Weather Science

White Paper Review on the Verification of Warnings

Technical Report No. 546

D.B. Stephenson, I.T. Jolliffe, C.A.T. Ferro
Exeter Climate Systems,
University of Exeter

C.A. Wilson, M. Sharpe, M. Mittermaier,
T.D. Hewson (on secondment at ECMWF),
Met Office

email: nwp_publications@metoffice.gov.uk

Executive Summary

Many different types of warning are routinely issued by National Weather Services. This joint University of Exeter/Met Office white paper presents a summary of the types, form and criteria of warnings issued by the Met Office over the period 2007-2009. It then critically reviews the methods used to assess the quality of these deterministic forecast products. Recommendations are made as to how warnings may be better issued and evaluated in the future.

Some of the key recommendations are that:

1. Weather services should provide clearer scientific documentation of the methods used to define all warnings, verifying observations and compound events. It would be useful for users, forecasters, and other scientists to be able to access these documents easily via the public web site;

2. All quoted verification measures should be accompanied by estimates of sampling uncertainty (e.g. by providing probabilisitic interval estimates). Targets for measures should make allowances for sampling variations in measures that can occur due to intrinsic variations in weather over the short time periods used for verification;

3. The “confidence” (1-FAR) of warnings usually tends to zero for rarer events, and therefore high confidence is unachievable for rare events. High confidence targets of 0.8 are unachievable for warnings of rare events and therefore make no sense as performance targets;

4. The verification of warning events is complex and more fundamental academic research needs to be done to develop new more appropriate verification measures for warnings. Several suggestions are made as to how this might be achieved.

This review emerged from intensive discussions with Met Office staff over the period April 2007 – April 2009 and so reflects operational procedures over this period.

Acknowledgements

We are grateful to Dr. Harold Brooks, Dr. Bob Glahn, Dr. Martin Göber, and Dr Brian Golding for helpful comments during this work. DBS would like to thank the Met Office for joint funding of his Chair that has made this collaboration possible.

Table of Contents

1. Background

The Public Met Service is defined as 'The provision of a coherent range of basic weather information and *weather related warnings* [our italics] that enable the UK public (and professional bodies as appropriate) to make informed decisions in their day to day activities (to optimise or mitigate against the impact of the weather) and to contribute to the protection of life, property and basic infrastructure.' (Anon, 2006).

Weather related warnings are high-confidence deterministic forecasts of when future severe weather events are likely to occur with high confidence. Warnings are simple to communicate and are routinely issued by most National Weather Services. In 1861, the Met Office introduced the first British storm warning service for shipping in order to improve safety at sea. By 1911, in addition to coastal waters, the North Atlantic was covered by telegraphic broadcasts of gale warnings. The Met Office still currently issues many types of warning such as National Severe Weather Warnings; Marine Warnings; Heat Health Warnings; Aviation and Defence Warnings, and Open Road Warnings. These types of warning are briefly described in Appendix A.

The simple warning-in-effect/warning-not-in-effect binary format of warnings allows the public to take quick action to minimise potential losses. Ideally, one should issue probabilities as this provides the forecast user with more information and allows the user to combine this information with their own costs and potential losses to come to an optimal decision for that user. However, many users prefer the weather service to implicitly make the decision for them by issuing deterministic forecasts, so such forecasts are likely to continue for the foreseeable future. Despite their ubiquity, it is surprising to note that there are very few published articles dealing explicitly with how best to verify warnings.

It is therefore of interest to review how such warnings are defined and evaluated and how these operations might be improved. This white paper report summarises joint collaborative work undertaken in 2008/9 by the University of Exeter and the Met Office. The main aims of this report are to:

- briefly summarise the types and formats of warnings issued by the Met Office;
- review how warnings are currently verified at the Met Office;
- identify verification issues that need to be addressed;
- make recommendations for good practice and future work.

Section 2 of the report describes how warnings and verifying observations are defined. Section 3 then goes on to critically review how warnings are currently verified. Section 4 summarises the key issues and conclusions and recommendations are presented in Section 5.

2. Structure of warnings and verifying observations

2.1 Data structure of warnings

Table 1 below shows an example of coastal strong wind warnings issued for the inshore waters area of Rattray Head to Berwick in the first three days of January 2007.

Y0	M0	D0	HHMM0	Y1	M1	D1	HHMM1	Y2	M2	D2	HHMM2
2007	01	01	0432	2007	01	01	0432	2007	01	01	1800
2007	01	01	1624	2007	01	01	1624	2007	01	02	0600
2007	01	02	0409	2007	01	02	0409	2007	01	02	1800
2007	01	02	1616	2007	01	02	1616	2007	01	03	0600
2007	01	03	0429	2007	01	03	0429	2007	01	03	1800
2007	01	03	1451	2007	01	03	1451	2007	01	03	1800
2007	01	03	1512	2007	01	03	1512	2007	01	04	0600

Table 1: Example of imminent coastal strong wind warnings issued for the inshore waters area of Rattray Head to Berwick in the first three days of January 2007. The first 4 columns are when the warnings were issued, the next 4 columns are when the warnings are due to start, and the final four columns are when the warning is due to end. Y=Year, M=Month, D=Day, HHMM=hours:minutes.

Warnings generally have the following structure. A warning is issued at time t_0 that severe weather is likely to occur in a specific region during a period starting at time t_1 and ending at time t_2, where $t_0 \le t_1 < t_2$. Met Office warnings are generally issued when a meteorological variable, either within a geographical region (e.g. severe weather and marine warnings) or at a specific site (e.g. defence warnings), is forecast to exceed a pre-defined threshold. There is an interesting problem for heavy rainfall events of whether to define the start time by the time of the first threshold exceedance or by when the rainfall event (e.g. convective storm) first appears. The difference $(t_1 - t_0)$ is the lead-time of the forecast and (t_1, t_2) is the period for which the warning is in effect. Multiple warnings with different (t_0, t_1, t_2) can be in effect at the same time, for example, early (i.e. large lead-time) gale warnings can often overlap imminent (i.e. zero lead-time) gale warnings. Time t_0 is well-defined, but (t_1, t_2) may be less so. For marine and severe weather warnings t_1, t_2 are given as clock times, usually to the nearest hour. In summary, a set of m warnings is a set of times (t_{0j}, t_{1j}, t_{2j}) for $j = 1,2,\ldots,m$ where the warning-in-effect time intervals $[t_{1j}, t_{2j}]$ can in principle overlap. The lead-time $(t_1 - t_0)$ and the duration of the warnings $(t_2 - t_1)$ are of great relevance to the forecast user and so should be properly assessed by the forecast verification.

2.2 Verifying observations

Definition of observed binary events against which warnings can be verified is problematic for several reasons:

- **Observational uncertainty:** There are uncertainties in the observations due to measurement and representation errors. This is particularly important for remotely-sensed observations;
- **Sparseness of true observations:** Very few genuine observations may be available, and their distribution within regions may be far from ideal (e.g. wind-speed observations over sea). It may therefore be necessary to use surrogate observations based on Met Office analysis nowcasting products from UKPP (UK Post-Processing). UKPP is the 2km resolution nowcasting system which replaced the operational nowcasting system known as Nimrod) – it blends radar, satellite and mesoscale numerical model products to enable the real-time generation of short period precipitation forecasts;
- **Spatial coverage:** Similarly, in assessing whether an event has occurred in a warning region, is an exceedance at one point within the region sufficient to record occurrence, or should there be a more stringent criteria. Large differences in sizes of regions for which forecasts are issued again seriously complicate the issue (Wilson, 2008) and it might be a good idea to try to compensate for this by considering warning rates per unit area (Hewson and Waite, 2008);
- **Temporal coverage:** For threshold exceedance events, should a single exceedance between t_1 and t_2 be sufficient to record the occurrence of the event, or should the threshold be exceeded for more than a given proportion of the period (t_1, t_2)? This decision is complicated by the differing lengths of periods for which warnings are issued.

Observations may be sparse, poorly distributed in space and time, and prone to large measurement errors. Hence ‘surrogate’ observations are often used instead for verification purposes. Wilson (2008) describes four different types of ‘observation’, namely real observations at permanent stations, Nimrod analysis on a 15km x 15km grid, UKPP analyses on a 2km x 2km grid, and ‘virtual observations’ obtained by local adjustment to UKPP. He finds that the assessed value of forecasts seems to be very dependent on the type of verification observation (truth) used, so the choice of which ‘observations’ to use can make a large difference to the perceived quality of a set of forecasts. Such choice can also allow the dangerous possibility of choosing the observations in a way that gives the best verification scores.

A problem for warnings for geographical regions is that observations may only be available at fairly arbitrary, unevenly-spaced, points in space. A related matter, which is also relevant to discrete sampling in time, is that an event not being observed does not necessarily mean that it did not happen (e.g. small-scale features such as tornadoes can be missed). Hence, hit rates based on observation networks underestimate the true hit rate for small-scale systems. Although not relevant when observations are automatic, Barnes et al. (2007) suggest that when there is reliance on volunteers or unofficial observations (e.g. in N. America), there may be tendency to look harder for an event when it has been forecast than when it has not.

3. Verification of warnings

Verification has three potential audiences: administrators who want measures that are easy to use for setting performance targets, scientists who want measures that give useful feedback on how to improve the forecasting system, and forecast users who have to take actions based on the warnings so as to minimize potential losses. Both administrators and scientists would like robust verification procedures that are easy to implement. However, the ideal forecast format for users may be one for which verification becomes complicated. Therefore, verification procedures should not solely determine the forecast format, but conversely forecast formats that make verification awkward should be avoided if at all possible.

3.1 Which warnings are verified?

Not all warnings issued by the Met Office are currently verified or have verification results that are routinely interpreted:

- National Severe Weather Warning Service (NSWWS) – flash warnings for severe gales and heavy rain are verified (Wilson 2008). Despite the statement 'verification of warnings is an important element of the requirement [of the NSWWS]' (Anon, 2006), some other types of flash warning and early warnings are no longer verified;
- Marine warnings – gale and Coastal Strong Wind Warnings (CSWW) are currently verified (Sharpe, 2008a). Verification of the latter is likely to be phased out soon, as the customer no longer requires it. Storm warnings are not currently verified;
- Defence weather warnings – these are verified, except for thunderstorm warnings. The measures used are aggregated over all parameters forecast (except thunderstorms);
- OpenRoad Warnings – some aspects of these are verified, but because of their different nature they will not be discussed further.

Furthermore, little information on the Met Office's verification of warnings is readily available outside the Met Office. There is some information on the public web pages **www.metoffice.gov.uk/roads/openroad.html**, but some of this is out-of-date (**www.metoffice.gov.uk/corporate/verification/gale.html**). Part of the reason for this limited public information is that some of verification is paid for by specific clients.

3.2 Counting of compound events: hits, misses and false alarms

Rather than consider mutual timings of warnings and observed events, weather warnings are usually evaluated by counting the number of compound events:

- Hit – event observed while a warning is in effect;
- Miss – event observed while no warning is in effect;
- False alarm – event not observed while a warning in effect;
- Correct rejection – event not observed while a warning is not in effect.

If one can define unambiguously when a warning is/is not in effect and when the event is/is not observed, then it is possible to count the number of compound events to obtain four counts, $a, b, c,$ and d, which can then be used to define a 2×2 contingency table:

		Observed		
		Event	No Event	Total
	Warning in effect	$a = npH$	$b = np(B - H)$	$a + b = npB$
	Warning not in effect	$c = np(1 - H)$	$d = n(1 - p(1 + B - H))$	$c + d = n(1 - pB)$
	Total	$a + c = np$	$b + d = n(1 - p)$	$n = a + b + c + d$

Table 2: Contingency table of counts expressed in terms of base rate $p = (a + c)/n$, *hit rate* $H = a/(a + c)$, *frequency bias* $B = (a + b)/(a + c)$, *and total number of events* $n = a + b + c + d$.

The base rate, $p = (a + c)/n$, is the proportion of the total number of events when the event was observed. The base rate tends to zero for rarer events, which has important consequences for the commonly used verification measures (see Section 3.3).

Compound events are countable if the warnings and observations are recorded at regular sampling times. For example, for hourly observations of wind-speed above a predefined threshold, one can count the number of exceedances and non-exceedances when warnings are and are not in effect, and hence compile the contingency table. However, it should be noted here that the counts represent numbers of hours rather than the number of distinct meteorological events.

Meteorological events such as tornados and other storms occur sporadically in time. The rate of such point events can be estimated by dividing the number of events in a given time interval by the length of the time interval. Rates can also be calculated conditional upon when warnings are and are not in effect. However, it is problematic to calculate the rate of observed non-events since the duration of a non-event is undefined. Similarly, for irregular duration warnings, it is impossible to count the number of "no warning in effect" events since one doesn't know how long such non-events last. It may be easy to see that there are 5 tornados in a month for a given area, but much more difficult to say how many 'no tornado' events there were (Dr Harold Brooks, personal communication). There is no easy answer to this, though one *ad hoc* method is to count the non-events by dividing the length of the non-event periods by the average length of the event periods.

The problem with counting non-observed events and no-warnings, has led meteorologists to realise that the number of correct rejections can not be calculated reliably for weather warnings. In most of the Met Office verification of warnings no value of d is available. If d were available in addition to either b or c, then it would be possible to calculate the base rate, respectively, for observed non-events and no-warnings! Furthermore, if d were available it would open up the possibility of using many other verification measures such as Peirce's skill score, Heidke's skill score, the equitable threat score, the odds ratio and extreme dependency score – see Mason (2003), Stephenson et al. (2008). Wilson (2008) recommends that, in future,

forecasts and observations should be made in such a way that d is available e.g. by using regularly sampled observations and warnings.

Brooks (2004) discusses tornado forecasting and notes that given $a, b, c,$ then d can be found if the base rate p is known, so he attempts to find an estimate of the base rate from which to infer d. His base rate is not the overall occurrence of tornados, which is extremely small, but their occurrence in circumstances where a tornado might conceivably have been forecast (the 'difficult' cases). Mason (1989) also suggests restricting attention to only 'difficult' cases in order to estimate d, which he calls the 'no-no frequency'. An analogy for heat health warnings is that it is unreasonable to include winter months in compiling a contingency table, and indeed the Heat Health Watch system is only in operation during the summer months. In winter the chance of exceeding the required temperature threshold is vanishingly small and including winter months would inflate d and dominate many verification measures. Glahn (2005) points out two difficulties with the Brooks (2004) suggestion. There is the question of how to estimate the base rate for the 'difficult' cases and also how to decide the threshold between easy and difficult cases. The value of many verification measures will depend on the choice of two thresholds: where to draw the line between easy and difficult situations as well as how large the probability of the event needs to be before a warning is issued. If there is uncertainty about the choice of the easy/difficult threshold it would be possible to plot the value of a chosen verification measure as a function of base rate or, equivalently though perhaps less intuitively, as a function of d.

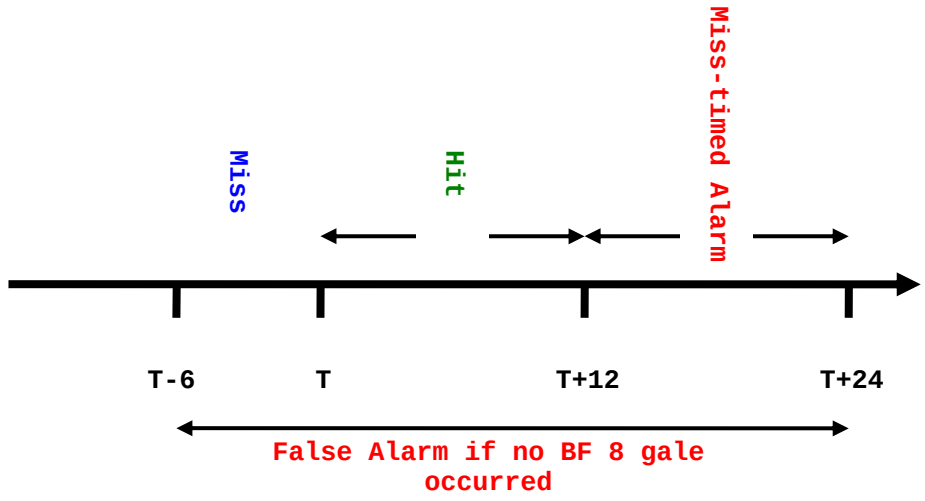

Figure 1: Example of how compound events are defined for an imminent gale warning. Issue and start time $t_0 = t_1 = T$, and end time $t_2 = T + 12$.

Source: Sharpe (2008).

The basic definitions of hits, misses and false alarms are often modified. For example, for an imminent Met Office gale warning shown in Figure 1, the period for judging a hit is from T to T + 12, whereas the period for judging a false alarm is T – 6 to T + 24. Hence, hits and misses are not evaluated on the same basis and so the resulting 2×2 table of counts is not a proper two-way cross-classification.

Furthermore, new types of compound event can be defined. For example, Sharpe (2008b) creates a large number of modified categories and measures based on amalgamating them in various ways, with a view to seeing which definition of a measure has better values. His argument is that with the traditional definitions, users feel that the quality of the warnings is better than the values suggest, but this could be because the measures are not being properly interpreted. Sharpe (2008a) introduces several new compound events: he subdivides missed events according to temporal distance from the warning, a category is introduced (near-hits?) for which the event occurred but in a smaller proportion of the area than required for a hit, false alarms are subdivided according to the maximum Beaufort strength recorded, and mistimed alarms are subdivided. Sharpe (2008b) gives a number of different definitions of the 'event' that corresponds to a successful forecast. The fact that some definitions give 'better' values of a verification measure than others means that the events corresponding to those definitions are more successfully forecast according to that measure.

Wilson (2008) notes that different wind gust thresholds may be used for hits and false alarms, and the definitions of hits and misses for the Met Office defence weather warnings are not straightforward, with near hits, and 'non-issues' also defined. For defence warnings, which have probabilities associated with them, the elements can even be non-integer. Hewson and Waite (2008) mention the idea of 'partial hits' when rainfall fails to reach the threshold indicated by a warning, but is only slightly below it. Barnes et al (2007) refer to false alarms, unwarned events (misses) and perfect warning (hits) and propose additional categories, namely 'underwarned events' and 'overwarned events', or even a continuum to replace the small number of categories. However, no detail is given on how to implement these ideas. Another complexity with defence warnings is that some stations are only "open for business", and therefore able to issue warnings, at certain times of the day.

Such extensions to the basic 'hit', 'miss' and 'false alarm' categories means that the rows and columns of a table may have more than two categories. Some of these extended categories are collapsed in order to give the 2×2 table format for Marine warnings and Defence weather warnings (Anon, 2003, Sharpe, 2008a). The statistical properties of the verification measures based on such collapsed tables will be different from those in a genuine two-way classification and merit further investigation.

3.3 Verification measures

Rather than consider mutual timings of warnings and observed events, weather warnings are usually evaluated by counting the number of compound events:Because estimates of d are either unreliable or unavailable, verification measures for warnings are often limited to those measures based only on a, b, c. According to Wilson (2008), the current measures used for verification of severe weather warnings at the Met Office are:

- Hit rate: $H = a/(a+c)$ (also known as probability of detection)
- False alarm ratio: $FAR = b/(a+b)$
- Threat score: $TS = a/(a+b+c)$ (also known as Critical Success Index)
- Frequency bias: $B = (a+b)/(a+c)$

Marine warnings and defence weather warnings use the first three of these (Anon, 2003, Sharpe 2008a), though the *false alarm ratio* is often erroneously referred to as the *false alarm rate* $b/(b+d)$ (e.g. Brooks, 2004). The frequency bias can be obtained from the identity:

$$B = \frac{1 - FAR}{H}$$

In other words, the *confidence*, $1 - FAR$, divided by the hit rate. Further information about these measures and their properties can be found in Mason (2003). A good warning system should ideally have a high hit rate, a low false alarm ratio (high confidence), a high threat score, and a frequency bias close to one.

FAR plays a special role in NSWWS, as the criterion for warnings to be issued is that the *confidence*, an estimate of the probability of an observed event given a warning, should exceed some target value. The target value can vary, but a typical value is 0.8, so that FAR is required to be smaller than 0.2 for warnings to be issued. Hewson and Waite (2008) note that for heavy rainfall warnings, FAR is often greater than 0.5, so that a confidence of 80% is currently unachievable and should be lowered. This is a consequence of the rarity of the event as will be demonstrated in the following section.

Hewson and Waite (2008) have suggested defining the *deterministic limit* as the lead time beyond which $b + c > a$ (or equivalently $TS < 0.5$). In other words, the lead time beyond which the number of misses plus false alarms exceed the number of hits.

3.4 Rare events: is high confidence achievable or even desirable?

By their nature the events for which warnings are issued are relatively rare at a given location. The degree of rarity varies for different events – for example, the events implied by 'red' early or flash warnings are more disruptive and will be rarer than those corresponding to 'amber' warnings.

Following Stephenson et al. (2008), Table 2 shows how the hit rate, confidence and threat score behave as a function of base rate. All three measures usually tend to the trivial limit of zero[1] for rarer events with $1 - FAR \to H / \beta$ and $TS \to H /(1 + \beta)$ where β is the finite value of the bias B in the limit as the base rate $p \to 0$. In order to find the non-trivial exponent, δ, it is necessary to have an estimate of the base rate p, which is not possible if we only have $a, b,$ and c. Since $1 - FAR \to H / \beta$, the only way to achieve a high confidence for increasingly rare events is for the warning system to have a decreasing frequency bias.

[1] *theoretically it is possible to have non-zero limits if there is strong extremal dependency – see Ferro (2007).*

Measure	Definition	Score in terms of bias and hit rate
H	$\frac{a}{a+c}$	κp^{δ}
1-FAR	$\frac{a}{a+b}$	$\frac{\kappa p^{\delta}}{\beta}$
TS	$\frac{a}{a+b+c}$	$\frac{\kappa p^{\delta}}{1+\beta-\kappa p^{\delta}}$

Table 2. Limits of some standard measures used for warnings as the base rate p tends to zero: Hit rate ($H = \kappa p^{\delta}$ where $0 < \delta \leq 1$), Confidence (1-FAR), and Threat Score (TS). The bias B is assumed to tend to a finite value of β as $p \to 0$.

The more disruptive the event, the more the benefit of over-forecasting it in order to reduce the risk of large losses caused by missed events. Anon (2003), in discussing defence warnings, writes:

> '*it is not the intention of those attempting to obtain a measure of the performance of an office to influence the process of issuing such warnings, which are expected to err on the side of flight safety'.*

In other words over-forecasting is acceptable if it increases safety, even if it degrades the apparent performance of the forecasts according to the chosen verification measures. A 'skewed' loss function leads to German forecasts of strong winds being over-forecast, especially for forecasts of storm and violent storm force winds (Dr Martin Göber, personal communication).

For NSWWS, users costs and losses are generally not quantitatively incorporated into the guidance for issuing warnings, but that guidance reflects that the ratio of cost to loss decreases for increasingly disruptive events. The following guidance comes from Anon (2006):

- Flash messages for wind, rain and snow will be issued when there is at least 80% confidence in the occurrence of [the event];
- An early warning of an 'amber' event will normally be issued … whenever the overall risk of [the event] is 60% or greater;
- An early warning of a 'red' event will normally be issued … whenever the overall risk of [the event] is 20% or greater.

The different thresholds here seem to reflect changes in costs and losses, partly due to severity of the event and partly due to lead-time. Interpreting 'confidence' and 'risk' as 'probability', the three thresholds correspond to FAR values of 0.2, 0.4, 0.8, respectively. The largest of these three values for FAR intuitively suggests very poor performance, indicating that the warnings often 'cry wolf'. However, this is not necessarily undesirable for severely disruptive events. Barnes et al. (2007) suggest that there is little evidence that a high value of FAR causes users to disregard warnings, when the warning is for a severe event.

Turning to the lowest value of FAR used (0.2) it is very likely that this is unachievable. As mentioned in Section 4.2, Hewson and Waite (2008) note that for heavy rainfall FAR is often greater than 0.5, so that a confidence of 80% is currently unrealistic and should be lowered. Barnes et al. (2007) quote FAR values for a number of different types of event, albeit all different from those described in this document, and all are well above 0.2.

In summary, high confidence (i.e. not crying wolf) can be achieved by issuing fewer warnings than observed events (i.e. frequency bias less than one), but then this under forecasting can compromise safety by failing to warn about events. Hence, overly high confidence targets for warnings of rare severe events might not be in the public interest.

4. Summary of key issues

From the previous discussion, it is clear that the production and verification of warnings are complex activities that involve many difficult choices. Some of the main issues that have emerged from this work are:

1. *How to define the warnings and verifying observations*
   Warnings are issued for weather in pre-specified spatial regions over specific time periods. It is not obvious how to choose either the spatial regions or the length of the warning periods.

   Geographical regions need to be small enough to represent homogeneous weather conditions, and also be useful to local forecast users. However, one also requires regions large enough to be able to capture observed events and hence allow verification. The size of the region will affect the base rate of the event – there will be more chance of severe weather occurring somewhere in an area if that area is large. This base-rate effect will cause large differences in verification measures when considering warnings over a set of very different sized regions (e.g. coastal wind warnings). Because of this, Wilson (2008) has suggested making the geographical areas for severe weather warnings more equal size, provided that this can meet administrative decision-making constraints (e.g. regional counties).

   Similarly, the longer the warning is in effect, $t_2 - t_1$, the more likely it is that the event will be observed, and hence the hit rate will be increased. The hit rate can also be increased by reducing the lead time for warnings (Dr Martin Göber, personal communication). To help discourage such practice, a warning with a lead time of less than 3 hours is deemed a ‘miss’ in most circumstances in the verification of Met Office defence warnings.

   Similarly, these spatial and temporal considerations make it problematic to define suitable verifying observations. In addition, there are also issues with observations due to sparseness, observational uncertainty, and choice of data set (see Section 2.2).

2. *Estimation of compound warning-observation counts*
   The conventional approach to verifying warnings is based on measures calculated from counting the number of compound warning-observation events i.e. the numbers of hits, misses, and false alarms. However, there are several major unresolved issues in how to obtain such counts:

- Definition of compound events. There are many different ways to define these based on the warnings and observed events, yet not all of these approaches lead to counts that are 2-way classifications of the events;
- Counting of events. There is a problem with how to count non-events for irregularly sampled data that leads to d being unavailable;
- Duration of events. Should one consider a single storm that lasts over several warning periods as one event or a collection of several shorter events?

See Section 3.2 for more details.

3. *Disadvantages of conventional warning verification measures*
The conventional measures of hit rate (H), false alarm ratio (FAR), threat score (TS) and frequency bias (B) have several disadvantages:
   - Any score based on a ratio of linear combinations of a, b, c can be written as a function of H and B. Hence, FAR and TS are re-expressions of H and B and so do not provide any additional information;
   - Only two numbers are required to describe any possible ratio of the three counts a, b, c and so H and B suffice;
   - H, 1-FAR, and TS all usually tend to zero for rarer events, and hence are not overly informative for warnings of rare events;
   - High confidence targets for 1-FAR are unachievable for rarer events unless the bias decreases below 1, which could result in few if any warnings ever being issued. Typical biases are found to be as high as 5 for severe weather warnings (Hewson and Waite, 2008). Barnes et al (2007) also criticise the use of FAR because apparently large values do not accurately reflect either the forecaster's or the public's views of the merits of forecasts;
   - H, 1-FAR, and TS are all strongly base-rate dependent (Mason, 1989) but unfortunately the base rate cannot be estimated unless we also have the number of correct rejections, d. Scores such as TS depend strongly on the number of observed events in the verification period, and hence are prone to large sampling errors (Wilson, 2008);
   - These count-based scores do not provide simple user-relevant feedback on the timing skill of the warnings. For example, how skilful were the warnings at different lead times, and are the durations of the warnings realistic?
   - The scores do not give an idea of the value of the warnings to forecast users since they ignore how users will use the warnings. Different forecast users will have different cost/loss ratios for warnings with different lead times.

   See Section 3.3 for more details.

4. *Sampling uncertainty on verification measures*
Due to the rarity of events, verification measures for warnings are prone to large amounts of sampling uncertainty. Despite this, verification measures are often quoted without any indication of their uncertainty. This is bad practice. Ideally, a confidence interval or some measure of sampling uncertainty should be given whenever the value of a verification measure is calculated. References on how to find such intervals are **www.ral.ucar.edu/~ericg/Gilleland2008.pdf** and Jolliffe (2007). Confidence intervals can be confusing, so special care is needed in explaining how they should be interpreted.

Sampling uncertainty should also be taken into account when assessing whether a performance target is met. For example, if a target is TS > 0.46 (Wilson 2008), a failure will be recorded if the sample TS < 0.46. An important question is 'what is meant by this target value of 0.46?'. Even if the (population) target is achieved exactly, the sample value will be lower than the population value roughly half the time, and failure to achieve the target (even though it has been achieved) will be recorded on some of these occasions. It would be worth taking into account the available sample size for verification when setting targets.

The simple example that follows illustrates the problem. Consider hit rate H, as the assumptions underlying probability calculations are somewhat simpler than for TS. Suppose that there is a target of 0.8 for H and that 1000 occurrences of the event of interest are available. If the 'true' population value of H is 0.8, there is a roughly 50% chance that the sample value will be less than 0.8, in which case it will be deemed that the target has not been met. For there to be a 90% chance or more that the sample value of H exceeds 0.8, the population value needs to be near 0.85. A sample size of 1000 is large – for extreme events it will be much smaller. If it is only 50, then a population value of H of nearly 0.87 would be needed in order to have a 90% chance of meeting the target in a verification sample of 0.8. Such calculations could be used to set thresholds for sample values of verification measures which have a high degree of confidence of being achieved, given a desired population value.

5. *More transparent reporting*
   It is interesting to note that not all issued Met Office warnings are verified, and of those that are, the verification results are generally not widely disseminated outside of the Met Office (Section 3.1). Furthermore, it has become clear while compiling this review that the methods used to define warnings, verifying observations and compound events could be documented more clearly in a way that would allow other scientists to *repeat* the procedures. The details of the verification approaches often exist in operational code without being clearly documented beforehand in a report. One possible reason for this is a current failure to allocate resources for such important activities.

5. Conclusions and recommendations

To conclude, this study has revealed that there is a wealth of complexity in how warnings can be created and evaluated, much of which has received very little academic attention. There are great opportunities to develop improvements in this important area of weather forecasting.

Our main recommendations for better practice and future work are listed below:

1. *Definition of the warnings and verifying observations*
   There should be clarity in the definition of 'a warning event' when it is being forecast. If nature of the event is clearly defined when forecast is made, it should be straightforward to know what is required in order for the forecast to be verified. There are two considerations to take into account when defining the 'warning event': what is convenient to verify and what is of most interests to users. The latter should take precedence, unless it poses exceptionally awkward problems for verification, though different users may, of course, not agree on what is of most interest.

   The choice of verifying observations can create large differences in the values of verification measures. Ideally, the same type of 'observations' should be used over a long period of time, but the models from which the observations are derived change frequently so this may not be possible. When changing between one type of 'observation' and another, extensive comparisons need to be made of the effect of the change on the verification measures, so as to avoid spurious increases or decreases in skill levels.

   Similar considerations hold for spatial issues as for time. However, an additional problem is the large differences in sizes between areas for which warnings are issued. It is highly desirable that for verification purposes areas are combined so as to give as near equal size as possible, but pooling over non-homogeneous areas should be avoided. Notwithstanding this, there are good administrative reasons for the widely differing sized areas, on land if not at sea, so that warnings may still be made for different sized areas even if some are pooled for verification purposes.

2. *Extended compound events*
   Properties associated with verification measures defined for (2 x 2) tables become more complicated when various extra categories are defined, such as 'near-hits' of various varieties, with measures then defined by amalgamating some of the categories. The only valid reason for switching to a new definition is if that definition is more relevant to the forecast user. Choosing a new definition of a measure simply to optimise the score seems unwise. It is likely that the definition giving the best value of a measure for one data set will be beaten by another definition when further data become available. Further research work is needed on investigating these extended categories.

3. *Warning verification measures*
   If the number of correct rejections *d* were available, a much greater number of verification measures would become possible, some of which are more informative than those that are currently used. Therefore, it would be useful to consider defining events and warnings so that the (2x2) cross-classification becomes appropriate. For example, every 6 hours, check whether or not an observed event has occurred in the previous 6 hours and whether or not a warning was in force during that period. If *d* is undefined because the definition of hits, misses, and false alarms can not be written as the 2-way intersection of warning and observed events, more research is then needed to find alternative verification scores to ones based on an inappropriate (2x2) cross-classification.

   A more general, but related, point, is that if a new verification score is suggested and deemed to have improved properties compared to an existing score but appears to give 'poorer' scores than the old one, this should not prevent its adoption. Retrospective use will show whether the latest forecasts really are worse

   To issue a warning, there most be a certain level of confidence that the warning event will occur. However, with the exception of Defence weather warnings, many of the warnings issued by the Met Office do not explicitly provide the user with this probability and so are deterministic. It would be worthwhile recording the 'confidence' of each forecast, so that this aspect of the warning could also be verified in order to help improve the forecasting systems.

   The implications of setting unachievable targets (e.g. confidence greater than 0.8) for rare events should be seriously considered from the viewpoint of the end-users of the warning systems. For good unbiased forecasting systems, the confidence of rare event warnings is likely to be considerably smaller than 0.5 because of the small base rate.

4. *Sampling uncertainty on verification measures*
   Any quoted value of a verification measure should be accompanied by information on the uncertainty due to sampling variation, or other reasons such as measurement error. Similarly when setting targets based on measures, such uncertainty should explicitly be taken into account e.g. by providing interval estimates.

   A promising way to obtain more precise estimates of verification measures is by pooling forecasts and observations over different spatial regions and different time periods. However, in doing so, one has to either pool relatively homogeneous data, or be careful to account for variations within the pool due to fixed effects such as spatial and temporal trends, annual cycle, etc. If it is not possible to achieve near-equal size areas, then weights may be applied, based on the sizes of areas, when computing verification measures aggregated over different areas (Sharpe, 2008b). This is a promising area of verification research that merits more careful attention.

5. *More transparent reporting*
   It would be good practice and would help increase user confidence, if the Met Office were to ensure that verification is performed for ALL warnings that are issued, and then make

these verification results widely available (e.g. via the web). It would also be good practice if all warning procedures and associated verification was clearly documented within the Met Office in such a way as to allow others to repeat the procedures.

6. *Ideas for future research*

This study has revealed several promising areas that could benefit from new research. For example, research that helps answer these questions could be of great relevance:

- How to define and count ALL four compound events (the missing d problem)
- How to develop a more user-relevant verification procedure for warnings that takes into account the timing of events and the forecast users' cost/loss ratios. The appropriate verification really ought to be concerned with the effectiveness of the actions triggered by the issued warnings rather than the forecasts on which they are based;
- How to account for spatio-temporal variations and trends when verifying warnings pooled over all locations and long time periods?
- How to interpret extended categories such as "near miss" and verify them using categorical data analysis techniques?

Unfortunately, despite its relevance, not much public funding is generally available to do verification research (e.g. from research councils). Since verification research would directly benefit National Weather Services in improving the products they deliver, it should perhaps be the responsibility of such organisations to fund this type of research. Some of the areas suggested above would make excellent 3-year projects for either PhD students or post-doctoral researchers co-located at the Met Office and the University of Exeter.

References

Anon 2003: The defence warning assessment scheme (DWWAS). Controlled Met Office document. Unpublished.

Anon 2006: National severe weather warning service. Met Office Internal Document. Unpublished.

Barnes L.R. et al. 2007: False alarms and close calls: a conceptual model of warning accuracy. *Wea. Forecasting*, **22**, 1140-1147.

Brooks, H. E., 2004: Tornado-warning performance in the past and future. *Bull Amer. Meteor. Soc.*, **85**, 837-843.

Ferro, C. A.T., 2007: A probability model for verifying deterministic forecasts of extreme events, *Weather and Forecasting*, vol. 22, no. 5, 1089-1100.

Glahn, B. 2005: Tornado-warning performance in the past and future – another perspective. *Bull Amer. Meteor. Soc.*, **86**, 1135-1141.

Hewson T. and Waite H. 2008. Exceptional rainfall in summer 2007 – an assessment of unified model guidance pertaining to short period warnings. Met Office Internal Document. Unpublished.

Jolliffe I. T. 2007: Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 137-150.

Mason I. 1989: Dependence of the Critical Success Index on sample climate and threshold probability. *Aust. Met. Mag.*, **37**, 75-81.

Mason, I. B. 2003: Binary events. In: *Forecast Verification A Practitioner's Guide in Atmospheric Science* (eds I. T. Jolliffe and D. B. Stephenson), 36-76. Chichester: Wiley.

Sharpe, M. A. 2008a: Marine forecast performance statistics between February 2007 and February 2008. Met Office Internal Document. Unpublished.

Sharpe, M. A. 2008b: Development of the National Severe Weather Warnings verification system: preliminary report. Met Office Internal Document. Unpublished.

Stephenson, D. B., Casati, B., Ferro, C.A.T. and Wilson, C. A. 2008: The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Met. Apps.*, **15**, 41-50.

Wilson, C. 2008: Comments on the verification of flash warnings. Met Office Internal Document. Unpublished.

Appendix A

A.1 National Severe Weather Warnings and Extreme Rainfall Alerts

These warnings are classified in two ways, depending on the lead-time of the forecast and on the anticipated severity of the event. *Early Warnings* are issued up to several days in advance, whereas *Flash Warnings* are issued close in time to the anticipated event. In issuing warnings, there are well-defined meteorological criteria for each type of event that determine whether or not a warning is issued. For Flash Warnings, meeting the meteorological criteria alone is sufficient for a warning to be issued. In the case of Early Warnings, there must be an expectation of severe disruption, in addition to meeting the criteria, before a warning is issued. This takes into account that severe weather, such as extreme wind gusts, may not necessarily cause severe disruption if it occurs at geographically remote locations and/or at certain times of day.

The severity of the event has two levels, *Amber Events*, and the more severe *Red Events.* Flash Warnings at the Red level are known as *Emergency Flash Warnings*. Amber level warnings may be issued for Severe Gales; Heavy Snow; Blizzards/drifting; Heavy rain; Freezing rain, glazed frost or widespread icy roads; Fog; (persistent low) Temperature. Red level warnings are only issued for (extreme levels of) the first three of these types of event. For all except 'Freezing rain, glazed frost or widespread icy roads' there are well-defined numerical thresholds and a warning is issued if the forecast probability of the thresholds being exceeded is greater than a given value. This value varies depending on whether the warning level is Amber or Red and whether it is Early or Flash (Anon, 2006). For example, for severe gales the requirement is that there is confidence of at least 60% that there will be two or more gusts of at least 70 mph within the warning period.

The warnings may be issued at *any* time of day or night. Both the lead-time and the duration of the warning can vary, and the exact start and end time of a warning may not always be precise, for example 'this afternoon', 'late evening'.

Extreme Rainfall Alerts are similar to Flash Warnings in that they are issued for counties and unitary authorities across the UK, and they can be issued at any time and last for any period. However they differ from Flash Warnings in the following ways:

- there are 3 thresholds; 30mm in 1 hour, 40mm in 3 hours and 50mm in 6 hours;
- each warning is for one of these thresholds and is issued with a probability of occurrence;
- daily 24hr advisories are issued with a lead time of 11 hours if the probability equals or exceeds 10%;
- early ERAs (lead time 8-11 hours) are issued if the probability is in the range 20-40%;
- imminent ERAs (lead time 1-3 hrs) is issued if the probability equals or exceeds 40%.

As well as being issued to a number of relevant agencies and authorities, warnings that are currently in force are displayed on the Met Office website at

www.metoffice.gov.uk/weather/uk/uk_forecast_warnings.html

A.2 Marine Warnings

There are three types of warnings, *Coastal Strong Wind Warnings, Gale Warnings* and *Storm Warnings.* Coastal Strong Wind Warnings (CSWWs) form part of the Inshore Waters forecasts that are issued four times per day. The forecasts are for 17 inshore waters areas and a warning is in force for an area if the forecast wind in that area exceeds a threshold (Beaufort Force 6 – BF6). The wording used in the forecasts is such that it can be deduced whether or not a warning is in force for a set of 6-hour periods. Unlike Gale Warnings, which are described next, CSWWs are issued at regular times – they are updated each time an Inshore Waters Forecast is issued.

Gale Warnings are issued separately for 31 sea areas and imply either mean wind speed exceeding a threshold (BF8), or gusts exceeding a higher threshold somewhere in the area. Although the wording of the forecasts may look imprecise, words such as 'imminent' and 'soon' have well-defined meanings, so that it is clear whether or not gales are expected in a set of 6-hour periods following the forecast. Gale warnings may be issued at any time of day or night, but are also included as part of the Shipping forecasts that are issued four times per day - Sharpe (2008a).

Storm Warnings form part of the High Seas forecasts that are issued twice a day, but like gale warnings they may also be issued at other times. The forecasts and warnings are for 13 separate sea areas in the North Atlantic, north of 45º N and east of 40º W. Storm warnings correspond to forecasts of wind speeds exceeding BF10. Storm warnings, gale warnings and CSWWs currently in force are readily available on the Met Office website at

www.metoffice.gov.uk/weather/marine.

A.3 Heat Health Warnings

A Heat-Health Watch system operates in England and Wales from 1 June to 15 September each year **www.metoffice.gov.uk/weather/uk/heathealth/print.html**. Warnings may be issued at three levels and depend on exceeding thresholds for maximum daytime and minimum night-time temperatures. These thresholds vary by region, but an average threshold temperature is 30°C by day and 15°C overnight.

Amber alerts are triggered as soon as the risk is 60% or above for threshold temperatures being reached in one or more regions on at least two consecutive days and the intervening night. *Red heatwave* action is triggered as soon as the Met Office confirms threshold temperatures will be reached in one or more regions. Finally, a *Red emergency* level is reached when a heatwave is so severe and/or prolonged that its effects extend outside the health and social care system. At this level, illness and death may occur among the fit and healthy, and not just in high-risk groups.

A.4 Defence and Aviation Warnings

Defence and aviation warnings are issued by specific Met Office forecast offices and cover a set of adverse weather conditions that partially overlap with those included in the National Severe Weather Warnings (NSWWS). Specifically, the conditions forecast are Strong Surface Wind, Gale, Air Frost, Snow, Snow Accumulation, Fog and Thunderstorms, Volcanic aerosol. Although these are ‘adverse’ conditions, the thresholds used in their definitions are generally not as extreme as the corresponding requirement for a warning in the NSWWS. They also differ from NSWWS in that the forecasts issued are probabilities of the adverse weather event. Except for snow accumulation, this probability forecast consists of a single number, the probability that the adverse conditions will occur. For snow accumulation four probabilities are issued for the four categories Negligible, Light, Moderate and Heavy. Defence forecasts are made for specific stations, rather than for areas as in the NSWWS and Marine Warnings. For more information, see Anon (2003).

A.5 Open Road Warnings

OpenRoad warnings give warnings of conditions likely to disrupt the smooth running of the roads **www.metoffice.gov.uk/roads/openroad.html**. They are somewhat different from other warnings in that various alert states are defined depending to some extent on the users. Direct road temperatures, rather than warnings based on the temperatures, are verified. Hence these warnings will not be discussed further.