18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

A framework for neighbourhood verification of high resolution spatial forecasts

Ebert, E.E.

Centre for Australian Weather and Climate Research (CAWCR), Bureau of Meteorology, Melbourne, Vic.
Email: e.ebert@bom.gov.au

Abstract: High resolution spatial forecasts from numerical models can look quite realistic and provide the user with very valuable guidance. However, when verified using traditional metrics such as root mean square error, threat score, etc., they often score quite poorly because of the difficulty of predicting an exact match to the observations at high resolution. Recent years have seen the development of neighbourhood (also known as "fuzzy") verification approaches that reward closeness by relaxing the requirement for exact matches between forecasts and observations. Some of these neighbourhood methods compute standard verification metrics for deterministic forecasts using a broader definition of what constitutes a hit. Other neighbourhood methods treat the forecasts and/or observations as probability distributions and use verification metrics suitable for probability forecasts. Each method is distinguished by a unique decision model for defining a "good" forecast. All methods can be implemented using a common moving-window framework in which the forecasts within a certain space-time window are compared to observation(s) either in the same window or in the center of the window. By varying the threshold intensity used to define an event and the size of the space-time window, the dependence of forecast accuracy on intensity and scale can be evaluated. The strength of the methodology is in identifying those spatial and temporal scales for which the forecasts have sufficient accuracy for a particular application.

This paper briefly describes the neighbourhood verification approach and demonstrates it on high resolution rainfall forecasts from the Bureau of Meteorology's ACCESS weather prediction model.

Keywords: *Spatial verification, forecast verification*

1. INTRODUCTION

Precipitation forecasts from high resolution numerical models tend to look more realistic to the human eye than smoother forecasts made at lower resolution, with better depiction of spatial variability and intensity range. However, when verified at the scale of the model grid, higher resolution forecasts are at a disadvantage because it is more difficult to match the observations at fine scale. Small errors in location can lead to poor values of traditional verification metrics such as root mean square error (RMSE), correlation coefficient, and equitable threat score (ETS)[1].

Weather forecasters appreciate the realism of high resolution numerical forecasts. But rather than interpret the numerical guidance literally, forecasters compensate for small scale errors in location and timing by interpreting weather features in a slightly looser way, as "located in this area" and "at about this time". From experience they may know that a certain model typically predicts the location of rain systems correctly to within about 40-70 km, and the timing to within 1-2 hours. Their conceptual interpretation of the model output tends to be probabilistic rather than strictly deterministic. The use of forecasts in this way calls for a method of evaluation that allows for a level of uncertainty in the forecast, and given that uncertainty, measures the closeness of the forecast to the observations.

A number of recently developed spatial verification methods can be classed as "neighbourhood" verification methods (sometimes known as "fuzzy" verification) because they verify forecast values within a space-time neighbourhood surrounding the observation, instead of requiring a point-to-point match. Neighbourhood methods have been developed by Brooks et al. (1998), Zepeda-Arce et al. (2000), Atger (2001), Damrath (2004), Germann and Zawadzki (2004), Yates et al. (2006), Theis et al. (2005), Marsigli et al. (2005), Rezacova et al. (2007), Segawa and Honda (2007), and Roberts and Lean (2008). Each method is distinguished by a unique decision model for defining a "good" forecast. For example, Atger's (2001) method asserts that a useful forecast is one that predicts an event close to an observed event. Ebert (2008) showed that the neighbourhood verification methods could all be implemented within a common framework that computes a variety of scores (corresponding to the various methods) within a spatial window that is sequentially moved over each grid box or observation. The size of the window is systematically increased from grid scale to some large number of grid squares, and the intensity threshold used to compute probabilistic and categorical verification scores is varied from low to high values. This process allows the user to identify at which spatial and temporal scales the forecasts have sufficient accuracy for a particular application.

This paper briefly describes and demonstrates four neighborhood verification methods that evaluate different aspects of forecast "goodness". They are used to verify high resolution quantitative precipitation forecasts (QPFs) from the Australian Community Climate and Earth-System Simulator (ACCESS), the new numerical weather and climate prediction system recently implemented at the Bureau of Meteorology and CSIRO (Puri 2005).

2. NEIGHBOURHOOD VERIFICATION METHODS

Neighbourhood verification computes error metrics for the set of all neighbourhoods, or space/time windows, in a domain, rather than the set of all individual grid boxes. The size of the local spatial neighbourhood around a grid box is increased monotonically from 1x1, 3x3, …, to (2^n+1)x(2^n+1), approximating the size of the domain. If a temporal domain is used then t time windows are increased in the same way. Since many neighbourhood verification methods use the concept of an "event", i.e., the occurrence of a value greater than or equal to some threshold value, the intensity threshold for an event is also varied from small to large values, $T_1,\ldots, T_m$. Thus, instead of the single score that is normally reported for grid box-scale validation using a single event threshold, neighbourhood verification provides an m x n x t array of scores for varying scales and thresholds. It is then possible to examine the array of scores to determine which space and time scales have useful skill for forecasts exceeding various intensities.

A well known verification approach that can be considered "neighbourhood" is *upscaling*, in which the estimates and observations at grid scale are averaged to larger scales before being compared using the usual continuous and categorical verification metrics (e.g., Yates et al. 2006). The upscaling method considers a useful forecast to be one that has the same average value as the observations. In the context of QPFs this is one criterion for judging whether precipitation estimates at the catchment scale are useful for hydrological purposes.

[1] See JWGV (2009) for definitions of traditional metrics used to verify forecasts.

The *fractions skill score* (FSS) method of Roberts and Lean (2008) considers a perfect estimate to be one with the same frequency of events as was observed within a neighbourhood. This neighbourhood method implicitly acknowledges that the observations are likely to contain random error at the grid scale, and asserts that a better approach to comparing high resolution forecasts with observations is to assess their similarity in terms of their fractional coverage of events (e.g., grid boxes with rain). The fractions skill score is a variation of the Brier skill score used to verify probability forecasts:

$$FSS = 1 - \frac{\frac{1}{N}\sum_N (P_{fcst} - P_{obs})^2}{\frac{1}{N}\left[\sum_N P_{fcst}{}^2 + \sum_N P_{obs}{}^2\right]} \qquad (1)$$

P_{fcst} and P_{obs} are the fractional coverages of forecast and observed grid box events, respectively, in each of the N neighbourhoods in the domain. The FSS varies between 0 for a complete mismatch and 1 for a perfect match. The target value of FSS above which the estimates are considered to have useful skill (better than the skill of a uniform forecast of f_{obs}, where f_{obs} is the fraction of observed grid box events in the full domain) is given by FSS_{useful}=0.5+f_{obs}/2. This leads to the concept of a "skillful scale", namely the smallest scale at which the FSS exceeds FSS_{useful}. This is a more meaningful concept for many users, and is now used at the Met Office to help forecasters understand the quality of high resolution model precipitation forecasts (Mittermaier and Roberts 2009).

The *multi-event contingency table* (MECT) method of Atger (2001) considers an estimate to be useful if at least one occurrence of an event is predicted close to an observed event. "Close" can refer to space, time, intensity, or any other important aspect. Closeness is an important criterion for weather forecasters using model output to prepare warnings of heavy rain and strong winds, and for emergency managers and disaster relief agencies planning emergency response activities. The MECT method compares a neighbourhood of forecasts to an observation in the center using traditional categorical metrics such as frequency bias, probability of detection, false alarm ratio, and so on. Whenever an event is observed in the central grid box of the neighbourhood and also predicted by the model in at least one grid box in the neighbourhood, this counts as a hit. If there is no event observed in the central grid box but one or more neighbourhood grid boxes with forecast events, a false alarm is counted. As the neighbourhood increases in size, it is easier to get a hit, but also easier to get a false alarm. Although any categorical verification score can be computed, the one most relevant to accuracy assessment in this case is the Hanssen and Kuipers discriminant HK, which measures the difference between the probability of detection (rewarding hits) and the probability of false detection (penalizing false alarms).

Germann and Zawadzki (2004) proposed a neighbourhood verification method that uses as its criterion for goodness, "A forecast is useful if it has a high probability of matching the observed value." Called the *conditional square root of RPS* (CSRR), it uses a probabilistic approach to compare forecasts to the observed rain in the center of each neighbourhood. The square root of the ranked probability score (RPS) can be interpreted as the standard error of the probability estimates across the full range of observed and forecast intensities. Normalizing by the observed event fraction in the domain enables performance to be compared for different cases:

$$CSRR = \frac{\sqrt{RPS}}{f_{obs}} \qquad (2)$$

Likely users of CSRR information would include emergency managers and other decision makers concerned about the effects of local high impact events.

The first two neighbourhood methods described in this section, upscaling and FSS, compare neighbourhoods of forecasts against neighbourhoods of observations. Ebert (2008) calls this strategy "model oriented", meaning that the observations are manipulated to represent the scales resolved by the model. This gives a fair assessment of the forecasts in the sense that they are being evaluated only on scales that they claim to resolve. The last two neighbourhood methods, MECT and CSRR, compare each a neighbourhood of forecasts against the single observation in the center of the neighbourhood. Although this may seem "unfair", many users wish to know the accuracy of the forecast at a particular location. Note that this "user oriented" philosophy is more demanding than the model oriented one, but not as tough as the traditional grid box-to-grid box verification, since skill can still be demonstrated when the forecast detects events close to the observation.

3. VERIFICATION OF MODEL PRECIPITATION FORECASTS AGAINST BRISBANE RADAR

In mid-November 2008 a series of severe thunderstorms brought widespread rain, hail, and strong winds to the Brisbane area. The left panel of Figure 1 shows hourly rainfall accumulation measured using a combination of radar and gauge data and analysed using the Bureau's operational Rainfields system (Seed and Duthie 2007). The right panel shows the 7h forecast valid for the hour ending at 0700 UTC on 16 November, predicted using the ACCESS model with 0.05° grid spacing (~5 km). Both maps show a rain system near the coast at around 28°S, with the forecast field displaced slightly to the east of the observations. To prepare the data for verification, the original 2 km resolution radar grid boxes were spatially averaged to the 0.05° scale of the model. In spite of this averaging, the radar field has greater spatial variability than the model field.

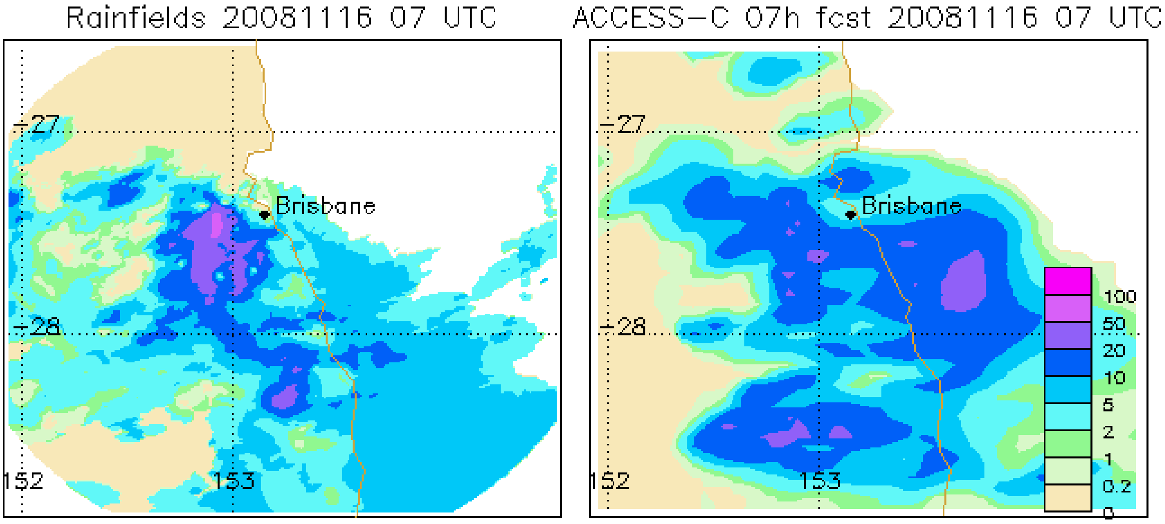

Figure 1. Hourly rainfall estimated from a radar-gauge analysis (left) and the 7h forecast from the ACCESS 0.05° model (right), valid at 0700 UTC on 16 November 2008. The colors indicate rain rate in mm h^{-1}.

Traditional verification of this rainfall forecast against the Rainfields data yields an RMS error of 6.0 mm, a spatial correlation coefficient of 0.45, and equitable threat scores of 0.27 for rain exceeding 0.2 mm h^{-1} and 0.15 for rain exceeding 10 mm h^{-1}. Since the perfect value of ETS is 1.0, the low values achieved for this forecast are rather discouraging, especially considering that a weather forecaster in Brisbane would have found this model rainfall guidance exceedingly useful. In fact, this example was chosen because it was subjectively judged to be a good forecast.

Neighborhood verification scores were computed for eight window sizes ranging from 1x1 to 33x33 (the largest that the radar analysis could accommodate) and eight intensity thresholds ranging from 0.2 to 50 mm h^{-1}. No time window was used. The results are shown in Figure 2 as a function of the rain intensity threshold (x-axis) and spatial scale (y-axis). In these plots the shading and the number show the value of the score, with red and pink values indicating good performance and yellow and blue values denoting poor performance. The value in the lower left corner is the score that would be achieved using traditional grid box matching and a very low (essentially rain/no rain) threshold.

For the upscaling approach the equitable threat score was chosen as the error metric since it penalizes both misses and false alarms, and therefore rewards correct placement of forecast features. The 0.05° ACCESS model showed optimal skill at a spatial scale of 45 km and an intensity threshold of 2 mm h^{-1}. This indicates that a broader than grid scale interpretation is beneficial, and that the moderate rain was located more accurately than rain of other intensities (Fig. 2a).

A different picture emerges when comparing the estimated and observed fractional coverage of rain grid boxes using the fractions skill score. The FSS improves with increasing spatial scale and decreasing rain intensity, and tends to be characterized by higher numerical values than the ETS. The FSS puts greater emphasis on the rain intensity *distribution* than does the upscaling approach, and is more sensitive to bias. The forecast was useful according to the target criterion (FSS > $0.5+f_{obs}/2$) at all scales when light rain thresholds were considered, and at scales of 45 km and above for intensities exceeding 10 mm h^{-1} (Fig. 2b).

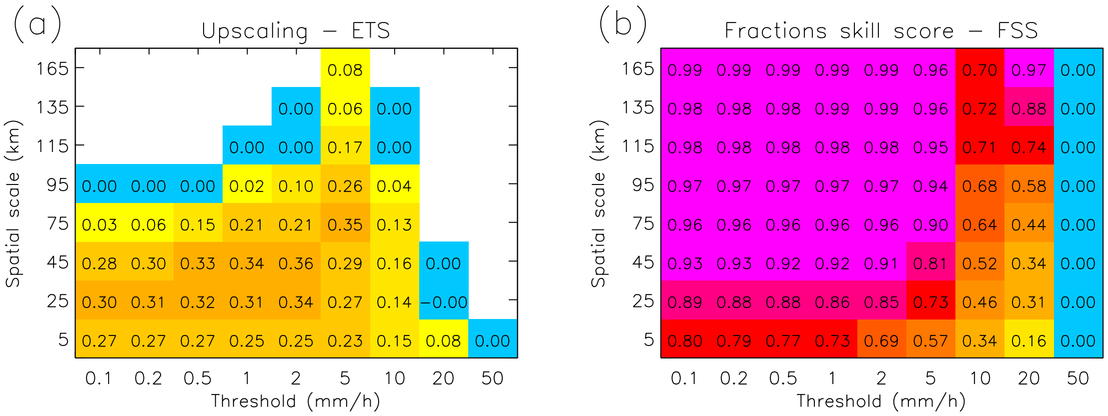

Figure 2. Neighbourhood verification scores using the (a) upscaling and (b) Fractions Skill Score methods for 7h ACCESS forecast of 1h accumulation valid at 0700 UTC on 16 November 2008, verified against hourly Rainfields radar-gauge data near Brisbane.

4. VERIFICATION AGAINST AUSTRALIAN DAILY GAUGE DATA

In a different experiment, forecasts from a slightly coarser resolution (0.11° grid) version of the ACCESS model were verified directly against gauge observations over Australia to demonstrate the MECT and CSRR verification methods. Both methods compare neighbourhoods of forecasts with observations in the centers of the neighbourhoods. Although there is a mismatch in spatial scale between the model and observed precipitation values, it should be noted that, in practice, forecasts made at this resolution are often used to predict values at point locations. 24h forecasts of daily rainfall on 42 days between 1 September 2008 and 1 March 2009 were verified using the MECT and CSRR methods. (QPFs were not available on all days because the model was still being tested during this period.) Figure 3 shows results that were aggregated over all days in the sample.

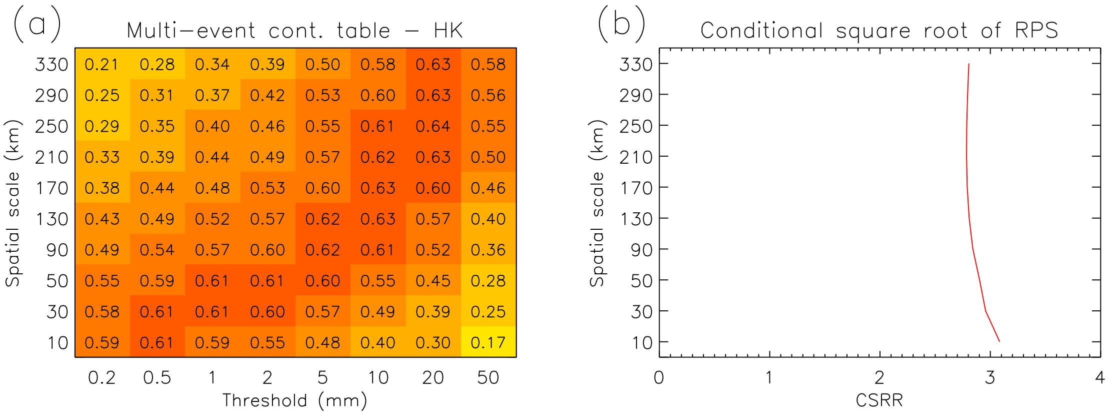

Figure 3. Neighbourhood verification scores computed using the (a) MECT and (b) CSRR methods for 24h ACCESS model forecasts of daily accumulation at 0.11° grid scale, verified against daily rain gauge observations over Australia.

The MECT method measures the ability of the model to predict rain of a given intensity close to where it was observed (Fig. 3a). A good way to use the results from this method is to choose a threshold of interest, say 5 mm, and then scan vertically to see at which scales the best scores were achieved. In a forecasting context this would tell the user how large the neighbourhood surrounding the point of interest should be to provide a useful indication of rain, with some hits and not too many false alarms. The maximum values of HK were found at small scales for light rain thresholds, increasing to larger scales as the rain threshold is increased. Put another way, the optimum "search radius" increases with increasing intensity threshold, which is not surprising since higher intensities are often associated with smaller scale features that are more difficult to pinpoint exactly. For 24h ACCESS model forecasts of moderate rain exceeding 5 mm h^{-1} a neighbourhood of 100 km (effective radius of about 50 km) would have given useful guidance during this period.

The final method used to evaluate the ACCESS QPFs is the probabilistic CSRR method, which has a low (good) value when the estimated intensity distribution within a neighbourhood peaks near the observed value. The focus of the CSRR on the distribution is similar to that of the FSS, but neighbourhood estimates are compared to point observations rather than observation neighbourhoods. Unlike the scores previously shown, the CSRR assesses the full intensity distribution rather than a binary (yes-no) threshold exceedance; in this case the eight intensity thresholds in Fig. 3a were used to define the rainfall intervals over which the score was computed. Fig. 3b shows that the CSRR for the ACCESS model improved slowly with increasing scale up to about 200 km, which was the optimal scale for representing the 24h forecast intensity distributions of daily rainfall. Comparing the CSRR and MECT results, the forecast scale at which the intensity distributions were most accurate was about twice the scale required for finding one or more moderate intensity events close to observed events.

4. DISCUSSION

As the production and distribution of high resolution model-based precipitation products becomes increasingly common, the need to evaluate them appropriately becomes more important. Standard grid box-by-grid box verification can suggest that high resolution forecasts are not as accurate as lower resolution forecasts, yet most users intuitively feel that the high resolution products should be more useful. Neighbourhood verification gives credit to estimates that are "close" to the observations, thus offering an alternative to traditional verification approaches. This is achieved by looking in space/time neighbourhoods surrounding the observations and evaluating the degree of "closeness" according to various criteria. By evaluating the accuracy as a function of both intensity and spatial scale, neighbourhood verification gives information about which scales have useful skill. This helps users to decide whether to use the estimates at face value at full resolution, or spatially transform the values to give more accurate and useful information.

To the majority of users of high resolution forecasts who are not very familiar with objective verification techniques and scores, neighbourhood verification may seem somewhat daunting. Two new metrics, namely the FSS and the CSRR, have only recently been introduced into the meteorological literature and are not yet found in standard textbooks on verification. Even those who are comfortable with verification methods and scores may find it overwhelming to interpret the results from several neighbourhood methods, each of which produces a large array of scores. The key is to first identify which is the most important aspect of the forecast to get right – is it the spatial average, the fractional area, the presence of one or more high intensity events nearby the location of interest, the rain rate distribution, etc.? Then choose the neighbourhood verification method that addresses this aspect. Focusing on an intensity threshold of interest and condensing its scale-dependent performance into a single easily-interpreted value like the "skillful scale" can help make the verification results much more accessible. As neighbourhood verification becomes more widely used, new approaches will certainly emerge for interpreting the results in ways that intuitively meet the needs of specific users.

REFERENCES

Atger, F. (2001), Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlin. Proc. Geophys.*, 8, 401-417.

Brooks, H.E., Kay, M. & Hart, J.A. (1998), Objective limits on forecasting skill of rare events. *19th Conf. Severe Local Storms, Amer. Met. Soc., Minneapolis, MN, 14-18 September 1998, AMS*. 552-555.

Damrath, U. (2004), Verification against precipitation observations of a high density network – what did we learn? *Intl. Verification Methods Workshop, 15-17 Sept. 2004, Montreal, Canada.* [Available online at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Workshop2004/presentations/5.3_Damrath.pdf]

Ebert, E.E. (2008), Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appl.*, 15, 51-64.

Ebert, E.E. (2009), Neighborhood verification – a strategy for rewarding close forecasts. *Wea. Forecasting*, accepted.

Germann, U. & Zawadzki, I. (2004), Scale dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *J. Appl. Meteorol.* 43, 74-89.

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

JWGV (Joint Working Group on Verification) (2009), Forecast verification: Issues, methods, and FAQ. [Avaliable at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html.]

Marsigli, C., Boccanera, F., Montani, A. & Paccagnella, T. (2005), The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. *Nonlin. Proc. Geophys.* 12, 527–536.

Mittermaier, M. and N. Roberts (2009), Inter-comparison of spatial forecast verification methods: Identifying skillful spatial scales using the Fractions Skill Score. *Meteorol. Appl.*, accepted.

Puri, K. (2005), Project plan for ACCESS. Bureau of Meteorology, September 2005.

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.

Seed A., and E. Duthie (2007), Rainfields: A quantitative radar rainfall estimation scheme. *33rd Conf. Radar Meteorology, Amer. Met. Soc., Cairns, Australia, 6-10 August 2007.*

Segawa, T. & Honda, Y. (2007), The verification of high-resolution precipitation forecasts of the operational JMA mesoscale model. *3rd Int'l Verification Methods Workshop, Reading, UK, 31 January-2 February 2007.*

Theis, S.E., Hense, A. & Damrath, U. (2005), Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorol. Appl.* 12, 257-268.

Yates, E., S. Anquetin, V. Ducrocq, J.-D. Creutin, D. Ricard and K. Chancibault (2006), Point and areal validation of forecast precipitation fields. *Meteorol. Appl.*, 13, 1-20.

Zepeda-Arce, J., Foufoula-Georgiou, E. & Droegemeier, K.K. (2000), Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.*, 105, 10,129-10,146.