

# VERIFYING FORECASTS OF HIGH IMPACT WEATHER

Elizabeth E. Ebert\*

Centre for Australian Weather and Climate Research (CAWCR), Melbourne, Australia

Barbara G. Brown

National Center for Atmospheric Research (NCAR), Boulder, CO, USA

Marion Mittermaier

The Met Office, Exeter, UK

## 1. INTRODUCTION

Much of the benefit to society through improved weather forecasts will come from advances in our capability to forecast high impact weather and take mitigating action. High impact weather can come in many forms including rare and severe weather such as thunderstorms, tropical and extra-tropical cyclones, strong wind events, heavy rain events, fog, extreme heat and cold, floods and droughts, to name a few. Due to the relatively rare and often extreme nature of these events, meaningfully measuring the quality of high impact weather forecasts requires different approaches than are normally used for standard weather forecasts.

The goal of verification is to measure the quality (goodness) of the forecasts by comparing them to observations. This paper discusses several issues concerning the verification of high impact weather, and presents some strategies for evaluating high impact weather forecasts that give useful information for forecast users and forecast providers.

## 2. USER ISSUES

A good forecast should help the user to make a better decision than s/he could otherwise make in the absence of the forecast. What is meant by a "good" forecast depends to some degree on how the forecast is to be used. For example, emergency managers preparing to respond to victims of a tropical cyclone want good predictions of the location of landfall, but pinpointing the timing may be less critical. Event managers need to know the onset and duration of a heavy rain event but are less concerned about the amount of rainfall, whereas flood hydrologists need accurate

predictions of rainfall amount. Developers of weather forecast systems, including numerical modelers, want their products to simulate realistic weather, including distributions and extremes of temperature, wind, and precipitation. Operational forecasters understand the limitations of numerical model output and interpret it in the light of systematic errors learned over time.

Appropriate verification information can guide improvements in the forecasting systems, help forecasters interpret objective guidance, and inform end users regarding how to respond to warnings. Clearly no single verification approach can simultaneously meet the needs of all forecast users. Thus, verification approaches must be selected or developed to evaluate those aspects and attributes of the forecast that are important to the particular users in question. For the evaluation of the end-to-end forecast process the different components (guidance, warning, and so on) require different but complementary strategies for verification of the same outcomes.

In order for verification information to be useful, it must be communicated in an effective and timely manner. The evaluation methods must be clearly understood and relevant to the user. In the case of external users, this usually requires consultation between verifiers and users to establish meaningful quality standards and metrics. The verification information must be easily and quickly accessible, and ideally updated on a routine basis.

For rare events, the skill of a forecast system is often not well known. Providing relevant forecast verification information for similar events can increase the confidence both of forecasters in using objective guidance products in high impact situations, and external users in taking action based on the forecast.

---

\* *Corresponding author address:* Elizabeth E. Ebert, CAWCR, GPO Box 1289, Melbourne, VIC 3001, AUSTRALIA; e-mail: [e.ebert@bom.gov.au](mailto:e.ebert@bom.gov.au)

### 3. FORECAST ISSUES

There is often a mismatch between the events that forecasters need to issue warnings for, and what the numerical model guidance can provide. Some types of weather responsible for damage (lightning, gusts, fog) may not be explicitly predicted by the model, and must therefore be diagnosed from other variables. Even if they can be explicitly predicted (e.g., heavy rain), the model resolution may not capture the intensity of the experienced weather, and the processes associated with the variable are often sub grid scale. Some mesoscale models are being run experimentally at resolutions of 1-2 km, but most operational mesoscale models have grid scales of 10-20 km, and global models are coarser still. This means that the operational models cannot explicitly resolve the moist convection that leads to thunderstorm generation and associated severe weather.

When the quantities are fairly unpredictable, as in the examples just described, probabilistic predictions are more appropriate than deterministic ones. Nowadays, warnings are often issued as areas with medium or high risk (probabilities exceeding certain critical values) of experiencing a particular type of high impact weather. In the short range (1-3 days) and medium range (4-7 days) NWP ensembles are used to generate probability forecasts of the occurrence of extreme temperature, heavy rain, strong winds, and other high impact events. However, it is not possible to verify a probability forecast for a particular high impact event, however much one may be tempted to do so. Verification of probability forecasts requires many matched forecasts and observations. This may be difficult to achieve for high impact weather, which is often rare by definition. Verification using only a small dataset leads to results with large uncertainties.

A very useful ensemble-based forecast product for high impact weather is the Extreme Forecast Index (EFI; Lalaurette 2003), which is a spatially mapped index with values between -1 (when all ensemble members have lower values than the minimum of the model's climatological distribution) and +1 (all ensemble members greater than the model climatology's maximum value). The EFI indicates how extreme the forecast is, relative to the model's own distribution. Verification of the EFI must be done against the "extremeness" of the observed weather, i.e., the percentile of the

observation relative to its own climatology. This will be discussed further in Section 5.

### 4. OBSERVATION ISSUES

Ideally the forecasts and observations should correspond to the same quantity (e.g., accumulated rainfall) and have the same temporal and spatial scales. This is the case when comparing forecast and observed daily quantities at a site, or verifying NWP forecasts against analyses on the model grid. Otherwise it is necessary to match the forecasts and observations by sampling, interpolation, aggregation or averaging. These operations usually alter one or both of the values being compared, and can introduce an element of uncertainty to the verification results. Care must be taken to choose an appropriate matching method that minimizes the impact of the spatiotemporal remapping on the verification results (e.g., Accadia et al. 2003).

The quality of the observations affects the validity of the verification, with poorer quality observations introducing an artificial error component into the verification results. Observation-related errors including sampling error and measurement error, and both can be more pronounced for high impact weather. Extreme events, in addition to being rare, are often highly localized, which means that the observation network may not sample the event well in space and/or time. An event may have been forecast but in the absence of sufficient observations the good forecast performance cannot be confirmed. Quantitative observations may be difficult or impossible to obtain in developing nations where networks and infrastructure may be somewhat inadequate but the population is nevertheless vulnerable to devastating weather. In this case anecdotal evidence must be used to verify forecasts. Even when observation networks are well developed, strong winds and flooding associated with thunderstorms and tropical cyclones may disable instruments and communication networks.

Measurement errors such as rain gauge undercatch in strong winds, and attenuation of radar reflectivity in heavy rain, can seriously degrade the quality of high impact weather observations. While the effects of random error in the observations can be somewhat ameliorated by averaging in space or time, bias errors are often less easy to correct (frequently they are not well understood) and if left

uncorrected will be misinterpreted as bias error in the forecast.

Recognizing the many uncertainties associated with observational data, some investigators are now exploring the concept of "probabilistic observations" (e.g., Bellerby and Sun, 2005). Observations-based PDFs can be used to drive downstream applications such as hydrologic models; their use in forecast verification is a new area of research (e.g., Roberts and Lean 2008).

### 5. SIMPLE VERIFICATION APPROACHES

As discussed in Section 2, different users require different kinds of verification information. External users of high impact forecasts mainly need advice how much trust they can place in the forecast – how much uncertainty is normally associated with the forecast in similar situations (non-systematic error), and what biases it is likely to have. Simple metrics like mean error and mean absolute error are often quite effective for conveying this type of information.

High impact events may be defined categorically, often as the occurrence of a variable exceeding some dangerous threshold (e.g., rainfall exceeding 50 mm in 6 h). Categorical verification results displayed in a 2x2 contingency table (shown below)

		Observed events		
		yes	no	
Forecast events	yes	<i>hits</i>	<i>false alarms</i>	<i>forecast yes</i>
	no	<i>misses</i>	<i>correct rejections</i>	<i>forecast no</i>
		<i>observed yes</i>	<i>observed no</i>	<i>total number of fcsts</i>

address questions that are directly relevant for many types of decisions and responses to the warnings, such as:

- (a) when a warning is given, what is the chance that the event will actually occur?
- (b) when an event occurs, what is the chance of a warning being issued?

The contingency table elements indicate the ability of a warning system to discriminate between events and non-events. Multi-category contingency tables show greater detail on forecast performance while still being easy to understand.

A multitude of scores derived from the contingency table summarize particular aspects of forecast performance. Some simple ones, namely the frequency bias (FBI – ratio of forecast events to observed events), probability of detection (POD – fraction of observed events correctly predicted), false alarm ratio (FAR – fraction of predicted events that were false alarms), and the threat score (TS – fraction of observed and/or predicted events that were hits), are commonly used in high impact weather verification. (See Jolliffe and Stephenson (2003) for definitions and more information on categorical scores.) Hewson (2007) proposed a conceptually attractive "deterministic limit", defined as the time into the forecast (hours) when the number of hits falls below the total number of misses and false alarms.

Most of the frequently used summary scores that take into account errors due both to misses and false alarms (threat score, Heidke skill score, Peirce's skill score, Gilbert skill score) have the undesirable property that for skilled forecasts they converge to zero as the event becomes rarer (Stephenson et al. 2008), thus ceasing to provide useful information. The proportion correct (PC) converges to 1 for increasingly rare events, reflecting perfect forecasts of non-events. Since this is unhelpful information and can easily be misleading, PC should not be used to verify rare events.

Stephenson et al. (2008) recommend verifying binary forecasts of rare events using the Extreme Dependency Score (EDS):

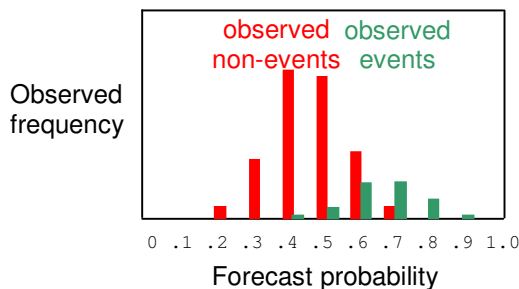
$$EDS = \frac{2 \log\left(\frac{hits + misses}{total}\right)}{\log\left(\frac{hits}{total}\right)} - 1$$

which measures the association between forecast and observed rare events, and converges to a value between 0 and 1 for skilled forecasts. EDS is independent of the forecast bias, so should be presented together with FBI. Other scores that are "friendly" to extreme forecasts (don't converge to 0) include the odds ratio skill score (ODSS), the

linear error in probability space (LEPS), and quantile-based categorical statistics (Jenkner et al. 2008). These scores take account of the climatological frequency of the event.

For probabilistic verification, reliability diagrams and relative operating characteristic (ROC), provide a wealth of information on forecast bias and discrimination. For verifying rare events and higher probabilities, statistical fitting methods can be used to better estimate the reliability and reduce the uncertainty associated with small samples (Bradley et al. 2003; Atger 2004).

If probabilistic forecasts are difficult for many users to understand, the verification of probabilistic forecasts is even harder to comprehend. A simple and meaningful approach is to tabulate the number of observed events and non-events associated with each probability forecast, or present them in a discrimination diagram as shown in Fig. 1 (S. Mason, personal communication).



**Fig. 1. Sample discrimination diagram showing the performance of probability forecasts.**

For any quantitative evaluation it is important to ensure that the verification dataset contains samples from same regime, to avoid false skill (Hamill and Juras 2006). This may be difficult in the case of rare events, where it is tempting to combine data from different regimes in order to create a larger sample size.

Finally, not all verification needs to be quantitative. Anecdotal / qualitative / survey assessment of forecast usefulness is a valid way to learn about forecast quality. This was conducted during two recent World Weather Research Program (WWRP) forecast demonstration projects, namely the Severe Weather Forecast Demonstration Project (SWFDP) in southern Africa and the MAP D-PHASE project in central Europe (Coiffier and

Chen 2008; Ambrosetti et al. 2007), where timely feedback from forecasters provided not only extremely useful assessments of forecast value but descriptions of forecast deficiencies.

## 6. DIAGNOSTIC VERIFICATION APPROACHES

Several new spatial verification methods have become available in recent years. These methods were developed largely to evaluate the capabilities of high resolution mesoscale models, but are applicable to any matched forecasts and observations on a grid. They account for the spatial coherence in weather events and evaluate collections of related gridpoints rather than single gridpoints. A few spatial verification methods that are particularly informative for high impact weather will be briefly discussed here; Gilleland et al. (2009) give a more complete review of spatial methods.

Features-based verification methods such as the Contiguous Rain Area method (CRA; Ebert and McBride 2000), the Method for Object Based Diagnostic Evaluation (MODE; Davis et al. 2006, 2009), and the Composite Method (Nachamkin et al. 2005) evaluate the forecast location, size, and other properties of identifiable features such as heavy rain areas, and are therefore quite intuitive for the user. These approaches can highlight systematic errors in forecasting systems, provided the events being forecast can be reasonably represented as objects. Features-based methods are starting to be applied to estimate timing errors as well as spatial errors.

Neighborhood (sometimes called "fuzzy") verification methods give information on the spatial and temporal scales at which forecasts have certain levels of accuracy and skill, helping decision makers to use the forecasts appropriately (Ebert 2008; Mittermaier and Roberts 2009). The neighborhood approach is probabilistic in nature, reflecting the way forecasters actually interpret high resolution guidance. Rather than traditional point-to-point matching of forecasts and observations, which can impose quite severe penalties for small-scale errors in high resolution forecasts, neighborhoods of forecasts are matched to the observations. Some methods such as the Fraction Skill Score (Roberts and Lean 2008) evaluate forecast neighborhoods against neighborhoods of observations, implicitly accounting for observational uncertainty as well.

Object and neighborhood verification approaches have been used in the MAP D-PHASE project in Europe and the WWRP Beijing 2008 Olympics Forecast Demonstration Project, and have been shown to provide useful information on high impact weather forecast quality (Ament et al. 2008). So far the users of this information have been mainly researchers. This is probably because these spatial verification techniques are not yet considered "mainstream", and are methodologies may still be in the process of being optimized. The effective communication of spatial verification information to operational forecasters and external forecast users still requires improvement. While the location and intensity errors and diagnosis of useful scales are clearly useful information for users, the methods for getting them can be somewhat complicated (particularly for the object-based techniques). For users who are not comfortable with a "black box" approach to forecast evaluation, the added complexity may make these techniques less attractive than the more traditional approaches.

## 7. FINAL REMARKS

This paper gave a brief overview of many of the issues that must be considered when verifying forecasts of high impact weather, and presented some methods that can be used to evaluate various aspects of forecast performance.

Since much high impact weather is relatively rare and often extreme, it is a challenge for models and human forecasters to predict correctly. Getting it wrong may be understandable, and to some extent also inevitable given the inherently unpredictable nature of many severe events such as thunderstorms. At the same time, errors in forecasts and warnings of high impact weather (particularly when the errors involve missing an event or significantly underestimating its severity) can lead to wrong responses on the part of emergency managers and other decision makers, and subsequent losses of property and even life. For this reason it is important that users of high impact weather forecasts and warnings are equipped with appropriate information on their limitations. Verification should be conducted as often and as carefully as possible in order to understand and quantify the errors in forecasts and warnings, and the results should be conveyed to users in a user-friendly and timely manner.

## 8. REFERENCES

- Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918-932.
- Ambrosetti, P., U. Germann, A. Hering, L. Fontannaz, M. Stoll, 2007: MAP D-PHASE severe convection forecasts. *4th European Conf. Severe Storms, 10-14 September 2007, Trieste, Italy*.
- Ament, F., M. Arpagaus, and M. W. Rotach, 2008: Quantitative precipitation forecasts in the Alps – first results from the Forecast Demonstration Project MAP D-PHASE. *Geophys. Res. Abstr.*, **10**, EGU2008-A-08259.
- Atger, F., 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. *Quart. J. Royal Meteorol. Soc.*, **130**, 627-646.
- Bellerby, T.J. and J. Sun, 2005: Probabilistic and ensemble representations of the uncertainty in an IR/microwave satellite precipitation product. *J. Hydrometeor.*, **6**, 1032-1044.
- Bradley, A.A., T. Hashino, and S.S. Schwartz, 2003: Distributions-oriented verification of probability forecasts for small data samples. *Wea. Forecasting*, **18**, 903-917.
- Coiffier, J. and P. Chen 2008: Severe Weather Forecasting Demonstration Project – Regional Subproject in RA I – Southeast Africa. Final Report. WMO Secretariat, 27 February 2008.
- Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methods and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772-1784.
- Davis, C.A., B.G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) applied to WRF forecasts from the 2005 SPC Spring Program. *Wea. Forecasting*, submitted.
- Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appl.*, **15**, 51-64.

Ebert, E.E. and J.L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrology*, **239**, 179-202.

Gilleland, E., D. Ahijevych, B.G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, submitted.

Hamill, T.M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteorol. Soc.*, **132**, 2905-2923.

Hewson, T., 2007: The concept of 'Deterministic limit'. *3rd Intl. Verification Methods Workshop, 31 January-2 February 2007, Reading, UK*.

Jenkner, J., C. Frei and C. Schwiertz, 2008: Quantile-based short-range QPF evaluation over Switzerland. *Meteorol. Zeitschr.*, **17**, 827-848.

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

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J. Royal Met. Soc.*, **129**, 3037-3057.

Mittermaier, M. and N. Roberts, 2009: Inter-comparison of Spatial Forecast Verification Methods: Identifying skillful spatial scales using the Fractions Skill Score. *Wea. Forecasting*, submitted.

Nachamkin, J. E., S. Chen, and J. Schmidt, 2005: Evaluation of heavy precipitation forecasts using composite-based methods: A distributions-oriented approach. *Mon. Wea. Rev.*, **133**, 2163-2177.

Roberts, N.M. and H.W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78-97.

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