A comparison of two radar rainfall ensemble generators

Katie Norman1, Alan Seed2, Clive Pierce3

1Met Office, FitzRoy Road, Exeter, Devon, EX1 3PB, UK, katie.norman@metoffice.gov.uk
2Centre for Australian Weather and Climate Research, Bureau of Meteorology, GPO Box 1289K, Melbourne, Victoria, 3001, Australia, A.Seed@bom.gov.au
3Joint Centre for Hydro-meteorological Research, Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB, UK, clive.pierce@metoffice.gov.uk

1. Background

As radar becomes more influential to those responsible for water management and short range precipitation nowcasting, an estimate of the space-time uncertainty associated with the radar precipitation estimates is needed. The Met Office radar precipitation rate product is generated every 5 minutes at 1, 2 and 5 km resolution, as described by Harrison et al (2000 and 2009). The 2km product is used by The Met Office Short Term Ensemble Prediction System (STEPS) (Bowler et al., 2006) to generate precipitation nowcasts out to T+6 hours. The current uncertainty product that accompanies radar data gives an estimate of data quality (a function of beam height); however this does not give a quantitative estimate of the error.

Recently, ensemble generation has been suggested as a method of quantifying uncertainty. These models fall into two categories: modelling individual sources of error, (Lee et al., (2007), Lee and Zawadzki (2005a, 2005b, 2006), Jordan et al (2003), Berenguer and Zawadzki (2008)) and statistical descriptions of the total error in the radar data (Germann et al (2009), Ciach et al (2007), Llort et al (2008)). The difficulty with the first approach is that the error structure is complex and inter-dependent. The difficulty with the second approach is the requirement to have a reference field with adequate resolution, which is usually based upon a dense network of rain gauges.

STEPS already incorporates an observation uncertainty algorithm based upon an analysis of the error in the assumed Z-R relationship. This is used to generate an ensemble of perturbation fields. The perturbations are correlated in space and time to replicate the correlations in the random fluctuations about an assumed Z-R relationship.

The advantage of statistical models based on the error measured by comparing radar estimates to a reference, is the whole error is measured. Not just the error in the assumed VPR and Z-R relationship. Germann et al (2009) (hereafter GM09) looked at the location dependence of the errors in radar data as compared to gauge measurements. This has advantages where location dependent variables affect the measurement (such as beam shielding or orographic enhancement).

Here two models for ensemble generation, one falling into each category have been built and tested. Section 2 describes the two models, and the results of the verification of the models against rain gauge ground truth measurements are given in Section 3. In order to be used within STEPS, the ensemble generator needs to generate an ensemble time series of fields of instantaneous rain rate at time intervals less than 15 minutes suitable for the following purposes: a) for use as initial conditions for generating STEPS ensemble nowcasts of rain rate and accumulation; b) to generate ensembles of T+0 rainfall accumulations for use in flash flood forecasting and other downstream user applications; c) for wider use to generate a range of QPE uncertainty products tailored to customer requirements. STEPS currently issues extreme rainfall alerts (ERAs) to the Environment Agency (EA) when rainfall amount and accumulation time thresholds are exceeded (a part of (a) and (b)). The statistics used for the comparison will focus on this requirement.

2. Descriptions of the modelling approaches

2.1 Overview of the model using a statistical description of the differences between radar and gauge

An ensemble generator has been built following the methodology described by GM09. This approach allows a time series of ensembles of radar accumulation fields to be generated by adding perturbations to the radar data. Historical rain gauge data is analysed to build a statistical model of the errors and this model is then used to generate the ensembles. The mean, covariance and temporal correlation of the errors are calculated using historical data, where the error at location \((x,y)\) at time \(t\), is defined in Equation 1 and \(R_{true}\) denotes the ground truth measurement (in this case, gauges).
\[ c(x, y, t) = 10 \log_{10} \left( \frac{R_{\text{true}}(x, y, t)}{R_{\text{radar}}(x, y, t)} \right) \]

*Equation 1*

\( R_{\text{radar}} \) is used as a weight when calculating the covariances and mean error vector. This ensures that errors due to the discrete nature of the tipping bucket rain gauge accumulations are not given undue influence. However this means that spurious echoes in the radar data that are coincident with precipitation will be given a large weight.

GM09 calculate a covariance matrix of \( \varepsilon \) (with \( R_{\text{radar}} \) weights) once, offline, using historical data and the decomposition of \( C \) is stored and used to generate every ensemble. However the covariance can only be measured at gauge locations so either the covariance matrix needs to be interpolated over the remaining pixels, or this can happen after the perturbations are generated. Here, the random component of the perturbation has been generated at gauge location and interpolated and then combined with an interpolated mean error field.

The perturbations are applied in logarithmic units as in *Equation 2*, where \( i \) refers to ensemble member \( i \).

\[ 10 \log_{10} \left( R_{i,j} \right) = 10 \log_{10} \left( R_{\text{radar}} \right) + \delta_{i,j} \]

*Equation 2*

An autoregressive lag-2 (AR(2)) model is used to impose temporal correlation on the perturbations generated using this method.

2.2 Data used for error characterization

Six months EA gauge data (15 minute accumulations) from 2006 were used to calculate the errors. The error structure of 15 minute accumulations will be different to that of instantaneous rates. However, for the purposes of this trial, the error characterization for 15 minute accumulations has been used and applied to precipitation rate data in order to produce the ensembles.

FIG. 1 shows the mean error vector calculated at each gauge location and subsequently interpolated over a suitable domain using Delaunay Triangulation. It also serves to show the domain over which the test ensembles were generated. To prevent large areas being interpolated over using data from few gauges, the output of the Delaunay triangulation was limited so that any triangles with a facet longer than 42 km were not used.

Some quality control of the covariance matrix was necessary for use in generating the ensembles. Gauges with fewer than 200 samples used to calculate the variances were removed from the matrix (1 row and 1 column removed). Singular value decomposition was used to decompose the covariance matrix:

\[ C = U W V^T \]

*Equation 3*
The numerically ordered vector $W$ allows some data to be discounted in order to gain a better representation of $C$ in the generated ensembles. This was optimised to produce the most representative covariance in the generated perturbations.

### 2.2 A model based upon the characterization of $Z$-$R$ and range-height error sources

This approach (ZRVPR model hereafter) follows the cascade methodology as used in STEPS. Two noise cascades are created using space-time models for the sampling errors relating to the height of the radar beam relative to the freezing level and the fluctuations in the $Z$-$R$ relationship, the error term has been modelled as a stochastic field with the same space-time characteristics as the rainfall. Errors arising from fluctuations in the $Z$-$R$ relationship vary both in space and time on scales that are about 40 km in space and 60 minutes in time. Lee and Zawadzki (2005a, 2005b, and 2006) modelled these fluctuations using a multiplicative cascade structure and this component of the observation error has been included in the STEPS algorithm.

Errors that arise from the $Z$-$R$ relationship are not necessarily the major source of error in a QPE product, particularly on days when the bright band is close to the surface and errors due to the vertical profile of reflectivity dominate at all but the closest of distances from the radar. Errors in estimates of surface rainfall increase rapidly when the base scan of the radar observes the snow above the wet-bulb freezing level and not the rainfall below. The $Z$-$R$ relationship assumes a drop size distribution and fall speed that is only appropriate for rainfall. The fall speed of snow is very much less than that of rainfall, so snow can drift significant distances in the horizontal during the time it takes the snow that has been observed at some height above the wet-bulb freezing level to melt and fall to the ground as rain.

Three cases of rainfall in Australia were used to explore the correlation between the radar reflectivity at the ground and a radar observation at a height which is above the wet bulb freezing level. The cases were selected to include two storms from Sydney and Brisbane where the wet-bulb freezing level was greater than 3 km, and a storm from Melbourne where the wet-bulb freezing level was close to 1 km.

**FIG. 3.** Correlation as a function of height separation for pairs of radar observations where one observation is at the base scan and the other is below the wet bulb freezing level.

**FIG. 4.** Correlation as a function of height separation for pairs of radar observations where one observation is at the wet-bulb freezing level and the other above it.

FIG. 3 and FIG. 4 show the correlation as a function of height separation for pairs of observations that are below the wet-bulb freezing level and pairs of observations that are at and above the wet-bulb freezing level. A model for the correlation of the form

$$\rho = a \exp \left[ - \left( \frac{h}{b} \right)^c \right]$$
where \( h \) is the height separation and \( a, b, \) and \( c \) are constants, has been fitted to the data and is also shown in the figures. A model of the form described in Equation 4 is a reasonable representation of the correlation between two points that are vertically separated. The model used to describe the errors is given in more detail by Seed (2010).

2.3 Comparison of the two models

One of the primary differences in the two models is the treatment of biases (systematic errors) in the radar data, the ZRVPR model tries to model two of the sources of random error in the radar data, whereas the GM09 model also models the systematic error. The systematic errors can be caused by orography, beam shielding and height of the radar beam to name but a few.

The GM09 model uses rain gauges as a ground truth measurement, although efforts have been made to quality control the data, there may be errors present in the gauge data, especially for lower precipitation amounts, where the discrete nature of the data from tipping bucket rain gauges introduces some error. The reliance on rain gauges also limits the domain over which you can generate the ensembles: The perturbations are interpolated over regions where there are no co-located gauges, so it is assumed that the perturbations vary smoothly over these regions.

3. Evaluation of the models

3.1 Overview of the verification methodology

The summary statistics presented in this section relate to the validation of 30 member QPE ensembles generated using the two models. Accumulation data from the EA tipping bucket rain gauges, integrated over a 60 minute period were used as a reference to produce the statistics. The models generated QPE ensembles for five UK based case studies, selected to encapsulate a range of precipitation events. These events are summarised in Table 1 and comprise of approximately 5600 non-zero rain gauge samples. Four statistics are compared here: the reliability, the resolution, the relative operating characteristic (ROC) and the ensemble spread-RMS error relationship.

<table>
<thead>
<tr>
<th>Date of event</th>
<th>Event description</th>
<th>Features of interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 July 2009</td>
<td>Heavy, thundery showers over England, heaviest in the south-east around early evening.</td>
<td>Very intense precipitation in thunder storms.</td>
</tr>
<tr>
<td>29 July 2009</td>
<td>Broad band of precipitation associated with developing frontal wave over the Midlands.</td>
<td>Widespread rain with relatively small variance in intensity within the raining area.</td>
</tr>
<tr>
<td>27 August 2009</td>
<td>A narrow band of showers moving eastwards across England and Wales during the afternoon and evening.</td>
<td>Showers demonstrating interesting organisation along trough/frontal line.</td>
</tr>
<tr>
<td>6 October 2009</td>
<td>Cold frontal precipitation moving south-south-eastwards across England and Wales during the afternoon and evening.</td>
<td>A large drop in wet-bulb freezing level across the frontal zone.</td>
</tr>
<tr>
<td>1 November 2009</td>
<td>Widespread precipitation associated with a deep depression moving north-eastwards over the UK.</td>
<td>Widespread and intense rainfall caused localised flooding in many areas of the UK.</td>
</tr>
</tbody>
</table>

3.2 Comparative verification statistics for 60 minute rain accumulation

A perfect ensemble generator should produce QPE probabilities that match the observed frequencies of particular events. Attributes diagrams are plotted in FIG. 5 for both of the models for events greater than 1 mm. The ZRVPR model is under-confident of events with low frequencies and over-confident of events with higher frequencies. The statistical model is consistently over-confident, which is indicative of insufficient ensemble spread, though has a slope closer to the perfect reliability line. However, both the ensembles have similar Brier, Resolution and Reliability skill scores, where the sample climatology was used as the reference, and both ensembles show skill above the reference.

A skillful ensemble should deviate from climatological probabilities to resolve the uncertainties at a particular time (resolution). However the resolution component of the Brier skill scores for both ensembles is similar, and at
all thresholds are below 0.03 (where 0 indicates perfect resolution). As the radar data itself is used as input to both models, it is unsurprising that they show good resolution.

**FIG. 5.** Attributes diagram for statistical model (left) and ZR VPR model (right) run on 5 case study events where hourly accumulation is greater than 1 mm.

**FIG. 6.** RMS error of the radar compared to gauge (labelled “radar”), RMS error of a random ensemble member compared to the ensemble mean (labelled “perfect”) and randomly ordered ensemble spread and RMS error data (labelled “no skill”). All plotted against the spread of the ensemble for the statistical model (left) and for the ZR VPR model (right) run on 5 case study events.

**FIG. 7.** ROC curves for the statistical model (left) and the ZR VPR model (right) for different thresholds. Lines denote ROC of the ensembles (probabilistic); points are the ROC of the radar (deterministic).

The ensemble spread should be representative of the actual error in the measurement, in order to test this, ensemble spread/ RMS error pairs have been put into bins containing equal amounts of data and plotted in FIG. 6. These graphs also show the RMS error of an ensemble with perfect spread and with no skill. Both ensemble generators exhibit insufficient ensemble spread (the measured RMS error is greater than the corresponding ensemble spread).
spread). The statistical model generated more ensembles with larger spread, as demonstrated by the RMS error of the no-skill line, the values of standard deviation bins, which are smaller for the ZRVPR model than the statistical model, and the RMS error curve, which is slightly closer to the perfect RMS error curve.

The ROC curves of the ensembles are shown in FIG. 7, the statistical model has slightly better ROC scores than the ZRVPR model, but both are relatively consistent over several thresholds and have ROC scores above 0.87. This shows both ensembles have the ability to discriminate between two outcomes (exceedance / non-exceedance of a threshold) with some skill.

4. Evaluation of the models

Both of the ensemble generators have shown skill when compared with the sample climatology. Overall, the statistical model shows slightly more skill, though it would be useful to test this over a larger sample. The ROC curves and the reliability of the ensemble are the statistics most relevant to the issuing of ERAs, for both of these the statistical model outperforms the ZRVPR model. However, there are limitations to the further development of the statistical model: one way to improve the statistical model would be to assess the error structure under different weather regimes, wet bulb freezing level heights, convective or stratiform precipitation, or for different seasons. This would make the statistical ensemble generator more dynamic; however there is a limitation on how many categories you can use, due to the amount of historical data available. Schemes which model individual sources of error have much more scope for improvement; the ZRVPR ensemble generator shows a good deal of skill, given only errors in the VPR and Z-R relationship are modelled. The primary advantage of the statistical model is likely to lie in the location dependence of the error and the modelling of the systematic errors in the radar data which are highly correlated in space, as shown by FIG. 1.

Acknowledgment

The authors would like to thank the Environment Agency (England and Wales) and the Scottish Environment Protection Agency for providing the rain gauge data.

References


