, i=1,…,50, where $M_n^{(i)}$ is the *i*-th largest block maxima. (Source: Sisson *et al*, 2006).

As an illustration, we return to the Vargas Tragedy study. Figure 6 displays Bayesian predictive return level plots for four different models. The lower two curves denote the Gumbel and GEV predictive densities based on annual maxima data. The top two curves represent generalised Pareto threshold exceedance models, corresponding to a two-season rainfall structure (top) and a homogeneous model (lower). The Bayesian homogeneous Pareto model predicts the return period of the observed 1999 to be 260 years. The effect of integrating out parameter uncertainty (rather than conditioning on fixed maximum likelihood estimates) is clear when comparing to the classical predictor return level of 752 years.

We previously discussed that the meteorology of the Caribbean can be broadly classified into "wet" and "dry" seasons, in which it may be reasonable to assume that the process exhibits different behaviour, and accordingly it should be modelled by different extreme value parameters. Exactly where the change point between the two seasons occurs is unknown. However, under the Bayesian framework, we can introduce two new parameters that specify the boundaries of one season, and let these be estimated by the data (Coles and Pericchi, 2003). Each possible combination of seasonal change points represents a different model as data are then allocated to different seasons. Posterior distributions on the likely location of the seasonal change points are immediately available. Incorporating this parameter/model uncertainty into the predictive process gives the "seasonal" generalised Pareto model in Figure 6 Under this more realistic model, the 1999 event is even more likely, with a return period of 131 years.

In a second analysis, Sisson et al (2003) did not model the daily rainfall data directly, but rather a series derived from the 3-day aggregate rainfall data set. Their data then corresponded to the maximum observed 3-day "storms." Given that the December 1999 event was immediately

flanked by two additional days of extreme rainfall, this is probably a more realistic dataset to consider. Fitting the Bayesian seasonal generalised Pareto model as the model of choice yielded a return period of 308 years for the whole 3-day storm. Wieczorek et al (2001) revealed that the mudslides of the 1999 event exposed very large riverbed stones in a region far from their natural location. From their size and location, they could only have been moved there by a rainfall/mudslide event in the past at least as severe as the December 1999 event. Based on geological and sediment analyses, Wieczorek et al (2001) concluded that the exposed event occurred within the last 500 years or so. In view of this, the estimated return period of the 1999 event, 308 years, seems quite realistic, thereby giving at least weak confirmation of the analysis with an independent source of evidence (while emphasising again the sensitivity of results to individual data points).

Sisson et al (2006) took this modelling a little further in two ways. Firstly, they permitted the number of seasons itself to be an unknown number to be estimated. In this setting the posterior model probabilities gave overwhelming support to a two-seasonal model.

In another example of Bayesian inference allowing the practitioner to avoid specifying a single model, Coles and Pauli (2002) developed a bivariate extreme value distribution that, under differing parameterisations, incorporated both asymptotic dependence and asymptotic independence. In this analysis, the authors were able to use a Bayesian approach to weight the two dependence classes, and thereby average over their uncertainty as to the true asymptotic state.

In the Vargas Tragedy analyses, the parameter prior distributions were specified as independent "vague" priors for each parameter (e.g. a Gaussian distribution with a large variance), in an attempt to be uninformative in order to provide an objective analysis. Objective priors permit benchmark analyses without subjective bias, a valid criticism in subjective prior specifications. However it is currently unclear how to specify a truly objective prior – independent vague parameter priors are unlikely to provide this, as they are dependent on the model parameterisation. This is a subject of current research.

A number of useful prior formulations have been developed. The first allows subject matter experts to provide their opinion on the return periods/levels of extreme events (Coles and Tawn, 1996). These opinions, effectively treated as priors on extreme quantiles of the process, may be transformed into priors on the extreme value parameters, from which the Bayesian inference proceeds as before. Thus, the awkwardness of the extreme value model parameterisation is avoided, in favour of more conceptually accessible quantities.

Botollo et al (2003) illustrate the higher-level structure of a hierarchical prior. Assume that the observed data may be partitioned into different data "types" e.g. different categories in insurance loss data. One approach to modelling is to fit different extreme value models to each data type independently. With few data observations in each model, parameters will be poorly estimated. An alternative is to pool all available data and estimate a single set of parameters. This will improve parameter estimate variability, but will also ignore the varying data types. The hierarchical model provides a compromise between the two. Firstly consider fitting different extreme value distributions to each data type. Now state that all the (say) tail shape parameters are related, in that they all derive from some common distribution. This common distribution induces dependence between the tail shape parameters, thereby permitting the information within each data type to help estimate the parameters of all other data types. This is a useful representation for structured data, and may be combined with (say) the quantile prior approach of Coles and Tawn (1996).

There is a greater computation requirement for Bayesian inference over maximum likelihood estimation. However given even a moderately powerful desktop or laptop, many routine Bayesian

analyses can be performed in minutes. Exceptions to this can include poorly written code, or highly complex models with strong parameter dependencies. Some current research has developed Bayesian inference procedures that admit inference under models with analytically or computationally intractable likelihoods (Bortot et al 2007, Sisson et al 2007). Such methods require repeated simulation of datasets from the model in lieu of likelihood evaluation, and have natural application in extreme value modelling (Bortot et al 2007) and extreme risk estimation (Peters and Sisson, 2006). These procedures are very highly computationally intensive, and depending on the nature of the problem, the time to perform the calculations can be measured in hours, days or even weeks.

In summary, the power of the Bayesian approach to inference makes it highly competitive with, and often superior to classical approaches to extreme value theory. On the down side, while the incorporation of prior information into the analysis is a powerful benefit, it is not always clear on the best way this might be achieved, and the computation requirements can sometimes be high. Overall though, the benefits of the Bayesian approach probably outweigh its disadvantages.

5.5 Robustness: imprecise probabilities, sensitivity analysis, InfoGap

In traditional statistics, probabilities are based on large sets of data or on physical considerations and are thus quite precise numbers – for example, the bias of a coin is a definite number that may be approximated very precisely by observing many throws of the coin. That is not the case in areas such as extreme risk evaluation or criminal trials, where small data sets and opinion combine to produce high or low probabilities, but where any attempt to impose numerical precision on the probabilities results in (potentially dangerous) distortion. The law, for example, has solidly resisted quantifying the criminal standard of "proof beyond reasonable doubt": any attempt to lay it down as a precise number will not actually lead to any consistency in decision-making, as it is impossible to determine a numerical probability that a defendant is guilty on the evidence (even if it is clear that he is guilty "quite certainly" or "only on the balance of probabilities"). (Franklin, 2006)

There is always some psychological resistance to dealing explicitly with imprecision in probabilities. Surely a probability is already a measure of uncertainty, so dealing with "uncertainty in uncertainty" or "probabilities of probabilities" is over-elaborate and too confusing in practice? That is not correct. For the brain and natural language, it is imprecision that is natural and easy and precision unnatural and costly. The ubiquity of fuzzy language in discussing probability, such as "extreme risk", "quite likely", "a remote chance", is a sign that people are comfortable with imprecision and find it adequate in representing their ideas on probability.

There are four ways of dealing with imprecision in probabilities. All have value and one (or more) can be chosen according to the pragmatic needs of the problem, such as how much the imprecision matters in the decision to be reached. In increasing order of sophistication they are:

- Keeping to fuzzy natural language and studying its grounding in numerical probabilities
- Restricting numerical probabilities to one significant figure
- Representing imprecise probabilities in some simple way such as by probability bounds or triangular distributions and using them to conduct sensitivity analyses
- Using InfoGap theory to study directly the robustness of decisions to imprecision in the probabilities

It is beyond the scope of this report to discuss these in detail, but we comment briefly on what is achievable by each method.

People certainly operate naturally with fuzzy probabilities such as "very likely" and prefer to use them for reporting so as to avoid precision in which they do not believe (or to "maintain deniability") (e.g. Olson and Budescu, 1998). Their prevalence in scorecards as well as in informal risk discussions makes it imperative to determine whether there is consistency among different people's understanding – do all parties to a discussion mean the same numerical range by "extreme risk", for example? Biosecurity Australia's apple risk analysis (2006, p. 43) uses the following table to translate probability words into numerical probability ranges:

Table 1: Nomenclature for qualitative likelihoods, corresponding semi-quantitative probability intervals (Source: Biosecurity Australia, 2006, Table 12.

Likelihood	**Qualitative descriptors**	**Probability interval**
High	The event would be very likely to occur	$0.7 \rightarrow 1$
Moderate	The event would occur with an even probability	$0.3 \rightarrow 0.7$
Low	The event would be unlikely to occur	$5 \times 10^{-2} \rightarrow 0.3$
Very low	The event would be very unlikely to occur	$10^{-3} \rightarrow 5 \times 10^{-2}$
Extremely low	The event would be extremely unlikely to occur	$10^{-6} \rightarrow 10^{-3}$
Negligible	The event would almost certainly not occur	$0 \rightarrow 10^{-6}$

Although by and large reasonable, such translation tables encounter the problem that the fuzzy words of natural language are in general highly context-sensitive (a small elephant is bigger than a big mosquito because being a small elephant is being small for an elephant – with reference, that is, to the mean in the appropriate, context-dependent, reference class). Research shows that there is some consistency in how subjects translate verbal to numerical probabilities, but some individual variability (Wallsten, Budescu and Zwick, 1993; further in Caponecchia, 2007, section 4) and sensitivity to context. (Fox and Irwin, 1998) The upshot is that verbal probabilities and translation tables to numerical probabilities are only usable in the elicitation and communication of risk judgments with extreme care. One will have to check very carefully whether experts and non-experts mean the same by such expressions as "extreme risk" and whether they mean the same in one risk setting as in another – and even if an organisation achieves standardisation internally, it has little control over the use of words by its stakeholders or audience. As an illustration, one may compare the translation table above with one in Burgman (2005, p. 77), taken from a paper on geological risk in petroleum exploration:

Table 2: A Kent scale used to evaluate geological risk of petroleum exploration prospects (Source: Burgman 2005, p. 77, from a 1998 paper by P. Watson)

Expression	**Synonyms**	**Percent probability**

Proven	True	98-100
Virtually certain	Convinced	90-98
Highly probable	Strongly believe, highly likely	75-90
Likely	Probably true, chances are good	60-75
Even chance	Slightly better, slightly less than even	40-60
Probably not true	Unlikely, chances are poor	20-40
Possible but very doubtful	A slight chance, very unlikely	2-20
Proven untrue	Impossible	0-2

Thus a 1% probability counts merely as "low" in Biosecurity Australia's table but as "proven untrue" in the petroleum exploration table. That is natural in the different contexts, since BA's table needs to differentiate between very low probabilities while petroleum exploration is more concerned with high probabilities (of striking oil). So, obviously the numerical meaning of natural-language probability expressions cannot be simply taken over from one context to another. A particular contextual matter, often commented on in bank operational risk, is the need for clarity in the time period to which the risk refers: a loss that has one-in-a-thousand chance of happening in a day is quite likely to happen in a year. It is much easier to clarify such matters with numbers than with words. (Relevance to Biosecurity reviewed in McCarthy et al., 2007)

Reporting numerical probabilities to only one significant figure (for example, 0.4 or $2 \times 10\text{-}6$ but not 0.41 or $2.4 \times 10\text{-}6$) is a common practice but one usually done unreflectively. (But see Phillips and LaPole 2003 for some efforts at using restrictions on significant figures to report uncertainty.) It relies on the fact that it is quite rare for decisions to be sensitive to differences in probability of less than one significant figure: a chance of three-in-a-million may warrant higher precautions than a chance of one-in-a-million, but it is hardly likely that one will take much notice of the difference between one-in-a-million and 1.3-in-a-million, even if one is convinced that the difference in the chances is real and not just measurement error. (But see Caponecchia, 2007, section 4 for the salience of relative extreme risks such as "the risk has increased by 30%.")

To report a probability to one significant figure is to make implicit use of an interval-valued probability, since by "probability 0.4" one means "probability in the range 0.35 to 0.45". It is possible to make more explicit use of bounded probabilities (Walley, 1991; Ferson et al., 2004). One may either use the interval, perhaps with the implicit assumption of a uniform distribution between the bounds, or use a triangular distribution, with a "midpoint" that is the best estimate of the probability and a (not necessarily symmetric) range of uncertainty on either side (some sceptical comment in Burgman, 2005, pp. 78-9). The use of interval probabilities encourages sensitivity analyses, since it is easy to calculate what would happen if the ends of the ranges were used. For example, Biosecurity Australia's apple risk analysis (pp. 114-5) concludes that "a maximum value three times larger than the value agreed by the IRA team for every exposure value results in an overall risk with the recommended risk that just exceeds Australia's appropriate level of protection." There is however a conceptual difficulty with the idea of interval-valued or bounded or triangular probabilities – the ends themselves appear to be precise but of course really are not (since if the probability itself is not known precisely, it is hardly likely that bounds on it will be), and normally harder to estimate than the central value. That translates into a practical difficulty for any sensitivity analysis based on the bounded probabilities, since one has little confidence that the probability is really bounded between the values stated.

The Info-gap decision theory of Yakov Ben-Haim (Ben-Haim, 2006; Regan et al., 2005) also deals with the sensitivity of decisions to uncertainty in the inputs to a problem, including the probabilities. After the input-output function of the problem has been modeled, info-gap theory

explores the sensitivity of the output to the full range of uncertainty in the inputs (rather than merely looking at the change in the outputs to several possible perturbations of the input as sensitivity analyses normally do). A decision maker can thus impose a range of acceptable possibilities for the output and make a map of what range of inputs would lead to outputs within the acceptable range.

All these techniques will be needed in the suite of methods in the toolkit of the analyst of extreme risk, where in the nature of the case there is normally considerable doubt as to the numerical value of the risks involved. One must report the uncertainty in the risk and examine how sensitive decision-making is to that uncertainty.

5.6 Strengths of commonsense reasoning under uncertainty

We summarise from Franklin (2001, pp. 324-5), some of the reasons for believing that the untutored brain is an excellent organ for probabilistic reasoning, in many circumstances (further references are available therein). Those circumstances are many, but the extreme risk situation is particularly important because of the lack of directly relevant data and hence the need to rely on intuition and analogy:

> That the vast majority of probabilistic inferences are unconscious is obvious from considering animals. For it is not just the human environment that is uncertain, but the animal one in general. To find a mechanism capable of performing probabilistic inference (as distinct from talking about it), one need look no further than the brain of the rat, which generates behaviour acutely sensitive to small changes in the probability of the results of that behaviour. Naturally so, since the life of animals is a constant balance between coping adequately with risk, or dying. Foraging, fighting and fleeing are activities where animal risk evaluations are especially evident; in general, the combining of uncertain information from many sources is of the essence of brainpower in the higher animals. Some further light on what the brain does is cast by the simple "artificial neural nets", whose behaviour after training on noisy data can be interpreted as implicit estimates of probabilities. These animal and machine studies confirm in the most direct possible way that to behave probabilistically, it is not necessary to have anything like explicit estimates of probabilities or ways of talking about them.

The human species inherited the mammal brain, with these abilities already loaded and in automatic use. In human life, the only certainties, proverbially, are death and taxes, and of these, the time and amount, respectively, are quantities rarely known. The "Iceman" discovered in the Alps in 1991 was certainly one who took a calculated risk. It is clear at least in principle how evolutionary pressures select for rational techniques of risk management; the roads continue to select against those whose evaluation of risk is below par. There are psychological studies which show how much of cognition generally is "intuitive statistics". Even such a basic operation as the discrimination of stimuli (for example, in deciding whether two sounds are the same pitch or not) is a probabilistic process of extracting a signal from a noisy background. And perceiving and remembering both involve unconscious testing of hypotheses on the basis of imperfect correlations. Very nearly all uncertain inference is unconscious, performed at the sub-symbolic level by the neural net architecture of the brain …

Human subjective assessments of risk expressed in words are reasonably accurate in many circumstances. Indeed, in business forecasting of such movable quantities as stock prices, human "judgmental forecasting" is still generally comparable to the best statistical methods (and it is possible to say which statistical methods it resembles). Child development studies show the gradual development of reasonable risk estimates in words. However, there are some strange features of the brain's implementation of probabilistic reasoning that result in systematic deviations from rationality. Estimates are age-specific, for example. They rely on mental "models"

or "prototypes" in some sense, leading to such problems as overconfidence in estimates and over-sensitiveness to the order of presentation of evidence. The relationship between words used to express risk estimates and what the brain is really doing is a problematic one: driver behaviour is related more closely to objective risk than to stated risk, and learning of relative frequencies is often best in the absence of clumsy attempts to make conscious statements about them. On the other hand, having a theory can help bring order into data and correct mistakes in it. Of interest in connection with the relation between numerical and purely linguistic expressions are experiments that show a reasonable consistency between phrases like "very likely" and the "fuzzy" numerical estimates of probabilities that they mean. Since people often use words in preference to numbers in discussing risks, to avoid committing themselves to accuracy they do not possess, it is fortunate that probabilistic words are reasonably well calibrated.

These considerations suggest this important conclusion, which is central to the point of view of this report:

> *It is reasonable to give human intuition the "last word" in risk assessment, while at the same time trying to use formal statistical methods as a kind of prosthesis to supplement its known weaknesses.*

A problem where the superiority of human intuition over formal methods is especially evident – and one very relevant to extreme risks – is the "reference class problem" (also called in artificial intelligence "multiple inheritance"). The most basic evidence for probabilities in an individual case is observation of a relative frequency (in a class of which the case is a member). For example, the probability that Tex is rich, given that Tex is a Texan and 90% of Texans are rich, is 0.9. But typically, a case is a member of very many classes, in which relative frequencies vary. And there is no useful theory explaining how to combine the probabilities arising from the different "reference" classes. For example, if the evidence is that Tex is a Texan philosopher, that 90% of Texans are rich and 10% of philosophers are rich, then it is impossible to say how to combine these two numbers to achieve a numerical probability that Tex is rich, on the given evidence. (Hájek, 2006) The problem has caused a great deal of trouble in, for example, the law of evidence, where there is often evidence of different classes but it is of dubious legal relevance (Colyvan *et al*, 2001; Tillers, 2005), and in attempts to construct medical diagnosis expert systems, where combining evidence from different symptoms is essential but how to do it is theoretically poorly understood. (See also Caponecchia, 2007, section 4 for its relevance to communicating probabilities.)

Yet humans are very good at combining different kinds of evidence. Where they have an advantage over formal methods is that they can learn from long experience the comparative relevance of different reference classes. For example, they can learn enough about being Texan, being a philosopher and being rich to have some sense of whether being Texan or being a philosopher is more likely to be relevant to being rich. The vocabulary of natural languages is already attuned to naming concepts that are relevant to living, that is, are positively relevant in probabilistic inferences; which of them are most relevant to a particular inference is something that itself can be learned – but only over a long period, and in the context of very many other concepts.

That wide base of experience and the resultant tuning of concepts is not something that should be put aside when it comes to extrapolating from experience when evaluating extreme risks. On the contrary, is it a foundation that must be built on. It is the wide base of analogous cases that can compensate for the lack of data of directly relevant cases that is a feature of extreme risk analysis.

An "advocacy" model, we argue in section 6 below, is ideal for taking advantage of the strengths of innate human probabilistic reasoning, since on the one hand it gives human intuition the last word in combining the evidence to reach a final conclusion, but on the other hand gives maximum space for the use of any technical methods on the way.

5.7 Psychological evidence on strengths and weaknesses of expert opinion

In extreme risk evaluation, especially when performed by a committee or formal process, the intuitions relied on will often be those of experts. Much has been written on the weaknesses of expert opinion, rather less on its strengths. The systematic errors of experts are well documented – their overconfidence, inability to know where their expertise ends, sensitivity to framing effects, confusion over base rates and conditional probabilities, and so on (survey in Burgman 2005, sections 4.5-4.6). The recent study of Tetlock on medium-term political judgement also supports a very pessimistic view of the quality of expert opinion, especially when it is expressed confidently. (Tetlock, 2005) Those consistent results certainly imply that one should not trust experts in general.

One could however suspect a certain negative bias in the reporting of expert opinion, since it is common experience that in certain areas experts can perform well, and much better than non-experts. Of day-old chicks labeled female by expert chicken-sexers, 98% grow up to lay eggs, though they look exactly the same as male chicks to the untrained eye. (Martin, 1994; Biederman and Shiffrar, 1987) Surgeons often find that patients have the conditions diagnosed by physicians, the voting in elections is rarely far from that predicted by experts, biologists are much better at identifying species than non-experts, the predictions of the Manhattan Project physicists on the size of the atomic bomb explosions were close to the truth, and there is reasonable correlation between exam markers (at least in the more technical disciplines). (e.g. Caryl, 1999) Even the much-maligned weather forecasts (which come from expert opinion on the basis of computer extrapolations of data) are of reasonable quality – for the rather variable British weather, the UK Met Office achieves a little over 80% accuracy for its next-day maximum temperature forecasts (within 2°C), which is well above what is possible with simplistic methods like persistence forecasting (Met Office, 2006).

There is also reason to believe that some of the heuristics on judgement that are errors in general work quite well in the contexts where they are normally used. That may explain away some of the findings on human irrationality in probabilistic reasoning, though far from all. (Gigerenzer and Todd, 1999)

Further, Tetlock's careful and extensive study in the notoriously unpredictable area of political judgement found that some experts were better than others. The difference between better and worse ones was not what they thought (for example, left versus right, or "Doomster" versus "Boomster"). It was rather a difference when it came to cognitive style. Experts who were (self-rated) "Hedgehogs" – who applied a one-size-fits-all pet theory to all cases and stuck to it – were less successful at prediction that "Foxes" – who know "many little things", hedge their bets when they should, change their minds in response to evidence, and are less inclined to invoke excuses when they get it wrong. (Tetlock, 2005, ch. 3)

Plainly, there is a need not so much for scepticism in general about expert opinion, but for an understanding of where experts can be trusted and how to improve the performance both of those that are not trustworthy and those who are. From the examples, it is clear that some sort of pressure has to be exerted on the experts to punish bad judgment and reward success. The media

political pundits in Tetlock's study, for example, received nothing but positive reinforcement from being confidently and consistently wrong on television.

The best tutor is feedback from experience. It is certainly possible for experience to improve performance, at least if it is based on enough data for there to be significant estimation of whether the experts' probability estimates were reasonably accurate. (Benson and Önkal, 1992)

Unfortunately, the area of extreme risks is one where such feedback is not available, since in the nature of the case the events in question occur very rarely (and the non-occurrence of the events will merely tend to reinforce the optimism of experts).

That is why we suggest replacing the ideal feedback of real experience with the "virtual" feedback provided by the advocacy model – scrutiny of experts' assessments by a neutral panel of "judges", informed by the scenarios and reasoning put forward by possibly hostile stakeholders.

Justification and accountability improve judgement (Hagafors and Brehmer, 1983). Lee et al. (1999) write:

> Accountability (or the need to justify one's judgments and decisions to others (Tetlock and Tetlock)) motivates complex and effortful information processing and encourages decision makers to engage in cognitive activities that promote (or at least seem to promote) high-quality decision making. Accountability also increases decision makers' concern about committing potentially costly judgmental errors (Kruglanski and Kruglanski) and encourages decision makers to engage in more analytic and less intuitive cognitive processes (Hagafors and Brehmer). Moreover, accountability has been found to influence persuasion by causing accountable message recipients to hold flexible, moderate positions on an issue when an unknown audience will evaluate their position (Cialdini and Leippe).

But true accountability requires that the person to be held accountable fears his judges. He must be motivated by anxiety as to what their views might be. As Tetlock (1983) puts it, "Findings suggest that accountability leads to more complex information processing only when people do not have the cognitively lazy option of simply expressing views similar to those of the individual to whom they feel accountable." Facing an audience of known views simply leads most people, especially the socially anxious, to move their estimates towards those of the audience. That is not an option where the audience's views are either not known or, as in the advocacy model, are known to be varied. It is no use asking for the justifications of the decision later, either, as that merely leads to the generation of reasons why the original decision was right all along and sometimes to more extreme positions (Lerner and Tetlock, 1999). Also significant is the distinction between outcome accountability (rewards for getting the decision right) versus process accountability (rewards for showing that one's decision process was justified); it appears in general that process accountability leads to more productive effort (Simonson and Staw, 1992; Siegel-Jacobs and Yates, 1996). Very relevantly for the advocacy model, the authority to which justification is submitted must be perceived as legitimate and itself having the expertise to evaluate the justification. Lerner and Tetlock (1999) summarise:

> Self-critical and effortful thinking is most likely to be activated when decision makers learn prior to forming any opinions that they will be accountable to an audience (a) whose views are unknown, (b) who is interested in accuracy, (c) who is interested in processes rather than specific outcomes, (d) who is reasonably well-informed, and (e) who has a legitimate reason for inquiring into the reasons behind participants' judgments. But even among studies that incorporate this very specific kind of accountability, effects are highly variable across judgment tasks and dependent variables …

They add an impressive table of the cognitive biases that are found to be attenuated by accountability, including hastiness in judgment, lack of awareness of one's own judgment processes, overconfidence, over-sensitivity to the order in which information appears, pursuing sunk costs and groupthink (but accountability was not helpful with some of the other classic cognitive biases, such as insensitivity to base rates and insensitivity to sample size).

So accountability is not perfect as a device for improving probabilistic thinking, but the psychological findings on its advantages provide a solid theoretical foundation for believing that an advocacy model will have benefits.

6. Adversary and advocacy models of public judgements

We now provide some more details on the advocacy model and the reasons for favouring it. We argue that a model of risk evaluation similar to those used in the Basel II compliance regime for bank operational risks and in Biosecurity Australia's Import Risk Assessments is a good approximation to best practice in the area, as it permits the diversity of relevant evidence to be presented and soundly evaluated.

The essential idea is that a well-tried method of finding what is wrong with someone's product (for example, their risk analysis) is to have an *adversary* look for its weaknesses. An independent or antagonistic consultant can discover what the creator of the product will never see in his pet project.

Such methods have proven useful in, for example, software testing, where a team independent of the developers of software is employed solely to find bugs (Myers, 2004, p. 15), or in computer security where one can employ teams of hackers to conduct tests of the vulnerability of one's system to penetration. (Klevinsky *et al.*, 2002) These methods are especially applicable to software because of the possibility of non-destructive testing. Though less easily applicable elsewhere, the success of adversarial approaches to finding "unexpected" risks is still relevant. As a writer on software testing puts it, "removing wallpaper is not easy, but it is almost unbearably depressing if it was your hands that hung the paper in the first place. Similarly, most programmers cannot effectively test their own programs because they cannot bring themselves to shift mental gears to attempt to expose errors." Unexpected risks can best be found by someone who wants to find them, that is, by an adversary of the system's makers or guardians (operating in an overall environment where both sides will in the end be listened to).

The most developed and best-known use of adversaries in is the system of legal trials in Anglo-American law. The two sides are represented by counsel who have wide discretion to put their cases as they think fit, though the judge moderates the process to some degree. The final decision is made either by a jury which acts as a "black-box" fact evaluator which does not need to give any reasons for its decision, or by a judge or panel of judges who deliver reasons for their judgment. The model encourages effort to present a rational case that will be as convincing as possible, while leaving the final decision to disinterested parties. It is a problematic model where there is a need to evaluate technical complexity, for example in medical negligence or complicated financial cases where the evidence may be beyond the understanding of juries or legally-trained professionals. It also tends to be impervious to discoveries of systematic errors, for example, to psychological evidence on the low reliability of eyewitness identification evidence. (Wells and Olson, 2003)

Compliance regimes that regulate industries and ensure adherence to standards have come to adopt what might be called an *"advocacy model"*, which has some of the qualities and advantages of a trial but also some fundamental differences. Typically, a compliance body, such as the Australian Prudential Regulation Authority (APRA), Biosecurity Australia (earlier the Australian Quarantine and Inspection Service)'s, or the Aged Care Standards and Accreditation Agency, is a permanent authority that oversees the compliance with published standards by the players in the regulated industry. A body seeking a determination from the authority (for example that import of New Zealand apples should be allowed or that an aged care home should be allowed to continue to operate) submits extensive documentation, typically about risk measurement and mitigation. The documentation may be prepared by specialists, sometimes outside consultants who work with insiders on understanding the body's operations in detail. The documentation is examined by experts from the regulator, who can and typically do demand further documentation on matters they consider possibly suspicious. After some rounds of queries and possibly inspections, a decision is reached. A generally co-operative attitude is maintained between the regulator and

body regulated, except in extreme cases. The degree of confidentiality of the process varies; in cases of accreditation like APRA or in aged care, confidentiality is normal during the process to encourage honesty in sharing of data, but a public report is issued at the end of the process. The regulator is responsible to some outside body such as Parliament, and is also subject to embarrassment if a risk it has overlooked appears as a media scandal involving losses of millions in rogue trading or a cluster of deaths in an aged care facility.

The case studies described above in which an advocacy model was used in one form or another (bank operational risk, the Ernst & Young case, and Biosecurity Australia's apple risk analysis) show, we believe, how the model has acted to force the parties involved to work hard to identify and quantify all the risks and to honestly lay them out for inspection. A close study of what those cases have in common is the best way forward in creating on overall framework for best practice in the analysis of extreme risks. In planning the implementation of an advocacy model, a number of administrative issues arise such as the exact locus of final judgment, security of tenure, financial arrangements for tribunals, stakeholders and consultants, and the like. These are important issues in ensuring the independence and credibility of the decisions reached by the process – indeed, there are a few cases of spectacular failures of semi-judicial tribunals from problems in these areas. (Franklin, 2007) Research on these questions needs to draw on expertise in public administration and corporate governance.

7. Recommendations

In the light of what we have found we make these recommendations to teams involved in the evaluation of extreme risks and to ACERA. The essential thinking behind these recommendations lies in our earlier conclusion that we repeat here:

It is reasonable to give human intuition the "last word" in risk assessment, while at the same time trying to use formal statistical methods as a kind of prosthesis to supplement its known weaknesses.

The recommendations are:

Education of extreme risk evaluators in Extreme Value Theory, basic Bayesian theory and imprecision/robustness concepts

These technical methods have proved to have application in certain areas in evaluating and communicating extreme risks. They are not panaceas, but have advantages over older standard statistical methods in providing the necessary flexibility to deal with difficult extrapolations beyond the range of existing data. In particular they guard against dangerous illusions of false precision in extreme risk estimates and underestimates of tail probabilities. A team involved in extreme risk evaluation should have some general understanding of the scope and limits of such methods and be should be able to call on experts in those methods when the case indicates the need for them.

Education of statisticians involved in extreme risk evaluation in more qualitative legal perspectives, outlier detection, data mining methods of fraud detection, and methods of causal chain analysis

The strongly quantitative style of education in statistics, valuable as it is, can lead to a neglect of the more qualitative, logical and causal perspectives needed to understand data intelligently. That is especially so in extreme risk analysis, where there is a lack of large data sets to ground solidly quantitative conclusions, and correspondingly a need to supplement the data with outside information and with argument on individual cases.

Psychological study of the full advocacy model

Although we have provided reasons for thinking an advocacy model will lead to better risk analyses, and have described case studies where some approximation to an advocacy model is used, there needs to be much more rigorous research into whether it actually works. We recommend ACERA fund and oversee such research, employing advisers skilled in methods of psychological experiment. They will have the skill to devise experiments with proper controls to determine whether an advocacy model really does lead to better identification, evaluation and communication of extreme risks.

Investigation of newer statistical methods such as data-mining, spatial and spatiotemporal methods, capture-recapture methods and prediction markets

A number of new statistical (or marginally statistical) methods have emerged in recent years, which p*rima facie* have good possibilities for application to extreme risk analysis. We recommend that ACERA fund and oversee investigations into them. Data mining has shown the possibilities of extracting value from large data sets and has proved its value to business in understanding customer behaviour; its applications to fraud detection are especially relevant to extreme risks. Many risks are spatially variable (for example the chance of transfer of fire blight from discarded apple cores to hosts is very dependent on the spatial distributions of cores, hosts and vectors), and

the general inadequacy of coverage of the space by data means there is (or should be) strong interaction between the methods of spatial statistics and extreme risk analysis. The capture-recapture methods recently used to estimate populations from small samples and to estimate the numbers killed in human rights abuses (e.g. Silva and Ball, 2006, section 5.7.5) extrapolate beyond observations so hold promise of applicability to extreme risks. Prediction markets promise to hold those making predictions accountable for what they say, thus creating a public predictive mechanism more reliable than the "sum of its parts"; disaster prediction is among the leading potential applications.

Use of independent facilitators such as consultants to mediate between risk evaluator and stakeholders

We were impressed with the role of consultants in the Ernst & Young case in mediating between the final risk evaluator (APRA) and the client whose risk analysis had to pass inspection. The possibility of the consultant representing each side to the other over several rounds of negotiation was most valuable in bringing the risk evaluation "up to scratch". Consultants are expensive, but in cases where it is very important to achieve the best possible result, we recommend their use.

Adoption of more transparent attribution policies for authors of texts

The advocacy model relies on the authors of risk analyses "standing behind" their assertions. Accountability requires clear attribution, which is not always the case in reports. For papers and reports by academics and members of the CSIRO and similar organizations, it is clear who the authors are (and which one takes prime responsibility) and it is possible to find websites on each of the authors where their qualifications, list of publications and contact details can be found. That is not normally so with government and especially commercial organizations, where reports, if publicly available at all, are attributed to the organization as a whole or to some large and anonymous group. Even if authors are known, the organization's website does not usually give any information about them. Under those conditions, reports are in effect unattributed and hence accountability is poor. We recommend that where possible government and commercial organizations adopt academic practice in displaying personal websites of their report-writing staff (and former staff) with lists of what they have written.

8. References

Anon (2005), Inside NAB's nightmare, *Risk Management Magazine*, 21 Sept, http://www.riskmanagementmagazine.com.au/articles/25/0c036625.asp

AQIS (Australian Quarantine Inspection Service) (1998), Final Import Risk Analysis of the New Zealand Request for the Access of Apples into Australia, http://www.affa.gov.au/corporate_docs/publications/pdf/market_access/biosecurity/plant/ACF133.pdf

Bank For International Settlements, Basel Committee On Banking Supervision (2002), Sound practices for the Management and Supervision of Operational Risk, Dec 2001, revised July 2002.

Bank For International Settlements, Basel Committee On Banking Supervision (2004), Basel II: International Convergence of Capital Measurement and Capital Standards: a revised framework, June 2004, http://www.bis.org/publ/bcbs107.htm.

Barnett, V. (2004), *Environmental Statistics: Methods and Applications*, Chichester: Wiley.

Bedford, T. and R. Cooke (2001), *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge: Cambridge University Press.

Beirlant J., Y. Goegebeur and J. Teugels (2004), *Statistics of Extremes: Theory and Applications,* Hoboken NJ: Wiley.

Ben-Haim, Y. (2006), *Info-Gap Decision Theory: Decisions under severe uncertainty,* 2nd ed, Oxford: Academic.

Benson, P.G. and D. Önkal (1992), The effects of feedback and training on the performance of probability forecasters, *International Journal of Forecasting* 8, 559-73.

Biederman, I. and M. Shiffrar (1987), Sexing day-old chicks: a case study and expert systems analysis of a difficult perceptual-learning task, *Journal of Experimental Psychology: Learning, Memory, & Cognitio*n 13, 640-45.

Biosecurity Australia (2006), Final Import Risk Analysis Report for Apples from New Zealand, Part B.

Bolton, R.J and D.J. Hand (2002), Statistical fraud detection: a review, *Statistical Science* 17, 235-55.

Bortot P., S.G. Coles and S.A. Sisson (2007), Inference for stereological extremes, *Journal of the American Statistical Association* 102, 84-92.

Bottolo L., G. Consonni, P. Dellaportas and A. Lijoi (2003). Bayesian analysis of extreme values by mixture modeling, *Extremes* 6, 25-47.

Burgman, M. (2005), *Risks and Decisions for Conservation and Environmental Management,* Cambridge: Cambridge University Press.

Caponecchia, C. (2007) Strategies for the Effective Communication of Probabilities, ACERA report.

Caryl, P.G. (1999), Psychology examiners re-examined: A 5-year perspective, *Studies in Higher Education* 24, 61-74.

Clemen, R.T. and R.L. Winkler (1999), Combining probability distributions from experts in risk analysis, *Risk Analysis* 19, 187-203.

Coleman, Mr Justice (2000). *Report of the Re-opened Formal Investigation into the Loss of the M. V. Derbyshire*, London: Stationery Office.

Coles S.G. (2001), *An Introduction to Statistical Modelling of Extreme Values*, London: Springer.

Coles S.G. and L.R. Pericchi (2003), Anticipating catastrophes through extreme value modelling, *Applied Statistics* 52, 405-416.

Coles S.G. and E.A. Powell (1996), Bayesian methods in extreme value modelling: A review and new developments, *International Statistical Review* 64, 119-136.

Coles S.G. and J.A. Tawn (1996), A Bayesian analysis of extreme rainfall data, *Applied Statistics* 45, 463-478.

Coles S.G., J. Heffernan and J.A. Tawn (1993), Dependence measures for multivariate extremes, *Extremes* 2, 339-365.

Coles S.G. and F. Pauli (2002), Models and inference for uncertainty in extremal dependence. *Biometrika*, 89, 183-196.

Coles S.G., L.R. Pericchi and S.A. Sisson (2003), A fully probabilistic approach to extreme rainfall modelling, *Journal of Hydrology* 273, 35-50.

Colyvan, M., H.M. Regan and S. Ferson (2001) Is it a crime to belong to a reference class? *Journal of Political Philosophy* 9 (2), 168-181.

Dalal, S.R., E.B. Fowlkes and B. Hoadley (1989), Risk analysis of the Space Shuttle: pre-Challenger prediction of failure, *Journal of the American Statistical Association* 84 (408), 945-57.

De Haan L. and S.I. Resnick (1977), Limit theory for multivariate sample extremes. *Z. Wahrscheinlichkeitstheorie* 40, 317-337.
Earman, J. (1992), *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory,* Cambridge, Mass: MIT Press.
Embrechts, P., ed, (2000), *Extremes and Integrated Risk Management,* London: Risk Waters Group.
Embrechts, P., C. Klüppelberg and T. Mikosch (1997) *Modelling Extremal Events for Insurance and Finance*, New York: Springer.
Ferson, S., R.B. Nelsen, J. Hajagos, D.J. Berleant, J. Zhang, W.W. Tucker, L.R. Ginzburg and W.L. Oberkampf (2004), Dependence in probabilistic modeling, Dempster-Shafer theory, and probability bounds analysis, SANDIA report 2004-3072.
Fox, C.R. and J.R. Irwin (1998), The role of context in the communication of uncertain beliefs, *Basic and Applied Social Psychology* 20, 57-70.
Fox, D.R. (2007), Statistical Methods for Biosecurity Monitoring and Surveillance, I: Control Charting, ACERA Report.
Franklin, J. (2001), *The Science of Conjecture: Evidence and Probability Before Pascal,* Baltimore: Johns Hopkins University Press.
Franklin, J. (2005), Risk-driven global compliance regimes in banking and accounting: the new Law Merchant, *Law, Probability and Risk* 4 (4), 237-50.
Franklin, J. (2006), Case comment: quantification of the `proof beyond reasonable doubt' standard, *Law, Probability and Risk* 5, available at http://lpr.oxfordjournals.org/cgi/content/full/mgl017?ijkey=ny8A8TF7Lp8kezF&keytype=ref
Franklin, J. (2007), International compliance regimes: a public sector without restraints, *Australian Journal of Professional and Applied Ethics*, to appear.
Gamerman D. and H. F. Lopes (2006), *Markov Chain Monte Carlo*, Boca Raton: Chapman and Hall.
González M. and J. R. Córdova (2000). Consideraciones sobre la probabilidad de ocurrencia de lluvias máximas en la zona littoral del norte de Venezuela. Memorias del Seminario Internacional Los Aludes Torrenciales de Diciembre 1999 en Venezuela, Instituto de Mecánica de los Fluidos, Universidad Central de Venezuela, Diciembre 2000.
Gigerenzer, G. and P.M. Todd (1999), *Simple Heuristics That Make Us Smart*, New York: Oxford University Press.
Hagafors, R. and B. Brehmer (1983), Does having to justify one's judgments change the nature of the judgment process? *Organizational Behavior and Human Performance* 31, 223-32.
Hand, D.J. and R.J. Bolton (2004), Pattern discovery and detection: a unified statistical methodology, *Journal of Applied Statistics* 31, 885-924.
Hájek, A. (2006), The reference class problem is your problem too, in B. Brown and F. Lepage, eds, *Truth and Probability: Essays in Honor of Hugues Leblanc*, London: College Publications, available at http://www.sess.smu.edu.sg/events/Paper/hajek1.pdf
Heffernan J. and J.A. Tawn (2003), An extreme value analysis for the investigation in to the sinking of the M. V. Derbyshire, *Applied Statistics* 52, 337-354.
Heffernan J. and J.A. Tawn (2004), Extreme values in the dock, *Significance* 1 (1), 13-17.
Hellin, J., M. Haigh and F. Marks (1999), Rainfall characteristics of hurricane Mitch, *Nature* 399 (27 May), 316.
Hodge, V.J. and J. Austin (2004), A survey of outlier detection methodologies, *Artificial Intelligence Review* 22, 85-126.
Jenkins, T, J. Slade and A. Street (2005), A practitioner's guide to the Advanced Measurement Approach to operational risk under Basel II, Actuaries of Australia Biennial Convention.
King, J.L. (2001), *Operational Risk: Measurement and Modelling*, New York: Wiley.
Klevinsky, T.J., S. Laliberte and A. Gupta (2002), *Hack I.T.: Security Through Penetration Testing*, Boston: Addison-Wesley.
Knight, J. (2005), Underarm bowling and Australia-New Zealand trade, *Australian Review of Public Affairs* 18 July 2005, http://www.australianreview.net/digest/2005/07/knight.html
Kotz S. and S. Nadarajah (2000), *Extreme Value Distributions: Theory and Applications.* London: Imperial College Press.
Koutsoyiannis D. (2004), Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation, *Hydrological Sciences Journal* 49 (4), 575-590.

Lavine, M. (1991), Problems in extrapolation illustrated with Space Shuttle O-ring data, *J. of the American Statistical Association* 86 (416), 919-21.
Leadbetter M.R., Lindgren G. and H. Rootzén (1983), *Extremes and Related Properties of Random Sequences and Series,* New York: Springer Verlag.
Ledford A. and J.A. Tawn (1996), Statistics for near independence in multivariate extreme values, *Biometrika* 83, 169-187.
Ledford A. and J.A. Tawn (1997), Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society* B, 59, 475-499.
Lee, H., P.M. Herr, F.R. Kardes and C. Kim, Effects of choice accountability, issue involvement, and prior knowledge on information acquisition and use, *Journal of Business Research* 45, 75-88.
Lerner, J.S. and and P.E. Tetlock (1999), Accounting for the effects of accountability, P*sychological Bulletin* 125, 255-75.
Lighthall, F.F. (1991), Launching the Space Shuttle *Challenger*: disciplinary deficiencies in the analysis of engineering data, *IEEE Transactions on Engineering Management* 38 (1), 63-74.
Marrison, C. (2002), *Fundamentals of Risk Management*, Boston: McGraw-Hill.
Martin, R.D. (1994), *The Specialist Chick Sexer: A History, World View, Future Prospects*, Bernal Publishing: Melbourne.
McCarthy, M., M. Burgman and I. Gordon (2007), Use of period of trade and trade volume in import risk analysis, ACERA draft review.
Met Office (UK) (2006), Weather Forecast Verification, http://www.metoffice.gov.uk/corporate/verification/city.html
Myers, G.J. (2004), *The Art of Software Testing*, 2nd ed, Hoboken: Wiley.
O'Hagan A, C.E. Buck, A. Daneshkhah, J.E. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley and T. Rakow (2006), *Uncertain Judgements – Eliciting Experts' Probabilities,* London: Wiley.
Olson, M.J. and D.V. Budescu (1998), Patterns of preference for numerical and verbal probabilities, *Journal of Behavioral Decision Making* 10, 117-31.
Otey, M.E., A. Ghoting and S. Parthasarathy (2006), Fast distributed outlier detection in mixed-attribute data sets, *Data Mining and Knowledge Discovery* 12, 203-28.
Peters, G.W. and S.A. Sisson (2006), Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk* 1, 27-50.
Phillips, C.V. and L.M. LaPole (2003), Quantifying errors without random sampling, *BMC Medical Research Methodology* 3, 9.
Phua, C., V. Lee, K. Smith and R. Gayler (2005), Comprehensive survey of data mining-based fraud detection research, *Artificial Intelligence Review*, draft,

Pickands, J. (1981), Multivariate extreme value distributions, In *Proceedings of the 43rd Session of the I.S.I.*, 859-878, The Hague. International Statistical Institute.
Regan, H.M., Y. Ben-Haim, B. Langford, W.G. Wilson, P. Lundberg, S.J. Andelman and M.A. Burgman, Robust decision-making under severe uncertainty for conservation management, *Ecological Applications* 15 (4), 1471-77.
Roberts R.G., C.N. Hale, T. van der Zwet, C.E. Miller and S.C. Redlin (1998), The potential for spread of *Erwinia amylovora* and fire blight via commercial apple fruit; a critical review and risk assessment, *Crop Protection* 17, 19-28.
Rogers, W. (1986), *Report of the Presidential Commission on the Space Shuttle Challenger Accident.*
Rosen, R.A. and A. Coreggia (2004), The New Basel Capital Accord: Part I: Environmental risks for banks, *Environmental Claims J.,* 16 (1), 93-101.
Senate Hansard (1997), Australian Parliament, Senate Rural and Regional Affairs and Transport Legislation Committee, 11 June 1997, http://www.aph.gov.au/hansard/senate/commttee/s1423799.pdf
Senate (2005), Australian Parliament, Senate Rural and Regional Affairs and Transport Legislation Committee, Administration of Biosecurity Australia: Revised draft import risk analysis for apples from New Zealand, March 2005, http://www.aph.gov.au/senate/committee/rrat_ctte/apples04/report/report.pdf
Sibuya M. (1960), Bivariate extreme statistics, *Annals of the Institute of Statistical Mathematics* 11, 195-210.
Siegel-Jacobs K. and J.F. Yates (1996), Effects of procedural and outcome accountability on judgment quality, *Organizational Behavior and Human Decision Processes*, 65, 1-17.

Silva, R. and P. Ball (2006), A Report by the Benetech Human Rights Data Analysis Group to the Commission on Reception, Truth and Reconciliation of Timor-Leste, http://www.hrdag.org/resources/Benetech-Report-to-CAVR.pdf.

Simonson, I. and B.M. Staw (1992), Deescalation strategies: a comparison of techniques for reducing commitment to losing courses of action, *Journal Of Applied Psychology* 77, 419-26.

Sisson S.A., Y. Fan and M.M. Tanaka (2007), Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of Americ,* 104, 1760-1765.

Sisson S.A., L.R. Pericchi and S.G. Coles (2006), A case for a reassessment of the risks of extreme hydrological hazards in the Caribbean. *Stochastic Environmental Research and Risk Assessment* 20, 296-306.

Smith R.L. (1985), Maximum likelihood estimation in a class of non-regular cases. *Biometrika* 72, 67-90.

Smith R.L. (1989), Extreme value analysis of environmental time series: An example based on ozone data (with discussion). *Statistical Science* 4, 367—393.

Tappin, L. (1994), Analyzing data relating to the *Challenger* disaster, *The Mathematics Teacher* 87 (6), 423-6.

Tetlock, P.E. (1983), Accountability and complexity of thought, *J. of Personality and Social Psychology* 45, 74-83.

Tetlock, P.E. (2005), *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press: Princeton.

Tillers, P. (2005), If wishes were horses: discursive comments on attempts to prevent individuals being unfairly burdened by their reference classes, *Law, Probability and Risk* 5, 33-49.

Tufte, E.R. (1997), *Visual explanations: images and quantities, evidence and narrative,* Graphics Press: Cheshire, Conn.

USNRC (United States Nuclear Regulatory Commission) (2006), Backgrounder on Nuclear Power Plant Fire Protection, http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/fire-protection-bg.html

Vaughan, D. (1996), *The Challenger Launch Decision: Risky technology, culture, and deviance at NASA*, Chicago University Press: Chicago.

Walley, P. (1991), *Statistical Reasoning With Imprecise Probabilities*, London: Chapman and Hall.

Wallsten, T.S., D.V. Budescu and R. Zwick (1993), Comparing the calibration and coherence of numerical and verbal probability judgments, *Management Science* 39, 176-90.

Wells, G.L. and G.A. Olson (2003), Eyewitness testimony, *Annual Review of Psychology* 54, 277-95.

Wieczorek G. F., M.C. Larsen, L.S. Eaton, B.A. Morgan and J.L. Blair (2001), Geological Hazards Team: Debris-flow and flooding hazards associated with the December 1999 storm in coastal Venezuela and strategies for mitigation. US Geological Survey. (Open File Report 01-0144).

9. Quantifying Bank Operational Risk (Supplementary Report)

Gareth Peters and Venta Terauds
School of Mathematics and Statistics, University of New South Wales, Sydney 2052

9.1 Executive Summary

The modelling of operational risk has taken a prominent place in financial quantitative measurement, as a result of the Basel II regulatory requirements on banks and similar financial institutions. This report details the modelling of extreme and rare events in the context of operational risk, with particular focus on the Australian financial sector. Initially the regulatory environment in banking within Australia is discussed in the context of operational risk. There is a focus on quantification requirements for operational risk and why such quantification is important in relation to regulatory standards. Definitions and discussion of operational risk and the associated risk categories introduced by Basel are detailed.

Then the different methodological and modelling approaches that are allowed for in the regulatory guidelines are presented from the most basic to most advanced approaches, including associated regulatory requirements. This is followed by an overview of the different phases that may be required to implement such a methodological and cultural change in a financial institution embarking on modelling operational risk.

Following this is a section which discusses the issues and difficulties associated with modelling operational risk. In particular aspects of operational risk which make modelling difficult at the most fundamental level are detailed. A strong focus of this section is on the different forms of data that can be incorporated in operational risk quantification. This includes an overview of issues associated with data collection and the analysis of data prior to modelling.

Preceding this is a section addressing the industry standard modelling framework, Loss Distributional Approach (LDA). This section includes discussion of popular statistical models utilised to model the annual loss distributions of risk profiles that fall under the banner of operational risk. To complete the discussion of quantitative approaches, a section on statistical models and methodology for particular data sources is presented.

Finally a section on the management aspects of operational risk is presented which relates the regulatory requirements for dealing with assessed and modelled risk profiles.

9.2 Background and Context within Australia's Financial Industry.

In January 2001 the Basel Committee on Banking Supervision proposed a New Basel Accord known as Basel II which was to replace the 1988 Capital Accord. This proposal considers three pillars which by their very nature emphasise the importance of assessing, modelling and understanding operational risk profiles. These 3 pillars are; minimum capital requirements (refining and enhancing risk modelling frameworks), supervisory review of an institutions capital adequacy and internal assessment processes and market discipline which deals with disclosure of information. Since this time the discipline of operational risk and its quantification has grown in prominence in the financial sector.

Operational risk, for a business or organisation, may broadly be defined as the risk involved in such an entity carrying out its normal operations. For a bank, the Basel Committee on Banking Supervision ("the Comittee") defines operational risk to be "the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events." (Basel Committee on Banking Supervision, 2006, p144)

So, operational risk is indeed a broad category. The Comittee gives a further classification into seven types of operational risk (Basel Committee on Banking Supervision, 2006, Annex 9):

Internal Fraud,
External Fraud,

Employment Practices and Workplace Safety,
Clients, Products and Business Practices,
Damage to Physical Assets,
Business Disruption and System Failure,
Execution, Delivery and Process Management,

which serves to further illustrate the disparate nature of events in this class. Reputational and strategic risk do not fall under the operational risk umbrella, and market and credit risks are treated separately, but almost any other event that may result in a loss to a bank, including legal action, may be termed an operational risk. In Appendix 1 we provide some specific examples of types of operational risk.

Basel II regulatory requirements have significantly changed the view that financial institutions have on operational risk. Under the three pillars of the Basel II agreement set out in the framework[1], internationally active banks are required to set aside capital reserves against risk, to implement risk management frameworks and processes for their continual review, and to adhere to certain disclosure requirements. These regulatory requirements, which are overseen and enforced in Australia by APRA, have encouraged many banks to deploy significant resources to the task of quantifying operational risk. Whilst many operational risk events occur frequently and with low impact (indeed, are `expected losses'), others are rare, and their impact may be as extreme as the total collapse of the bank. In any case, most institutions will not have sufficient internal data to accurately model their operational risks, especially with respect to extreme rare losses. The modelling and development of methodology to capture, classify and understand properties of operational losses is a new research area in the banking and finance sector.

Accordingly, the Basel II agreement incorporates a lot of flexibility. The Comitteee itself is made up of representatives of both central banks and banking supervisory authorities from each of the G10 countries, and the framework has been developed and revised in consultation with the authorities and the industry in member and non-member countries. The timing and degree of implementation is determined by the supervisory authorities in each country. Thus the agreement allows for differences in banking practices and regulations that occur across borders and evolve with time. It further allows for differences in size and activity of financial institutions by prescribing three different methods by which operational risk capital may be calculated. In order to implement each of the two more sophisitcated methods, a bank must meet certain qualifying criteria: in essence, it must prove to the regulator that it has sufficient resources and systems in place to properly carry out and audit/review the more sophisticated calculations. At the same time, institutions are expected to use (or to move towards using) the most sophisticated method that they `can'.

It is important to understand where operational risk fits into the overall risk picture within Australian institutions as this will help motivate the effort to spend time and resources on modelling many of these rare events in the presence of significant constraints. In Australian retail banking, the largest profit centre typically revolves around consumer credit and lending. The mortgage and home loan products represent the majority of profit. Other profit centres include markets and trading on the Australian stock exchange or global markets. Modelling of credit portfolios has been developed over many years and is reasonably well established in the banking sector. There are large databases, there are credit rating agencies, standards and rules for assessing and scoring credit ratings. The modelling of annual loss from a credit perspective, including rare event modelling, is well-established, and quantiles of the annual loss distribution are used to produce risk measures such as Value at Risk (VaR) figures which provide capital estimates. So in both methodology and in systems and process development, including accountability and incentives to report and maintain business processes, this area of modelling for extreme losses is highly developed.

At the other end of the spectrum one has operational risk. The infancy of modelling of operational risk relative to other risk disciplines was recognised by APRA early in the introduction of this new risk discipline: "...measuring and managing operational risk is still very much an emerging discipline" [Laker, 2006, p6] . . . "Unfortunately there is neither a history, nor broad agreement on the methodologies, for

[1]Note that the agreement covers market, credit and operational risk, however we shall restrict our consideration here to operational risk.

modelling operational risk." [Egan, 2005, p4]. As a response one of the key drivers put in place by APRA in Australia, in order to push the effort in developing a methodological framework for operational risk and implementation of this framework in an integrated manner throughout a financial institution, is the fact they have tied the accreditation of an advanced approach to credit modelling with an advanced approach to operational risk, typically termed Advanced Measurement Approach [APS 115].

We shall presently summarise the different approaches that are proscribed for quantifying operational risk. For now we note the significance of this move from APRA. The capital charge required to be held for credit lending typically dominates relative to other risk classes, certainly within retail banking. Hence the advanced approaches to modelling from a banking perspective are expected to lower this required reserve, freeing up capital to be used to grow the business. At least in Australia, this puts a very significant monetary incentive in place for financial institutions to adequately model operational risk over time. Looking at the picture from another perspective, banks should be very prudent in modelling rare events and developing understanding of the processes - such as system failure, infra-structure failure and rogue trading - that lead to massive losses, all of which have the potential to debilitate a financial institution or its subsidiariessiduaries.

The guidelines presented by APRA, as with those in Basel II, are not prescriptive in terms of implementation and methodological development. In particular, from a quantitative perspective, they do not advocate particular models for extreme or rare events. The most important quantitative guideline to date from APRA is APS 115. The key requirements specified in this standard are that a bank must have a "framework to manage, measure and monitor operational risk commensurate with the nature, scale and complexity of the institutions operations" and "approval from APRA to use an Advanced Measurement Approach to operational risk for determining the institution's operational risk regulatory capital requirements".

We now outline the three broad approaches that a bank may use to calculate its minimal capital reserve, as specified in the first pillar of the Basel II agreement.

The Basic Indicator Approach

Under this approach, capital is simply a fixed percentage of a bank's gross annual income. The gross income is taken as the average of that of the past three years, excluding any years in which the income was negative or zero. Currently, the committee has set the percentage at 15%.

The capital estimate provided by this method is likely to be an over-estimate, although, given the amount of resources required for the complex task of accurately quantifying operational risk, it may be the best method for some smaller banks. In general, internationally active banks and those with "significant operational risk exposures" are expected to use one of the more sophisticated approaches. In order to do so, certain systems must be in place. Specifically, before using the standardised or a higher-level approach, at a minimum, a bank must be able to show that

1. Its board of directors and senior management are actively involved in the oversight of the operational risk management framework;

2. It has an operational risk management system that is conceptually sound and is implemented with integrity; and

3. It has sufficient resources in the use of the approach in the major business lines as well as the control and audit areas. (Basel Committee on Banking Supervision, 2006, p148)

The Standardised Approach

This approach is similar to the basic indicator approach, in that gross income is used as the basic indicator of risk. In this approach, the gross annual income is considered separately for each of eight business lines - corporate finance, trading & sales, retail banking, commercial banking, payment & settlement, agency services, asset management, and retail brokerage – and a different percentage multiplier is applied to each

business line's income to give a business line capital charge. Again, the multipliers (termed "betas") are set by the committe; the current values are 12% (for retail banking, asset management and retail brokerage), 15% (for commercial banking and agency services), and 18% (for corporate finance, trading & sales and payment & settlement). Some analyses of these betas is provided in (Moscadelli, 2004) which estimates equivalent betas from a numerical study of many financial insitutions and then compares to the current values set by the Basel committee.

The business line capital charges are then summed to give a raw annual total for the bank. The minimum reserve in a given year is the average of this raw total over the previous three years, with any negative capital charge replaced by zero (rather than that year being excluded, as in the above approach).

The Alternative Standardised Approach (ASA)

The ASA is essentially the standardised approach, but with further flexibility, particularly in the treatment of the retail banking and commercial banking business lines. One option is to divide a bank's operations into two categories: "retail and commercial banking" and "other". The capital charge for the former is based on assets rather than income - it is the product of the total outstanding loans and advances of the section with the beta for commercial banking (15%) and a fixed factor m. The capital for the "other" grouped business lines is then just 18% of the gross total income, and the two are added to give total capital charge.

In Australia, APRA has proposed that banks adopt *at least* this option in the ASA. As discussed earlier, most of the activity of Australian banks (or ADIs: "Authorised deposit-taking institutions") occurs in retail and commercial banking, with the majority of assets (at least in smaller banks) in residential lending. Thus this approach is deemed to provide a more realistic estimate than the basic indicator and standardised approaches [Laker, 2006, p2-3]. Of course, the accuracy of the capital estimate is expected to increase with the complexity of the approach, and several of the larger Australian banks are in the process of adopting the Advanced Measurement Approach.

The Advanced Measurement Approach (AMA)

A bank adopting the AMA must develop a comprehensive internal risk quantification system. This approach is the most flexible from a quantitative perspective, as banks may use any methods and models they believe are most suitable for their operating environment and culture. However it is also the most restricted in that banks must gain supervisory approval before beginning to implement the AMA, and their models must satisfy further stringent qualitative and quantitative criteria outlined in the Basel II agreement. To start with, a bank is required to have an independent operational risk management section that is reponsible for the measurement of operational risk as well as the development of strategies for its management and mitigation. It must have an embedded `risk culture', where the day to day operations of the bank integrate risk control, measurement and reporting. All risk management processes must be well-documented and subject to regular internal and external audits... and so on. [Basel Committee on Banking Supervision, 2006, p150-2]. The key quantitative criteria are that a bank's models must sufficiently account for potentially high-impact rare events, and incorporate the use of each of

1. internal data;
2. external data;
3. scenario analysis; and
4. business, environment and control factors.

The detail in implementing these guidelines and meeting these requirements for a bank typically involves a strong interplay between the bank and the supervisory authority. Representatives from APRA regularly visit banks applying for the AMA approach to assess and provide feedback on all aspects of operational risk models and management frameworks being developed. The process also typically involves one or more outside independent parties such as KPMG, Ernst & Young, PricewaterhouseCoopers and Deloittes. These act as intermediaries, providing for APRA assurances and validations of approaches, models and implementations of business and data frameworks developed within banks.

This is an important part of the process in applying to the regulator for approval to use the AMA approach, since APRA is interested in external validation reports [see point 20 in attachment A of APS 115]. The

reason for this will become more clear in subsequent sections discussing modelling approaches and data management. Additionally, in October 2006 all banks applying to use the AMA in operational risk were required to take part in an exercise termed QIS5. This included producing a report on

1. the modelling approach implemented to date (including data management and recording systems covered in IT departments);
2. the road map for the following year leading up to deadlines for accreditation in the first round; and
3. (most importantly) VaR numbers for operational risk in Australia and any subsidiary holdings, including both standardised and AMA figures.

We note that the calculationualtion of Value at Risk as a risk measure is hotly debated in the academic community, especially relating to issues such as coherency in a risk measure [see Artzner et al, 2000] and the difficulty of estimating the quantile level of the annual loss distribution reported ($Q0.999$[2]). For such rare events as terrorist attacks, natural disasters and so on, these figures may not be sensible or stable over time. This was highlighted recently at the Quantitative Methods in Finance conference by a senior member of the German Financial Supervisory Authority. Dr. Gerhard Stahl discussed in his talk the concerns an ADI should have over the level of accuracy that is attainable for reporting at a 0.999 quantile level, and how stable this will be over time.

It is also worth considering what is involved in the practical implementation of an operational risk framework in a financial institution, as it is a massive undertaking. Crudely, the process can be separated into four phases. We shall briefly describe these phases before moving on to the core section of this case study, which revolves around one of these phases "methodological developments". By understanding how this framework needs to be integrated with the business, one gets a sense of the significance of developing models which will as best possible capture the behaviour of these rare events in operational risk. The business is now becoming accountable since managers will need to actively assess and manage their operational risk profile, which is passed to them from the models developed. Clearly this provides a significant incentive to ensure models are transparent and well understood by the risk community. Further this knowledge needs to flow on through the business managers who are being assessed on how well they actively manage such losses and events occuring from operational risks.

Phase 1 – The first step in the process is typically to build a core team for the development and implementation of the entire framework. This may include business representatives, risk specialists and quantitative analysts, policy developers and database experts, business analysts, auditors and validators.

Phase 2 – A key question faced by many financial institutions regards the development of an inhouse framework versus an "off the shelf" or "plug and play" solution, which would be modified for the given business model or hierarchy. The framework includes

1. areas of database design and the set up for the capturing of the Internal Loss Data;
2. choosing the desired modelling methodology and modelling of the actual annual loss distribution under this approach; and
3. reporting and integration of results from modelling into other sections of the institution, including education and assessment of risk profiles. [Cruz 2002]

Key reports and information flows in this space include the reporting of economic capital (an internal measure of capital – typically at a different quantile level to regulatory capital) and profit after capital to the bank's risk committee. In a truly integrated operational risk framework one could even go as far as assessing individual managers' Key Performance Indicators (KPIs) according to the performance of the operational risk capital charge on individual business units. Clearly this impacts the entire institution. Thus another aspect to consider, from a practical perspective, when developing these models is how to obtain substantial "buy-in" from business units who will want to understand how they are exposed to different levels of extreme events relative to other business units.

[2] APS 115 - Point 20

Phase 3 – Development of the model methodology. In this area APRA has given significant flexibility to Australian financial institutions. This is reflected in the many varied approaches implemented throughout Australia. In Australia the key requirements from the regulator are set out in a series of documents which include draft prudential standards, draft prudential practice guides, response to industry progress and discussion and guidelines. The most recent versions of the draft prudential standards released for Australia's financial industry are APS114, APS115 and APG115.

Of these documents the one which is directly relevant to operational risk quantitative methodology is APS115. In this document the first section outlines the process a bank must undertake to obtain approval for an AMA. From the perspective of modelling rare events, points 18 through to 26 provide the guidelines; the allocation of capital charge to business units according to their risk profile is then covered in point 27. Point 18 gives an indication of the level of detail provided: "the [bank's] operational risk measurement system must be sufficiently comprehensive to capture all material sources of operational risk across the bank, *including those events that can lead to rare or severe operational losses.*" In addition to this note, which refers briefly to the nature of the modelling required, another important point to be adressed regards soundness standards over a universal annual modelling period. Statements such as "This soundness standard provides significant flexibility for [a bank] to develop an operational risk measurement system that best suites the nature and complexity of the [bank's] activities" and "Given the subjectivity and uncertainty of operational risk measurement modelling, [a bank] must be conservative in the assumptions used in its operational risk measurement model, *including assessment and incorporation of severe loss events*", illustrate the significant challenge involved in constructing appropiate methodology. Further discussion on the soundness standards and the ten principles that underpin them can be found in [KPMG 2005]

Even before models can be developed, questions such as *how best to understand the nature of rare and extreme events that may lead to large losses* must be asked. The answers that a bank provides to these questions will dictate many of the modelling assumptions that can be made. Questions related to data sufficiency and validity, in addition to likely sources of information, and how best to integrate and fuse information on rare events, become critical to the process. In this context one needs to carefully assess how useful different data sources are for a given institution. Typically this requires thorough understanding of sources of bias present in data. Operational risk is inherently an area where data is still scarce and precious. Thus incorporation of expert opinion in many cases becomes a key driver in the measurement models.

The level of a business hierarchy at which relevant operational risk information can be extracted (and suitably modelled) directly affects the approaches many banks take in modelling, including the granularity of modelling for a given business hierarchy. In this respect, granularity is a term used to refer to the number of levels of the business unit risk type hierarchy used in modelling. For example a model which is not granular could model at the bank level by collecting all the loss data for a given risk type and combine it together then fit a statistical model. In Australia many banks model data at different levels of granularity ranging from assessment and modelling of internal loss data or external loss data at an institution level through to survey and scenario analysis at sub Business Unit and Risk Type (internal fraud, external fraud etc.) levels.

This then influences how easily expert opinion on extreme losses from different business units can be extracted, and how comprehensive this information is for a given business unit's risk profile locally within the business hierarchy. In turn, this affects how efficiently a business unit manager can understand, monitor and improve on their operational risk performance.

It should be noted that recent years have seen the emergence of typically 3 different data sources, combinations of which are used in different models implemented in banks. These data sources are

1. scenario analysis or survey data;
2. internal loss data collected to date (can be very scarce and typically does not contain any truly large losses); and
3. external data which comes from external companies such as FITCH.

However, the use of external data is severly hampered by the fact that many providers do not have complete records, and do not release institutional information. Hence scaling of loss amounts according to institution size - which is critical if data from external sources is to be combined with internal data - is very difficult, if not impossible in many cases. This in a sense compounds the problem, since many of the actual events recorded in these external data bases are the truly large or extreme losses that have been witnessed in the industry.

Once these questions are understood within the context of the bank's business framework, then the models for measurement of operational risk can be developed. The approaches taken will be elaborated on in future sections of the report.

Phase 4 – Calibration, sensitivity analysis and improvements to scenario analysis approaches. Again, this phase requires a lot of quantitative attention on how best to calibrate the model and how sensitive different modelling approaches are to key assumptions, inputs and approximations.

9.3 Model Frameworks for Operational Risk.

9.3.1 Issues Associated with Modelling Operational Risk.

It is relevant to start the consideration of operational risk models with an understanding of what makes developing quantitative models and methodology difficult. There are many reasons. Firstly the sheer size of financial institutions and their subsidiaries makes co-ordination and understanding of approaches to operational risk a practical challenge. This raises issues such as the need for different business units located in different sections of Australia and overseas to understand requirements of assessment, and to act to establish management frameworks. This is important as the line managers of such business units need to actively assess and manage risk according to the behaviour of their reported "modelled" risk profile. In this regard there is typically an information asymmetry, with much of the expertise in understanding the models developed - and therefore the key assumptions made in the process - located in centre functions, physically far away from many of the business units actually affected by operational risks.

The second issue is whether a bank is to implement in their models a "top-down" or a "bottom-up" approach. A top-down approach will do the mathematical modelling of the risk profile at a high level, for example the Bank level. All the loss data for the bank will be assumed homogeneous in terms of truncation and threshold levels and will be modelled as one set of data. This makes explicit assumptions about properties of the collected loss data, however it has the advantage of plenty of data for statistical modelling. For mathematical details see [Panjer 2006]. Once modelled at the top level of the hierarchy the capital results will be allocated to business units according to some weighting factors. A bottom-up approach will model data and expert opinion at much lower levels of the hierarchy. For example individual business units will assess material risk types for their business and any loss data associated with this business unit and risk type will be modelled at this level. Then an aggregation process will be performed to combine all the business unit risk type loss profiles to a bank level.

This again will significantly influence the types of models, and in particular, how data is used in such models. The chosen approach is usually dependent on how well a bank believes they can capture information from expert judgement and then integrate this with other loss data in the quantification process. Largely this process also involves significant business interaction, "buy-in", to develop a team of experts in the business unit who actively assess the local risk profile and take part in risk assessment exercises. This will be discussed in more detail in the next section, where we discuss the modelling of the individual data sources in operational risk.

Operational risk can borrow ideas from insurance mathematics in the area of methodological development. Many models and approaches which are based around the mature field of insurance mathematics have been advocated by researchers in academic institutions [Cruz 2002; Panjer 2006]. However, there are several key differences which will be explored in the context of operational risk. The most significant is the fact that operational risk is still a very new "science" and is inherently an inexact science where model assumptions

and expert opinions are critically important to capture. Understanding the implications for a model of such judgements and assumptions is also a key part of the model development journey.

9.3.2 Modelling Methodology for Operational Risk and the Loss Distributional Approach.

Once the level of framework granularity is decided for the operational risk model (as a function of the relevant data thats obtainable for each level of the hierarchy), the next step in the process is to apply a modelling framework. Of the methods developed to model operational risk, the majority follow the Loss Distributional Approach (LDA). The idea of the LDA is to fit severity and frequency distributions over a predetermined time horizon, typically annual as specified in the APS115 section on soundness standards.

The fitting of frequency and severity distributions as opposed to simply fitting a single parametric annual loss distribution involves making the mathematical choice of working with compound distributions. This would seem to complicate the matter, since it is well known that for most situations, analytical expressions for the distribution of a compound random variable is not attainable in an analytical form. The reason for modelling severity and frequency distributions seperately then constructing a compound process is summarised in detail in [Panjer 2006]. Some of the key points relating to why this is important in most practical settings are;

- The expected number of operational losses will change as the company grows. Typically growth needs to be accounted for in forecasting the number of operational risk losses in future years based on previous years. This can easily be understood when modelling is performed for frequency and severity separately.
- Economic inflationary effects can be directly factored into size of losses through scaling of the severity distribution.
- Insurance and the impacts of altering policy limits and excesses is easily understood by directly altering severity distributions.
- Changing recording thresholds for loss events and the impact this will have on the number of losses required to be recorded is transparent.

The most popular choices for frequency distributions are poisson, binomial and negative binomial. The typical choices of severity distribution include exponential, weibull, lognormal, generalised pareto, and recently in academic literature the g-and-h family of distributions [Dutta et al. 2006, Peters and Sisson 2006]. On the other side of the methodological divide there is a set of models being developed utilising concepts and ideas from Extreme Value Theory EVT [Embrechts et al 2006]. This divide mainly concerns approaches taken to fit such distributions and is discussed in detail in [Embrechts et al 2006].

A key note to make is that the most important processes to model accurately are those which have relatively infrequent losses. However, when these losses do occur they are distributed as a very heavy-tailed severity distribution. These processes are by their very nature the most difficult to model, due to scarcity of data. From a practical perspective, this is where the importance of eliciting expert opinion and performing surveys or scenario analysis becomes critical.

The reason why these simple parametric models are widely used is that from a practical perspective they are relatively simple to fit, and to apply goodness of fit tests to (for purposes of model selection). Additionally, given the scarcity of most data sources, the fitting of parametric distributions with more than two parameters can quickly become problematic and unreliable. This is a practical issue, however there is also the theoretical issue of whether this class of distributions adequately captures the true behaviour of the extreme events lying deep in the tails of these severity distributions. Industry consensus tends to suggest many of the extreme events, at least in the Australian financial sector, can be adequately modelled by lognormal and generalised pareto distributions. Returning again to EVT, in this space one can fit heavy tailed distributions for the severity distribution. Typically, fitting these models can be performed using either Points Over Threshold (POT) techniques of block maxima [Embrechts et al 2005]. There has also been some literature on fitting EVT models from a Bayesian persepective [Sisson et al 2006]. This approach will be discussed in another section of the report.

There are many approaches which can be used to fit these parametric distributions and the approach adopted by a bank will depend on the data source being modelled and how much confidence one has in the data source. This is highly subjective. Techniques commonly adopted to fit frequency and severity models include extreme value theory [Cruz, 2002], Bayesian inference [Schevchenko et al. 2006; Cruz, 2002], dynamic Bayesian networks [Ramamurthy et al. 2005], maximum likelihood [Dutta et al. 2006] and EM algorithms [Bee, 2006]. (In the next section we present a framework for modelling and a set of statistical tools which can be used to fit these distributions to different data sources and then select between the different proposed models.) After the best-fitting models are selected, these are combined to produce a compound process for the annual loss distribution:

$$Y = \sum_{i=1}^{N} X_i \,, \qquad (1)$$

where the random variable $X_i \sim f(x)$ follows the fitted severity distribution. The random variable $N \sim g(n)$, the fitted frequency distribution, is commonly modelled by poisson, binomial and negative binomial distributions [Dutta *et al.* 2006]. From this compound process, VaR and capital estimates may be derived.

Once compound processes have been fitted for each business unit and risk type, the next step in the process is to aggregate these annual loss random variables for each individual {business unit-risk type} combination, and thus to obtain the institution-wide annual loss distribution. This report will not discuss the issues associated with correlation and dependence modelling. For more information on typical approaches to introducing correlation in an aggregation process, including copula methods, correlation of frequency, severity or annual losses, see [Cruz 2002].

At a given level of the hierarchy structure, (which we may call a {business unit-risk type} tree), if there are M {business unit-risk type} combinations present[3], this process of determining the distribution of the annual loss involves an M-fold convolution:

$$Y_{levelM} = \sum_{i=1}^{M} Y_i \,,$$

Then the distribution of such an annual loss random variable will be given by,

$$f_{levelM}(y) = \int \int \ldots \int f_{BuRT(i)}(\tau_1 - \tau_2) f_{BuRT(i)}(\tau_1) d\tau_1 d\tau_2 \ldots d\tau_M \,,$$

Since each of these distributions $f_{BuRT(i)}$ for each {business unit-risk type}, at the lowest level of the business unit risk type tree, takes the form of a compound process developed from the LDA model framework, solving these convolution integrals for an analytic expression is not possible [Panjer 2006]. Hence, typically in practice, different forms of simulation are used to estimate these compound distributions. Then the convolved institutional level annual loss distribution, and finally the regulatory capital estimate are obtained (typically by using a VaR at the specified Q0.999).

An aside on approaches that have been used to simulate such compound processes to estimate the annual loss distribution can now be presented. The reason the compound distribution of Y has no general closed form is that it involves an infinite sum over all possible values of N, where the n^{th} term in the sum is weighted by the probability $Pr(N=n)$ and involves an n-fold convolution of the chosen severity distribution, conditional on $N=n$. Actuarial research has considered the distribution function of Y for insurance purposes through Panjer recursions [Panjer, 2006]. Other approaches utilize inversion techniques such as inverse Fourier transforms to approximate annual loss distributions, although they typically require assumptions such as independence between frequency and severity random variables [Embrechts *et al.* 2003].

[3]This number M will depend on the level of granularity of the model being used by the bank.

9.3.3 Modelling the Different Data Sources, Elicitation of Expert Judgement and Models to Fit this Information.

Before mentioning some techniques that can be used to fit parametric severity and frequency distributions to actual loss data or expert elicited judgements in the form of survey or scenario analysis, it is worth noting some recent theoretical results regarding the aggregation of compound processes.

Recent work found in "Multivariate models for operational risk" [Bocker *et al* 2005] provides some analytical results for the asymptotic quantiles of an annual loss distribution constructed by aggregating several compound processes. This is useful as it provides mathematical insight on bounds for the VaR at high quantiles (such as those required for operational risk capital reporting) after aggregation of several different types of compound process. In this paper, the key findings of the authors can be interpreted as stating that in the independent compound process case, the combined VaR measure at the next level of the hierarchy, after aggregation, will be asymptotically in the quantile level, convergent to the single compound process expression for the quantile with dominating VaR.

That is, if the VaR values of the individual ranked compound processes are dominated by one particular processes VaR then this will be the asymptotic VaR of the aggregated annual loss distribution at the next level, in the independent case. This study proves these results for classes of sub-exponential severity distributions which comprise Poisson processes. A subexponential distribution F satisfies, for $(X_i)_{i\in N}$ i.i.d. random variables,

$$\lim_{x\to\infty} \frac{P(X_1 + \ldots + X_n > x)}{P(\max(X_1,\ldots,X_n) > x)} = 1$$

for some value of n. So without concern over the mathematical technicalities presented above, broadly '*this translates into a statement that for sub-exponential distributions the sum of the random variables will be dominated by one single large loss and not by the summation of several small losses'*. This is particularly relevant to Operational Risk analysis. In [Bocker et al 2005] the authors provide an analytic result for an example of a bi-variate VaR, that is, one calculated from the aggregate of two compound processes. The frequency distribution for both processes is chosen to be Poisson; for the severity distributions the selected distributions are Weibul for one process and Lognormal for the other. When convolution of the two compound annual loss random variables is achieved, the VaR for the annual loss distribution at the aggregated level is dominated by the compound process produced by the Lognormal and Poisson distributions.

Additionally, in earlier work by the same authors they demonstrate in an LDA setting, that the tail quantiles of a compound process with sub-exponential severity distributions will be simply a multiplicative function of the mean of the frequency distribution and the quantile of the severity distribution. Hence showing asymptotically that the quantiles of the compound process will be independent therefore of the over-dispersion effects that can be added when including for example Negative Binomial processes. This is important from a modelling perspective as it indicates that when concern is in estimation of VaR, one can stick to fitting Poisson processes which come with well understood properties. These include independent modelling increments, exponentially distributed inter-arrival times for loss events both of which makes fitting such models to actual data significantly simpler.

These results clearly have implications which are yet to be realised and studied for the dependent processes case which is typically considered relevant in practical settings. An example of this is where frequency random variables for different business unit risk type processes are dependent on each other.

9.3.4 Survey Data and Scenario Analysis

From a practical perspective the most important data source in the Australian financial sector comes from survey or scenario analysis. This is largely a result of scarce data on rare events for process's such as terrorist attack, natural disasters, rogue trading and infra-structure failures. There is two broad approaches to dealing with expert opinions, Scenario Analysis and Survey Data. Scenario Analysis typically involves

setting up workshops with each business unit for which operational risk is being assessed and going through a sequence of exercises to assess potential loss amounts for each non-negligible risk type.

The term scenario analysis is used since a workshop facilitator will extract loss information from the business expert participants through a sequence of questions relating to internal events, external events both actual and hypothetical in the form of scenarios. An example of this would be to ask questions if assessing for example 'rogue trading'; What is the expected exposure and what is the worst case possible exposure? What systems are in place to set limits? What are the known and potentially unknown flaws in such systems? How are these being managed? How does the scale of operations in this bank compare to other known incidents in the financial sector that the bank is operating? What management frameworks are in place?

If an LDA approach is utilised then this information is considered by all participants in the workshops and questions directed at extracting information around the severity and frequency of such events are presented. What is the typical exposure? What is the expected exposure? What is the worst possible loss? What is the 1 in 10 or 1 in 20 year loss? How often does the loss occur per year ? etc...

These answers are then used to fit frequency and severity distributions. One way to do this is to use extracted measres of location and quantiles in dollar values to fit the severity distribution by solving the simultaneous equations relating the parameters to the quantiles or summary statistics elicited in the scenario analysis. The Frequency distribution can be fitted to rates of occurrence. Typically the fitted severity and frequency distributions and the simulated annual loss distributions would be played back to the business experts for each of the possible severity and frequency models considered. Then a feedback and refinement process is undertaken until there is comfort from the business and facilitators that the risk profile adequately captures the behaviour of the exposure for the assessed risk.

Other approaches include eliciting a sequence of quantiles or relative probabilities for different loss intervals. In general the following broad distributional summaries can be used as frameworks for developing scenarios and survey questions; Probabilities – extract individual probabilities of loss amounts based on actual industry losses, Quantiles – qth quantiles such as median, 1 in 10 loss or 0.9 quantile, Intervals – probability of losses above some threshold or in some dollar range, Location Measures - typical or representative measures of dollar losses (median, mode, mean), Scale and Dispersion Measures – how far from the (mean, median, mode) the loss might be, Measures of Shape – describing the density as unimodal, bimodal or multimodal, skewed left or right and kurtosis in the form of questions relating to tail behaviours.

The merits of each approach and an excellent discussion of such elicitation processes and the sources of inherent bias are presented in great detail in [O'Hagan 2006] . This text considers a very wide cross section of literature from Psychological expert elicitation and perception, practical elicitation and facilitation, statistical bias, survey development and modelling.

As pointed out in [O'Hagan 2006] it is important to understand that "the subjective perceptions and sensations are, in principle, measurable – and with some precision – but such measurements can only be interpreted relatively not absolutely". In this regard one needs to consider the possible impact of forcing the busines experts to conform to a certain summary of the severity and frequency distributions. Additionally, O'Hagan points out that when capturing expert opinion about some uncertain quantity in the form of a distribution it is important to recognise the two different forms of uncertainty, aleatory and epistemic. 'Aleatory uncertainty is induced by randomness such as when modelling uncertainty in one or more instances of a random process'. 'Epistemic uncertainty is due to imperfect knowledge about something that is not itself random and is in principle knowable.' Hence when developing models based on this survey and scenario analysis it is important to somehow consider separate variables or behaviour as a result of these two different uncertainties.

The second framework involves a Bayesian paradigm [Bayes 1763]. This approach from the perspective of operational risk is captured in [Peters and Sisson 2006, Shevchenko *et al* 2006]. To understand the difference between the Bayesian approach and the scenario analysis approach it is important to realise that

typically scenario analysis makes the assumption that the parameters of the severity and frequency distributions are deterministic. It then aims to extract what is equivalent to point estimates of the parameters required for the LDA approach. The Bayesian approach treats the problem from a different paradigm. The parameters are treated from a mathematical perspective as random variables and the survey and elicitation process now involves extracting information on the prior distribution for these parameters. This prior coupled with the likelihood model for the severity or frequency distribution is combined under Bayes law to produce a posterior distribution on the parameters. Hence, from this perspective the elicitation of prior information should follow a different route to the typical scenario analysis. More information of prior elicitation procedures is found in [O'Hagan *et al* 2007

9.3.5 Internal Loss Data and External Data

Typically the process involved in internal and external loss data is to firstly study the properties of the data in each risk category. This involves histograms, box plots, time series plots, all of which are used to identify and question trends present in the data which may be artificial. This could include misclassification of loss events, censoring and truncation etc.

Once the data is investigated, typically a maximum likelihood approach is used to fit the severity distributions. Other approaches could involve generalised moment matching or quantile matching. When mixtures of distributions are used then the popular approaches include Expectation Maximisation algorithm [Bee 2006]. This is particularly relevant when truncation is known to be present. For a review of each approach and the properties see [Panjer 2006].

If a Bayesian approach is used, typically this loss data would enter into the modelling through the evaluation of the likelihood when simulating from the posterior distribution of the LDA severity and frequency parameters. The simulation procedure in these cases typically involves development of sophisticated procedures such as Markov chain Monte Carlo (MCMC), importance sampling (IS) and sequential Monte Carlo (SMC) algorithms [Doucet *et al.* 2006; Peters, 2005]

Once the models for frequency and severity have been fitted, it is important to introduce some criteria to select the "best model". Typically this involves Kolmogorov-Smirnov or Anderson Darling tests for goodness of fit. Alternatively if a Bayesian approach is adopted one would consider Bayesian Information Criterion BIC or Deviance Information Criterion DIC as statistics to choose between different fitted frequency and severity models which best represent the data in the most parsimonious manner.

In summary, there are a number of pertinent issues in fitting models to operational risk data: the combination of data sources from expert opinions and observed loss data; the elicitation of information from subject matter experts, which incorporates survey design considerations; sample biases in loss data collection, such as survival bias, censoring, incomplete data sets, truncation and, since rare events are especially important, small data sets.

9.4 Managing Operational Risk

With so much effort going into the complex task of quantifying a bank's operational risk, it is important to emphasise that this is just one component of the overall task of managing operational risk. To be specific, the `management' of operational risk means the "identification, assessment, monitoring and control/mitigation of risk" [Basel Committee on Banking Supervision, 2003, p3]. As set out in the previous section, in order to use a quantification approach more sophisticated than the basic approach, a bank must fulfil certain requirements, many of which pertain to its risk management systems. The second pillar of the Basel II agreement [`Supervisory Review Process'] sets out a framework under which supervisors (of individual banks, and of the industry as a whole) must implement this process.

Many of the principles underpinning operational risk management have already been touched on. It is required that banks have a dedicated operational risk management unit, and much focus is on embedding a thorough awareness of operational risk in all levels of the bank's operations. Many of the aspects of risk management are a straightforward precursor to risk quantification: a risk can not be quantified until it is

identified; a risk cannot hope to be accurately quantified unless it is appropriately monitored and all incidences are reported. This is true within an individual bank and externally as well: banks are required to make "sufficient public disclosure" to allow other banks to compare and assess their operational risk [Basel Committee on Banking Supervision, 2003, p5], and supervisors are directed to compare the operational risk calculations of similar banks in their domain [Basel Committee on Banking Supervision, 2006, p217]. It is up to a bank to justify to the supervisory authority that its management systems, as well as its quantification processes, are sufficient, and industry-wide disclosure requirements can help a bank to ensure that it is of the required standard.

Other aspects of risk management have a less straightforward relationship with risk quantification. In many cases, it is desirable to reduce exposure to an identified risk. (Other risks may be taken on intentionally, as part of a wider strategy to reap certain rewards.) A mitigation strategy such as insurance against a particular risk will reduce the risk itself, but in itself introduce further risk, which must be measured, quantified, reported and so on. Thus there is a constant interaction between risk measurement and management. A bank will be constantly refining its risk models due to these internal interactions, as well as due to judgements and directives from the supervisory authorities.

9.5 References (for section 9)

APRA Prudential Standard APS 114 (January 2007), *Capital Adequacy: Standardised Approach to Operational Risk*. Draft Report.

APRA Prudential Standard APS 115 (October 2006), *Capital Adequacy: Advanced Measurement Approaches to Operational Risk*. Draft Report.

APRA Prudential Guide APG 115 (October 2006), *Advanced Measurement Approaches to Operational Risk*.

Artzner P., Delbaen F., Eber J. and Heath D. (2000), *Thinking Coherently*, Extremes and Integrated Risk Management, pp77-82, Risk Books, London.

Basel Committee on Banking Supervision, Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version, Bank for International Settlements, June 2006. Available from http://www.bis.org/publ/bcbs128.htm

Basel Committee on Banking Supervision, Sound Practices for the Management and Supervision of Operational Risk, Bank for International Settlements, February 2003.

Bayes, T. (1763). *An essay towards solving a problem with the doctrine of Chances*. Philos. Trans. R. Soc. London, **53**, 370—418.

Bee M. (2006). *Estimating and simulating loss distributions with incomplete data*, Oprisk and Compliance, **7** (7), 38-41.

Bocker, K. and Kluppelberg, C. (2005), *Multivariate Models for Operational Risk*, Hypo Vereinsbank.

Bocker, K. and Kluppelberg, C. (2005), *Operational Var: A Closed Form Approximation*, Hypo Vereinsbank.

Bortot P., S. G. Coles and S. A. Sisson (2006). *Inference for stereological extremes*. J. Amer. Stat. Assoc. In press.

Cruz M. (2002). *Modelling, Measuring and Hedging Operational Risk.* John Wiley & Sons, Chapter 4.

Dutta K. and J. Perry (2006). *A tale of tails: An empirical analysis of loss distribution models for estimating operational risk capital.* Federal Reserve Bank of Boston, Working Papers No. 06-13.

Egan B. (APRA May 2005), *Basel II Changes and Operational Risk*, Speech given at the Enterprise-Wide Risk Management Conference.
Available from http://www.apra.gov.au/speeches/05_04.cfm

Embrechts P., Degen M. and Lambrigger D. (2006). *The quantitative modelling of operational risk: between g-and-h and EVT*. Technical Report ETH Zurich.

Embrechts P., McNeil A. and Rudiger F. (2005). *Quantitative Risk Management, Techniques and Tools*, Princeton Series in Finance.

Embrechts P., H. Furrer and R. Kaufmann (2003). *Quantifying regulatory capital for operational risk*. Derivatives Use, Trading & Regulation, **9** (3), 217—223.

Garthwaite P. and A. O'Hagan (2000). *Quantifying expert opinion in the UK water industry: An experimental study.* The Statistician, **49** (4), 455—477.

KPMG (2005), *Financial Services Basel II: A Closer Look – Managing Operational Risk*, white paper from Advisory.

Laker J. (APRA April 2006), *Basel II - Observations from Down Under*, Speech given at the Second Annual Conference on the Future of Financial Regulation, London School of Economics.
Available from http://www.apra.gov.au/speeches/BASEL-II-OBSERVATIONS-FROM-DOWN-UNDER.cfm
Moscadelli M. (2004), *The Modelling of Operational Risk: Experience with the Analysis of the Data Collected by the Basel Committee*, Bank of Italy.
Available from http://ssrn.com/abstract=557214
O'Hagan A. (2006). *Uncertain Judgements: Eliciting Expert's Probabilities,*
Wiley, Statistics in Practice.
O'Hagan A. (1998). *Eliciting expert beliefs in substantial practical applications.* The Statistician, **47** (1), 21—35.
Panjer H. (2006). *Operational Risk: Modeling Analytics*, Wiley.
Peters G. (2005). *Topics in Sequential Monte Carlo Samplers*. University of
Cambridge, M.Sc. Thesis, Department of Engineering.
Peters G and Sisson S. (2006). *Bayesian Inference, Monte Carlo Sampling and operational risk.* Journal of Operational Risk, vol. 1, no. 3.
Ramamurthy S., H. Arora and A. Ghosh (2005). *Operational risk and probabilistic networks – An application to corporate actions processing*. Infosys White Paper.
Shevchenko, P. and M. Wuthrich (2006). *The structural modelling of operational risk via Bayesian inference: Combining loss data with expert opinions*. CSIRO Technical Report Series, CMIS Call Number 2371.

9.6 Appendix 1

In the table below are listed a number of kinds of operational risk along with some examples where those risks have been realised, and some applicable methodologies. (table from Franklin, 2005)

Type of risk	**Example**	**Methodology**
Internal fraud and human error	Barings rogue trader	Model pooled anonymised data, fraud detection
External fraud	Credit card fraud	Fraud detection analytics
Acute physical hazards	Tsunami,hail	Reinsurers' data + extreme value theory
Long-term physical hazards	Climate change	Climate modelling + work on effects on banking system
Biorisks	SARS, animal plague	Biomedical research + quarantine expertise
Terrorism	Bombing, Internet attack	Intelligence analysis
Financial markets risk	1997 Asian crisis, depression	Macroeconomic modelling, stock market analysis + extreme value theory
Real estate market risk	Home loan book loses value	Real estate market modelling
Collapse of individual major partner	Enron	Data mining on company data
Regulatory risk	"Basel III", nationalisation, government forces banks to pay universities for graduates	Political analysis
Legal risk	Compensation payouts for misinformed customers	Compensation law and likely changes
Managerial and strategic risk	Payout unwanted CEO, dangerous management decision	
Robbery	Electronic access by thieves	Model pooled data, IT security expertise
Reputational risk	Run on bank, spam deceives customers	Goodwill pricing theory + marketing expertise
New technology risk	Technology allows small players to take bank market share	"Futurology"
Reserve risk Interactions of all the above	Reserved funds change value Depression devalues real estate and reserves	Causal modelling of system interactions

(A few of these, such as reputational risk, are not recognized under Basel II's classification as "operational risk", but are important for a bank to evaluate nonetheless.)