Verification of Weather Warnings

Dr Michael Sharpe
Operational verification and systems team
Weather Science

Contents

Summary

Weather warnings are issued in the 147 counties and unitary authorities of the UK. These warning areas vary greatly in size and many of them do not contain any observations. Consequently the accuracy of weather warning services are usually measured by comparing them against hourly nowcast analyses, which are currently available on a 2km resolution grid across the UK. However nowcast analyses are not 100% accurate and may diverge from the truth in respect to intensity, timing and/or position. Therefore the newly developed Warnings Verification System (WVS) introduces near-hit categories to account for these discrepancies when measuring the performance of weather warnings. This has the added advantage of eliminating the double negative score (of a miss and a false alarm) which is often awarded to warnings that are almost correct. The WVS creates a final score by giving partial credit to weather events that are nearly correct in terms of intensity, time and position.

Difficulties Associated with the Verification of Warnings

Weather warnings comprise simply of the time of issue, a start time and an end time. Between the start time and end time of a warning a measurable weather component (such as rain or wind speed) is forecast to exceed a fixed threshold. The most easily measurable weather components are heavy rain and severe gales and the Met Office currently attempts to measure the quality of these warnings through a process of verification. In a simple verification process each event is classified as a

- *hit* if the threshold is exceeded between the start and end time of the warning,
- *missed event* if the threshold is exceeded when a warning is not in force or a
- *false alarm* if a warning is issued but the threshold is not exceeded.

Often warnings can be issued in any of the 147 counties and unitary authorities in the UK. Some counties are very large and some unitary authorities are extremely small. Simply verifying the warnings issued in these areas using the *hit*, *false alarm* and *missed event* definitions (given above) can lead to some very poor results because:

- If a warning is issued and the weather almost exceeds the threshold, the event is still classified as a *false alarm.*
- A severe weather event that occurs after the issue time of a warning but ends before the start time of the warning is classified as a *missed event* and a separate *false alarm* (i.e. a double negative score).

- A severe weather event that starts after the start time of a warning but continues just past the end time of the warning is reported as a *hit* and a separate *missed event*.
- Severe weather that occurs just outside an area in which a warning is issued is reported as a double negative; a *missed event* (in the neighbouring warning area) and a *false alarm* (in the area containing the warning).
- Due to a lack of observations, hourly nowcast analyses (available on a 2km grid of points across the UK) are used as the truth but there is less than 100% confidence in nowcast analysis data because it can contain timing, intensity and location errors.

Intensity, temporal, spatial and confidence issues can make verification scores misleadingly poor. The fully automated Warnings Verification System (WVS) has been developed to address these issues.

Methodology of the new Warnings Verification System

The Warnings Verification System (WVS) provides interactive forecaster feedback in addition to measurability scores for each warning area. The WVS uses the categories shown in Figure 1 to score weather warnings.

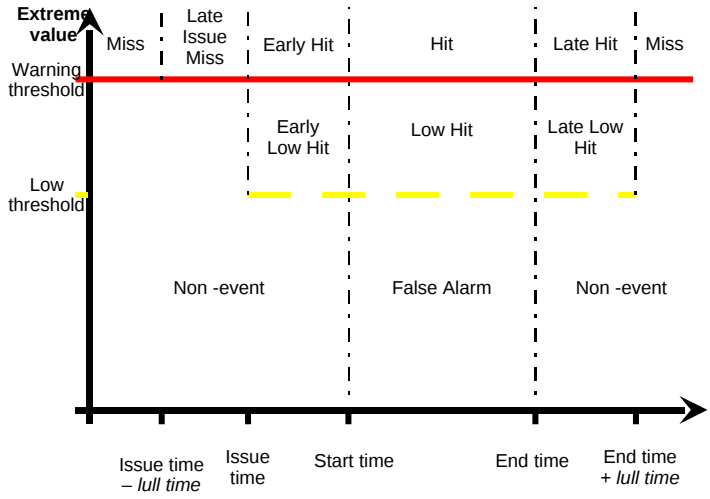

Figure 1: Event categories used by the Warnings Verification System

In this figure

- a *low hit* is when an event occurs (but just below the warning threshold),
- an *early hit* is when an event occurs between the issue and start time,
- a *late hit* is when an event occurs just after the warning period,

- a *late issue miss* is when an event occurs just before the issue time,
- the *lull time* is the average length of an event.

The remaining categories in Figure 1 (not shown in this list) are self explanatory. The settings of the WVS can be reconfigured to verify any warning product. It is possible that an event will not fall into just one of the categories shown in Figure 1. If a weather event starts within the warning period and continues into the *late hit* period, the WVS labels it as a *hit+late hit*. The introduction of compound events, such as this, avoids the double classification of events (which would otherwise have been classified as a *hit* and a separate *late hit*). A total of 20 meaningful compound events are possible using the event categories shown in Figure 1. Both temporal and intensity errors are addressed by expanding the simple *hit*, *miss* and *false alarm* definitions into those described by Figure 1.

Spatial discrepancies are addressed by looking at the severe weather that occurs close to, but just outside, each warning area. The largest warning area in the UK is the Highlands of Scotland (approx. 26,332Km^2) and the smallest is Slough (approx. 32Km^2). As the warning area size reduces, it becomes increasingly difficult to correctly issue severe weather warnings because a small positional error will turn a hit into a false alarm in Slough, but a much larger positional error is required to similarly transform a warning in the Highlands of Scotland. The user interface of the WVS shows an animation of each event. Figure 2 shows a single hour (17:00Z) from a WVS animation describing where the rainfall accumulation exceeded the heavy rain threshold during a National Severe Weather Warning that was issued for the period between 15:00Z and 20:00Z on 12/11/09 in the county of Windsor and Maidenhead (coloured dark blue in the figure). This warning was verified as a *false alarm*. The surrounding light blue area around Windsor and Maidenhead is a 25 mile extension of the county. The red points plotted on Figure 2 show the locations where the heavy rainfall warning threshold (of 15mm in 3 hours) was exceeded at 17:00Z on 12/11/09. The orange and yellow points are where the low threshold (of 13.5mm in 3 hours) was exceeded. It is clear from Figure 2 that the rainfall in Windsor and Maidenhead did not exceed the threshold, but just outside the county boundary the warning threshold was exceeded. Therefore if this county had been slightly larger the warning would have been verified as a *hit*. Consequently this warning is classified as a *nearby hit*. The number of *false alarms*, *missed events*, *non-events* and *hit type events* are currently recorded by the WVS at every 2 mile extension from 0 to 25 miles.

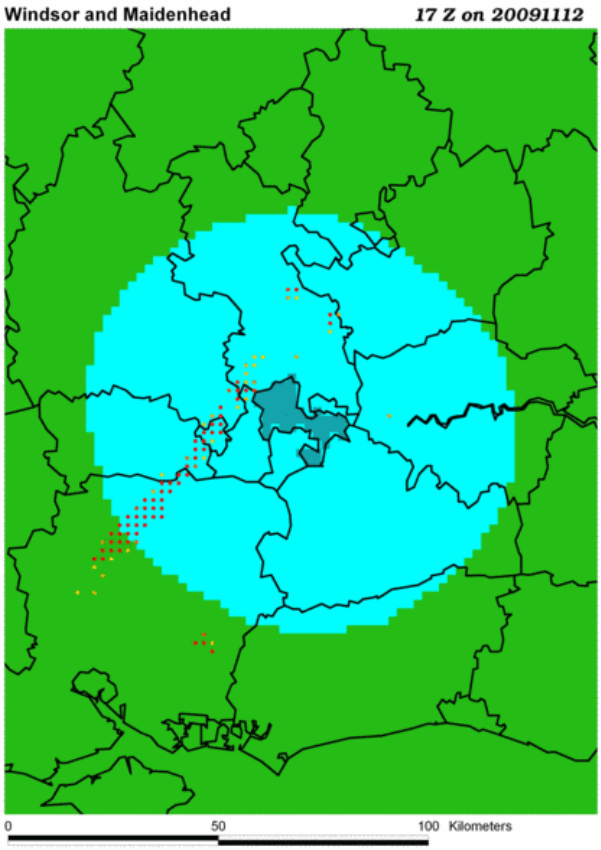

Figure 2: A screen shot from the WVS animation describing the heavy rain warning issued in Windsor and Maidenhead between 15:00 – 20:00 on 12/11/09

The final issue addressed by the WVS is confidence in the nowcast analysis data. If nowcast analyses were consistently 100% correct, just one grid point above the event threshold would be enough to be sure that an event had occurred. However because there is less confidence in the accuracy of nowcast analysis data it is prudent to look at how many observations exceed the warning threshold. The confidence that severe weather actually occurred increases with the number of grid points that exceed the warning threshold. Therefore it is appropriate to set a threshold on the number of grid points that exceed this threshold. However as warning area sizes vary significantly it is inappropriate to set this threshold to a fixed number of grid points. Instead a confidence threshold has been introduced based on a proportion of each warning area. The WVS currently uses confidence thresholds of 1%, 2%, 3%, 4% and 5%. When using a confidence threshold of 1%, at least 1% of the grid points in a warning area must be above the event threshold before a weather event occurs. So for a warning in the Highlands of Scotland (containing 6583 grid points) at least 66 grid points must be above the warning threshold for an event to have occurred. However it is by no means obvious which proportion of the warning area is the correct proportion and it is unlikely that this proportion

is fixed - it may legitimately vary from one event to the next. If the confidence threshold is set too high, events which should have been classified as *hits* are mistakenly classified as *false alarms*. If the confidence threshold is set too low events which are not severe enough to count as severe weather events are mistakenly classified as *missed events*. Therefore it is likely that the most appropriate overall result is the confidence threshold which gives the best overall result.

Scoring Methods used by the Warnings Verification System

It is important that any verification system gives at least one measure of performance. An overall measure should reflect the performance in each warning area. Three scores are used to measure the performance:

- The *Strict Score* is the score given when the scoring categories are restricted to *hit*, *miss* and *false alarm.*
- The *Flexed Score* is the score given when the strict definitions are flexed to define near spatial, temporal and intensity events as hits.
- A *Final Score* lies between the *Strict Score* and *Flexed Score* (discussed below)

The *Strict Score* and *Flexed Score* can be thought of as lower and upper bounds on the performance. Figure 3 shows the performance of the National Severe Heavy Rain Flash Warning Service, as measured by the ETS (Equitable Threat Score), using the *Strict* (red line), *Flexed* (green line) and *Final* (blue line) scoring methods. The *Final Score* shown in Figure 3 is calculated at the *optimal extension.* As a warning area is extended the score will improve. For small warning areas most of the improvement caused by extending the warning area contributes towards the optimal score. For large warning areas very little of the improvement caused by extending the warning area contributes towards the optimal score. Figure 4 shows what proportion of the improvement in the score (p_e) contributes towards the optimal value as a small, average and large warning each area is extended.

- Full credit ($p_e=1$) is given to all un-extended warning areas.
- Almost full credit ($p_e \approx 1$) is given to extended small warning areas until their extended radius reaches the average warning area radius.
- As the extended warning area radius varies between the average radius and double the average radius, the credit is reduced to zero.

Figure 4 is a hyperbolic tangent function.

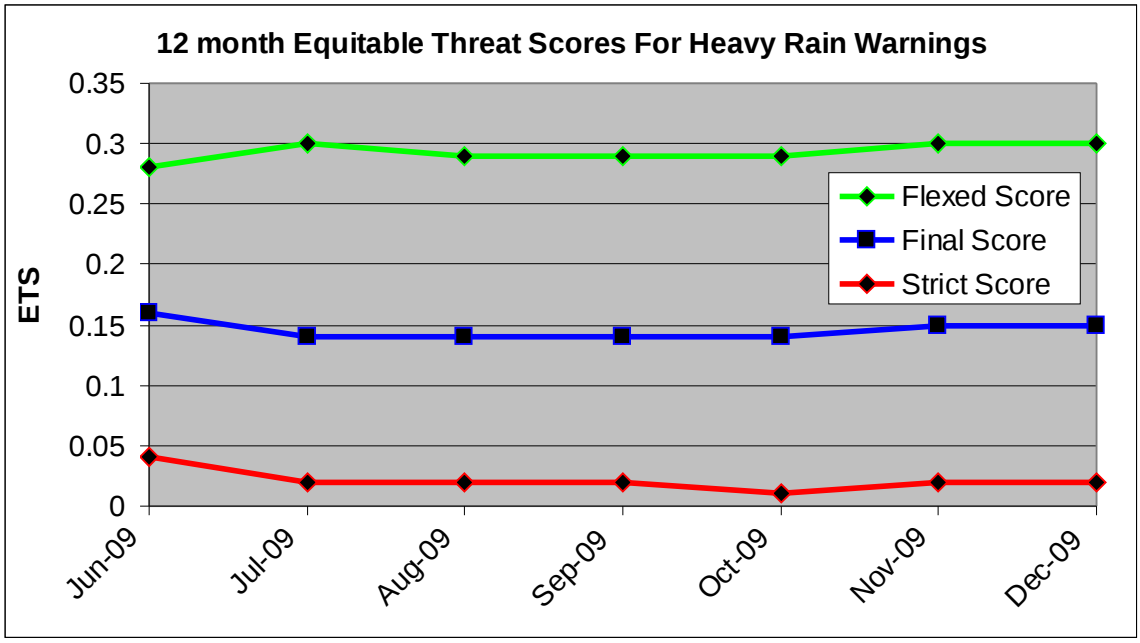

Figure 3: Rolling 12-month performance of Nsatioal Severe Heavy Rainfall Warnings in all UK areas

Equation (1) defines the number of hits at the *optimal extension*

$$h_E = MAX(h_0 + p_e(h_e - h_0), e = 1,2,\ldots,e_{\max}). \qquad (1)$$

The number of misses and false alarms at the optimal extension are calculated in a similar fashion. Figure 5 shows how the performance of the National Severe Weather Warning service increases as the warning areas are extensded. As the warning area is extended the ETS improves. The red and green lines show the improvement in the ETS as the *Strict* and *Flexed* scores are extended.

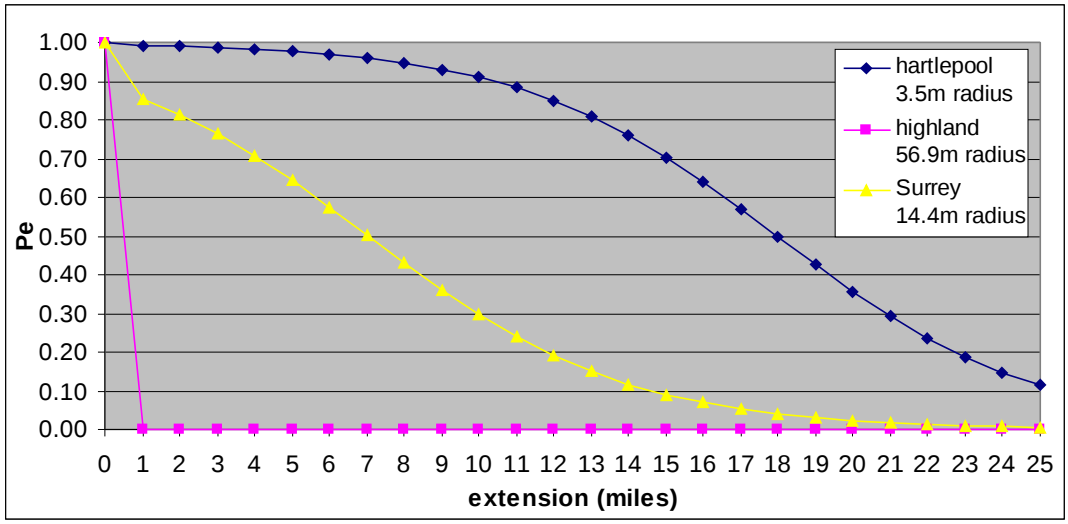

Figure 4: Proportion of the improvement in the verification score that is awarded as a warning area is extended

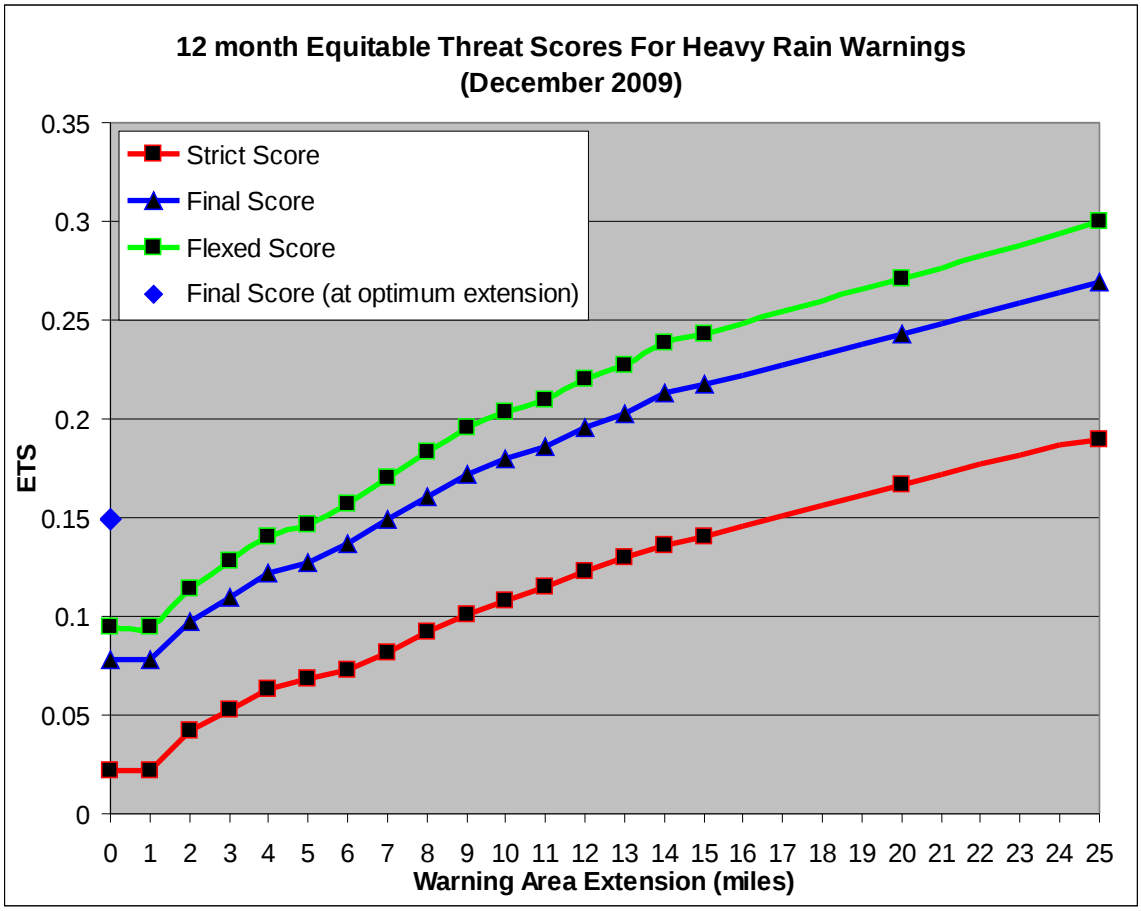

Figure 5: Improvement to the performance of the National Severe Heavy Rainfall Warning Service caused by extending the warning areas

The *Final Score* (blue line) is calculated from the *Strict Score*, a *Temporal Score* (when only temporal inaccuracies are treated as hits), an *Intensity Score* (when only intensity inaccuracies are treated as hits) and the *Flexed Score* (when both temporal and intensity inaccuracies are treated as hits) according to equation (2)

$$\begin{aligned} Final\ Score = MAX(\ & Strict + p_t(Temporal - Strict) + p_i(Best - Temporal), \\ & Strict + p_i(Intensity - Strict) + p_t(Best - Intensity), \\ & Strict + p_i \times p_t(Best - Strict)) \end{aligned} \qquad (2)$$

where p_i is the kept proportion of the improvement to the verification score caused by treating intensity inaccuracies as hits and p_t is the kept proportion of the improvement to the verification score caused by treating temporal inaccuracies as hits. (E.g., for the National Severe Heavy Rainfall Warning Service p_i=0.9, as the low threshold is 90% of the warning threshold, and p_t=0.55, as the warning period is typically 55% of the period between the *issue time* and the *end time* + *lull time*).

The ETS for the *Final Score* as a function of warning area extension is shown by the blue line in Figure 5. The blue diamond on the y-axis of Figure 5 shows ETS_E (obtained after the application of Figure 3). The blue line in Figure 3 shows ETS_E when the *Final Scoring* method is used.

Met Office
FitzRoy Road, Exeter
Devon, EX1 3PB
UK

Tel: 0870 900 0100
Fax: 0870 900 5050
enquiries@metoffice.gov.uk
www.metoffice.gov.uk