


**Project no. GOCE-CT-2003-505539**

**Project acronym: ENSEMBLES**

# Project title: ENSEMBLE-based Predictions of Climate Changes and their Impacts

Instrument: Integrated Project
Thematic Priority: Global Change and Ecosystems

# Deliverable D5.4: Scientific report on Verification methods for forecasts of extreme events

Due date of deliverable: February 2006
Actual submission date: 17 March 2006

Start date of project: 1 September 2004 Duration: 60 Months
Organisation name of lead contractor for this deliverable: UREADMM
Revision [draft, 1, 2, ..]

| **Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)** | | |
|---|---|---|
| **Dissemination Level** | | |
| **PU** | Public | **X** |
| **PP** | Restricted to other programme participants (including the Commission Services) | |
| **RE** | Restricted to a group specified by the consortium (including the Commission Services) | |
| **CO** | Confidential, only for members of the Consortium (including the Commission Services) | |

# Verification Methods for Forecasts of Extreme Events

David B. Stephenson

Department of Meteorology, University of Reading
d.b.stephenson@reading.ac.uk
28 February 2006

***Abstract***

Verification of extreme event forecasts has received surprisingly little attention with the exception of the work on threat scores for deterministic forecasts of binary events (e.g. threat scores). There is an urgent need to develop specific approaches better suited for the verification of such rare events, in order to assess the performance of current operational ensemble forecasting system.

This report attempts to establish a simple framework for verification of certain types of extreme event and then discusses some of the possible approaches in the hope that this will lead to more developments in this important area of verification.

This report is deliverable D5.4 of the WP5.3 verification work package in the EU project ENSEMBLES (http://ensembles-eu.metoffice.com). It is intended as a discussion paper rather than a final prescription and so comments would be most appreciated.

## *1. Aim*

Accurate prediction of rare high-impact events represents a major and pressing challenge for weather and climate forecasting. Timely forecasts of extreme meteorological events, such as tropical and extra-tropical storms and heat waves/droughts, can be used to inform decisions that help mitigate large economic losses and reduce human morbidity and mortality.

However, assessment of the skill at forecasting such events is problematic primarily because of the rarity of such events. For rarer events, many of the standard scores tend to trivial limits such as zero (e.g. the equitable threat score, the Brier score, and many others). In addition, the sampling uncertainty on estimated scores is large for rare events due to having few events in the verification period. In addition, some extreme events can be difficult to observe due to their small spatial scales (e.g. tornadoes) and observations can be error-prone (or even impossible!) due to the extreme meteorological values (e.g. wind speeds during severe hurricanes).

Despite or perhaps even because of these challenges, verification of extreme events has received surprisingly little attention with the exception of the work on threat scores for deterministic forecasts of binary events (see threat scores in Chapter 3 of Jolliffe and Stephenson, 2003). There is potential for developing specific approaches better suited for the verification of rare events. Such approaches urgently need to be developed to be able to assess the performance of current operational ensemble forecasting systems.

This report attempts to establish a simple framework for verification of certain types of extreme event and then discusses some of the possible approaches in the hope that this will lead to more developments in this important area of verification. No specific time scales are mentioned since verification of extreme events is applicable to weather and climate forecasts at all possible lead times.

The report is deliverable D5.4 of the WP5.3 verification work package in the EU project ENSEMBLES (http://ensembles-eu.metoffice.com). It is intended as a discussion paper rather than a final prescription and so any comments would be most appreciated (email: d.b.stephenson@reading.ac.uk).

## *2. Definition of extreme events*

Mankind is vulnerable to *severe* (also now referred to as *high-impact*; by e.g. WMO THORPEX programme) weather/climate events that cause damage to property and infrastructure, injury, and even loss of life. Although generally rare at any particular location, such events cause a disproportionate amount of loss. Prediction of such high-impact events can improve decision-making and disaster planning that then helps to mitigate some of the losses.

Severe events have high impacts generally because they are associated with extreme values of meteorological variables such as precipitation (e.g. floods), wind speeds (e.g. cyclones), temperatures (e.g. heat waves), etc. Such *extreme events* are multi-faceted and have various attributes such as:

- rate (frequency) of occurrence;
- magnitude;
- temporal duration;
- spatial scale;
- multivariate dependencies.

For example, a category 5 hurricane is quite rare, has large magnitude surface wind speeds, has a generally large spatial scale (with the exception of certain small events such as Hurricane Camille in 1969), and develops over synoptic time scales from hours to several days. In addition, the severity of such events can also depend on the combination of extreme behaviour in more than one variable, for example, much hurricane damage is due to extreme precipitation as well as extreme wind speeds. For example, a severe ice storm can involve conditions on all three variables: temperature, wind and precipitation amount.

IPCC (2001) defines *complex extreme* events as “Severe weather associated with particular climatic phenomena, often requiring a critical combination of variables’’. Prediction of such complicated and often rare spatio-temporal systems represents a major and pressing challenge for weather and climate forecasting.

In order to simplify the problem, it is useful at first to neglect certain attributes such as temporal duration, spatial scale, and multivariate dependencies. IPCC (2001) defines a *simple extreme* event to be “Individual local weather variables exceeding critical levels on a continuous scale”. This is a very simplified but often employed definition of an extreme event. Meteorological variables are invariably recorded and stored on computers at regular discrete time intervals rather than continuously and so a typical time series might look something like this:

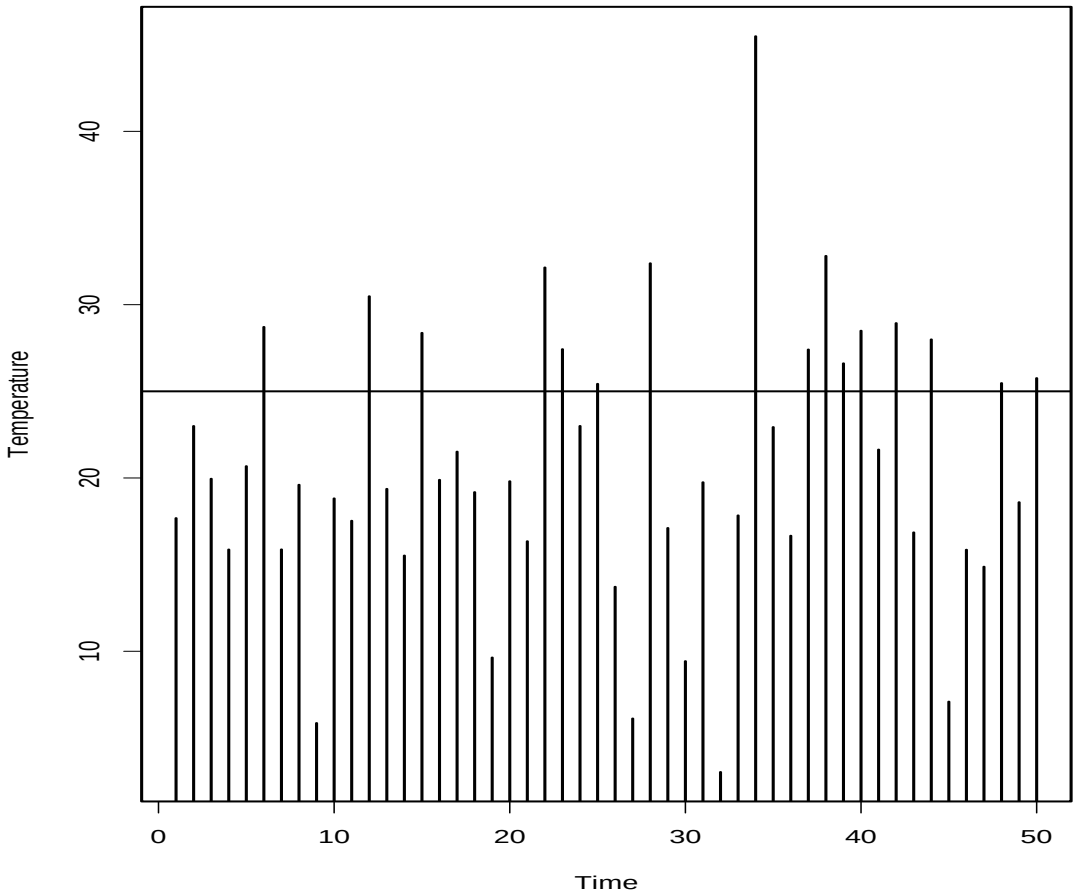


***Figure 1:*** *Artificial example of simple extremes in 50 temperature values generated randomly from a Normal distribution with mean of 20$^{o}$C and standard deviation 7$^{o}$C. The threshold is at 25$^{o}$C. Example generated using the R language commands: plot(rnorm(50,20,7),type="h",xlab="Time",ylab="Temperature"); abline(h=25)*

Several heat wave indices have been developed based on temperature exceedances above 25$^{o}$C (e.g. the NCDC/STARDEX indices). Temperatures above this threshold can cause human heat stress and can lead to extreme health situations. In the artificial example shown in Figure 1, the exceedances above the threshold temperature of 25$^{o}$C (the extreme events) occur at random times (*points*) and have randomly varying exceedances above the threshold value (*marks*). Such series are not amenable to the usual methods of regular time series analysis developed for a sequence of randomly varying values sampled at regular time intervals. However, such records can be considered to be a realisation of a stochastic process known as a *marked point process* – a process with random magnitude that occurs at random points in time (see Diggle, 1983; Cox and Isham, 2000). Strictly speaking, this is a special type of marked point process in which the points occur on a discrete set of regularly spaced times (e.g. daily values) rather than at any possible time. The occurrence of an exceedance in such cases can be modelled using a discrete-time Markov chain (see Section 5.2 of Lindsey 2004).

Point process methods have been widely used in various areas of science, for example, in providing a framework for earthquake risk assessment and prediction in seismology (Daley and Vere-Jones, 2002). Point process methods can be used to explore and summarise such records and are invaluable for making inference about the underlying process that gave rise to the record. Broadly speaking, this is performed by considering statistical properties of the points such as the number of events expected to occur per unit time interval (the *rate/intensity of the process*), statistical properties of the marks (e.g. how the average magnitude of the marks has changed over time), and joint properties such as how the marks depend on the position and spacing of the points, the magnitude of preceding events, etc.

There are two main types of approach for estimating the rate of a point process: the *counting specification* based on counting the number of points in fixed time intervals, and the *interval specification* based on estimating the mean time interval between successive points (see p.11 of Cox and Isham, 2000). The simplest counting specification involves dividing the time axis into a set of non-overlapping equally spaced bins and then counting the number of points that fall into each bin. This approach gives rather noisy results due to the sharp bin edges. More efficient and smoother rate estimates can be obtained using smooth local weighting based on a smooth kernel function rather than a sharp-edged bin. Such an approach has recently been used to investigate extreme flooding events in Eastern Norway as observed in paleoclimatic lake sediments (Boe et al., 2006).

For the sake of simplicity, in the following, we will only consider simple extreme events defined at regularly spaced time intervals. This is a necessary first step before addressing the more complex temporal duration, spatial scale, and multivariate attributes of extreme events that are often ignored when issuing forecasts of severe events. Despite the complex interaction of meteorological variables, meteorological hazards are often perceived as simple extremes in financial losses by the insurance industry and governments. In other words, despite their meteorological complexity, severe events can be defined as simple extreme events in variables such as total insured loss.

## *3. Types of extreme event forecast*

Several different types of forecast can be issued for simple extreme events defined by exceedance of a variable above a pre-defined threshold:

1. **Occurrence of Exceedance forecasts**
   Either deterministic or probabilistic forecasts of whether an exceedance will occur. Deterministic forecasts of whether or not an extreme event will occur within a specified time period. For example, temperature exceeding $25^{o}$C tomorrow for a given location. Probabilistic forecasts of the binary event X defined by the variable exceeding a pre-defined threshold (Y>u). At high thresholds, the series of X will contain a lot of zeroes (because the event is rare) and the forecasts p will generally be small.

2. **Number of Exceedance forecasts**
   Either deterministic forecasts of the number of extreme events or probabilistic forecasts of the distribution of the most likely number of extreme events that will occur within a specified time period. For example, the most likely number or a probability distribution of the number of hurricanes making landfall in the U.S. within a season.

3. **Time-To-Next-Exceedance forecasts**
   Rather than forecast the number of events in a fixed time period, forecast the time interval (or distribution of time intervals) before the next extreme event is observed. This type of forecast has so far been rarely used in weather and climate forecasting. The next category 5 hurricane to make US landfall has a probability of 0.9 of occurring within the next 3 days.

4. **Distribution of the Exceedance magnitude forecasts**
   Forecasts of the predictive distribution of the excesses (the marks). These can be issued as either quantiles of the excess for a given exceedance probability (*return levels*), the reciprocal of the exceedance probability P(Y>u) for a given excess (*return periods*), or alternatively a set of estimated distribution parameters for the predicted distribution of excesses (e.g. scale and shape parameter estimates for a Generalised Pareto Distribution model of the excesses). For example, a forecast for tomorrow of 10-year (specified return period) wind speeds exceeding 164m/s (forecast return level).

There are several important reasons why it is advisable to issue probability forecasts for rare events rather than deterministic forecasts. Whereas deterministic forecasts issue a specific value or category that is considered to be most likely to occur in the future and provide no estimate of forecast uncertainty, probabilistic forecasts make probability (risk) statements about the chance of different events occurring. Several reasons for preferring probabilistic forecasts are that:

- Future extreme events cannot be predicted with certainty and it is legally unwise to claim that there is no uncertainty in the forecasts;
- Probability forecasts are essential for quantitative assessment of risk;
- Probability forecasts allow different decision-makers (forecast users) to make their own optimal decisions, whereas deterministic forecasts are essentially a decision already made by the forecaster;
- Probability forecasts of rare events are harder to hedge for a specific user than are deterministic forecasts (see Murphy 1991b).

However, there are several difficulties when issuing probability forecasts such as:

- more information needs to be communicated so the forecasts can be difficult to communicate concisely (e.g. in short television broadcasts);
- the understanding and perception of probability and risk varies enormously from person to person;

- not all users want to make optimal decisions – they often prefer the forecaster to issue a definitive statement about what will happen (despite the fact that this is obviously impossible!).

In order to surmount these difficulties it is necessary for forecasters and users to work together at improving communication and understanding of what they are attempting to achieve.

## *4. Verification of extreme event forecasts*

Some possible scores for the different types of simple extreme forecast are summarized in Table 1. Much work needs to be done to improve the gaps in this table.

| Type of Forecast | Deterministic | Probabilistic |
|---|---|---|
| Occurrence of exceedance | Threat scores, EDS, ROC | Brier score, Mean Abs score, logarithmic score |
| Number of exceedance | MSE of square root of counts | ??? |
| Time-to-next-Exceedance | MSE of square root or log of time intervals | ??? |
| Distribution of exceedance | --- | Coverage (see text)<br>Interval scores |

Table 1: Summary of scores that can be used for the different types of extreme event forecast.

Deterministic forecasts of occurrence of exceedance can be assessed using the many different approaches developed for binary events (see Chapter 3 of Jolliffe and Stephenson, 2003). In particular, the threat score, and its variant the equitable threat score, have been widely used (and reinvented!) for rare events because of their ability to be defined even if one doesn't know the number of correct no-event forecasts (see Doswell et al. 1990; Schaefer, 1990; Murphy, 1991a; Marzban, 1998).

Probability forecasts of exceedance can be assessed by probability scores such as the Brier score, the mean absolute score, or even the more penalizing logarithmic score (if probabilities of zero are never issued).

Number of events is generally a positively skewed non-normal distribution of discrete values and so appropriate measures need to be used rather than simple measures such as mean squared error that is more suitable for normally distributed variables. Skewness can easily lead to large values that dominate the score and so either more resistant measures should be used (e.g. mean absolute error) or the counts should be normalized (e.g. by a

square root transformation) before verification. Time-to-next-event forecasts are rarely issued by weather and climate forecasters but such variables would also be strongly positively skewed and non-normal and so would also necessitate care in verification.

Return level forecasts can be considered to be the lower limit of an interval forecast in which the upper limit is the maximum value attainable by the observed variable. One thing to verify for such forecasts is the coverage: the fraction of times when the observed event falls above the forecast values should be close to the specified exceedance probability. This is a measure of forecast reliability not resolution –good forecasts should have good reliability yet should also give high return levels on occasions when extreme events occurred.

It should be noted that no single score is optimal for all people; users are often interested in very different aspects of the forecast system than developers of researchers. Users often want scores that show how how much money can be saved by making more optimal decisions based on the forecasts. Alternatively, users may want to protect themselves from damagingly large losses by using forecasts to take out protective cover. This is especially the case for extreme events where the costs of taking preventative action are generally much less than the losses that can be incurred if no preventative action is taken and the extreme event occurs (e.g. the costs and losses for New Orleans for Hurricane Katrina). An interesting limit occurs for rare catastrophic events that have vanishing probability of occurrence and vanishing cost/loss ratios – the value of forecasting systems for such events demands more attention.

## *5. Problems and possible new approaches*

Conventional verification scores can have several problems when used to assess forecasts of simple extreme events:

1. Trivial non-informative limits of skill scores going to either 0 (e.g. threat scores) or 1 (e.g. proportion correct) for rarer events. The majority of skill scores traditionally used to verify forecasts of rare binary events, such as the Equitable Threat Score and the Brier score, have the disadvantage that they all tend to zero for vanishingly rare events. This base-rate effect creates the misleading impression that rare extreme events cannot be forecast successfully no matter which forecasting system is used;
2. Large sampling uncertainty due to infrequent occurrence of the events. Because of the rarity of the events, much longer verification periods are needed to obtain reliable verification statistics. For extremely rare events (e.g. rarer than 20-year return periods) even historical records would be too short to allow a meaningful verification of the forecasts (if they had been made over the whole period!). The amount of data you need to verify depends strongly on the predictive power of the forecast. However, larger data sets also bring with them the potential problem of non-stationarity/inhomogeneity;
3. Verification statistics can have strange sampling properties when they are based on contingency tables having small (or even zero) cell counts (sparse tables; see section 9.8 of Agresti (2002). Sparseness in bin counts can also cause stratified statistics to be extremely noisy and non-informative (e.g. reliability diagrams);
4. Susceptibility to large biases. Rare event forecasts are often hedged to avoid missing warnings of severe events and this often leads to a large frequency bias;
5. The presence of extreme values can strongly influence statistics used to summarise exceedances. Non-resistant statistics such as mean and standard deviation can be unduly influenced by the presence of large outlier values.

Several approaches can be envisaged to try to alleviate such problems. One possible approach is to try to reduce the rarity of the event by considering large spatial regions and/or longer time periods. However, when doing this kind of pooling, one should take account of possible variations within either the spatial domain or time period. This requires the development of statistical models that have spatio-temporal explanatory variables. Pooling over inhomogeneous data can inflate skill (Hamill and Juras, 2006; submitted).

The trivial limit of scores can be avoided by using more appropriate association measures for extreme events such as the Extreme Dependency Score proposed by Stephenson et al. (2006). Stephenson et al. (2006) developed a simple asymptotic model for rare binary event forecasts and then used it to demonstrate that the vanishing of the scores is due to their explicit dependence on the base-rate. Inspired by recent work in bivariate extreme value theory (Coles et al. 1999), they proposed an alternative new skill score, known as the Extreme Dependency Score (EDS), for the assessment of skill in deterministic

forecasts of rare binary events. The EDS does not depend on the bias of the forecasting system and has the desirable property that it cannot be improved by hedging the forecasts (i.e. under- or over-forecasting the occurrence of the event). Despite scores vanishing for rare events, the relative skill as judged by association measures such as the odds ratio can actually improve compared to climatological forecasts (Goeber et al. 2004). The relative improvement over climatology has also been recently used to develop operational warning systems for extreme events (Lalaurette, 2003).

Another approach that has not yet been attempted (personal communication, Dr Chris Ferro), would be to *infer* the skill of rarer more extreme events from the skill of less rare moderately extreme events. Such extrapolation requires the use of regularity assumptions similar to those used in extreme value tail distributions such as Generalised Extreme Value and Generalised Pareto distributions. However, such assumptions are not valid for extreme weather that does not exist in less extreme forms (e.g. tornadoes, hurricanes), or that are qualitatively different due to non-linear feedbacks (e.g. heat waves, fog). In addition to giving some indication of skill for the higher thresholds, issuing forecasts of exceedances at lower thresholds can provide forecasters with useful experience and feedback in forecasting and interpreting more extreme events (personal communication, Kees Kok). In order to have sufficient numbers of events for longer range forecasting systems such as current operational seasonal forecasting systems, forecasts are issued for (moderate) extreme events that are not very rare. For example, tercile categories defined by the 0.15 and 0.85 empirical probabilities (the 6-year return period) are currently used at the Met Office and ECMWF for defining such events as extremes in summer mean temperatures (personal communication, F.J. Doblas-Reyes).

Various methods have been developed to deal with sparse contingency tables yet these have not yet been employed in verification studies. For example, one simple smoothing approach is to add a small constant to each cell count that can help reduce bias in measures of association (skill) such as the odds ratio (see section 9.8.7 of Agresti (2002)). More work needs to be done on developing and employing verification measures that are less sensitive to sparse counts and also large outlier values in excesses.

Surprisingly, many of these approaches have not yet received much attention from the atmospheric science community. There is much still to be done on the verification of extreme event forecasts.

## 6. Summary

This report has briefly defined what is meant by extreme events and has then considered how one could issue forecasts for a sub-class of such events known as simple extreme events. Even for simple extreme events, there are several different aspects that can be forecast either in a deterministic manner or probabilistically. Problems with conventional scores for such events have been presented such as trivial limits, sampling uncertainty, bias, and frailty. Several possible approaches have been presented for how one might try to alleviate such problems. With the advent of higher-resolution forecasting systems and the growing need for forecasts of high-impact events, there is a pressing need for more work on such novel verification approaches.

These problems will be addressed further in the ENSEMBLES project by partners in the validation research theme (RT5). More careful attention needs to be given to statistical inference of scores for extreme events and perhaps new approaches could be developed for the verification of probabilistic forecasts of marked point processes.

## Acknowledgements

This work has benefited from fascinating discussions with several colleagues. In particular, I would like to thank Francisco Doblas-Reyes, Chris Ferro, Martin Goeber, Ian Jolliffe, Caio Coelho, Kees Kok, Robert Mureau, Geert Jan van Oldenborgh, and Cristina Primo for their useful comments.

## *References*


Agresti, A., 2002: Categorical Data Analysis (2nd edition), Wiley, 710pp.

Bøe, A.-G., D.B. Stephenson, and S.O. Dahl, 2006: Point process methods for the diagnosis of extreme events in palaeoclimate records: Norwegian mega-floods in the Holocene, Quaternary Science Reviews (submitted).

Coles et al. (1999) "Dependence measures for extreme value analyses". Extremes 2:4, pp 339-365.

Cox, D. R., and Isham, V. (2000). Point Processes. Chapman & Hall/CRC, New York.

Daley, D. J., and Vere-Jones, D. (2002). An Introduction to the Theory of Point Processes. 2nd, Springer.

Diggle, P. J. (1983). Statistical Analysis of Point Processes. Chapman & Hall, London.

Doswell et al. (1990) "On summary measures of skill in rare event forecasting based on contingency tables" Wea. & For., vol 5, pp576-585.

Göber, M., C.A. Wilson, S.F. Milton, D.B. Stephenson, 2004: Fairplay in the verification of operational quantitative precipitation forecasts, J. Hydrology, 288, 225-236.

Jolliffe, I.T. and D.B. Stephenson 2003: "Forecast Verification: A Practitioner's Guide in Atmospheric Science" (Editors), Wiley and Sons, 240 pp.

Lalaurette F , 2003:Early detection of abnormal weather conditions using a probabilistic extreme forecast index, Q J ROY METEOR SOC 129 (594): 3037-3057 Part A .

Lindsey, J.K. 2004: Statistical Analysis of Stochastic Processes in Time, Cambridge University Press, 338 pp.

Marzban, C, 1998: Scalar measures of preformance in rare-event situations", Weather and Forecasting, vol 13, pp 753-763.

Murphy, A.H. 1991a: Comments on “On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables”. Weather and Forecasting: Vol. 6, No. 3, pp. 400–402.

Murphy, A.H., 1991b: Probabilities, Odds, and Forecasts of Rare Events. Weather and Forecasting: Vol. 6, No. 2, pp. 302–307.

Schaefer, J.T., 1990: The critical success index as an indicator of warning skill, Weather and Forecasting, vol 5, pp 570-575.

Stephenson, D.B., B. Casati, and C.A. Wilson, 2006: The Extreme Dependency Score: A New Non-vanishing Verification Measure for the Assessment of Determistic Forecasts of Rare Binary Events, Weather and Forecasting (to be submitted).