A new method for the characterisation and verification of local spatial predictability for convective scale ensembles

Article

Accepted Version

A new method for the characterisation and verification of local spatial predictability for convective scale ensembles

Seonaid R. A. Dey,[a]*Nigel M. Roberts,[b] Robert S. Plant[a] and Stefano Migliorini[c]

¡[a]*Department of Meteorology, University of Reading, Earley Gate, PO Box 243,Reading,RG6 6BB*

[b]*MetOffice@Reading, Met Office, Reading, UK*

[c] *Met Office, Exeter, UK*

*Correspondence to: Seonaid Dey, Department of Meteorology, University of Reading, Earley Gate, PO Box 243,Reading,RG6 6BB. E-mail: s.dey@pgr.reading.ac.uk

The use of kilometre-scale ensembles in operational forecasting provides new challenges for forecast interpretation and evaluation to account for uncertainty on the convective scale. A new neighbourhood based method is presented for evaluating and characterising the local predictability variations from convective scale ensembles. Spatial scales over which ensemble forecasts agree (agreement scales, S^A) are calculated at each grid point ij, providing a map of the spatial agreement between forecasts. By comparing the average agreement scale obtained from ensemble member pairs ($S_{ij}^{A(\overline{mm})}$), with that between members and radar observations ($S_{ij}^{A(\overline{mo})}$), this approach allows the location-dependent spatial spread-skill relationship of the ensemble to be assessed. The properties of the agreement scales are demonstrated using an idealised experiment. To demonstrate the methods in an operational context the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are calculated for six convective cases run with the Met Office UK Ensemble Prediction System. The $S_{ij}^{A(\overline{mm})}$ highlight predictability differences between cases, which can be linked to physical processes. Maps of $S_{ij}^{A(\overline{mm})}$ are found to summarise the spatial predictability in a compact and physically meaningful manner that is useful for forecasting and for model interpretation. Comparison of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ demonstrates the case-by-case and temporal variability of the spatial spread-skill, which can again be linked to physical processes.

Key Words: convective-scale ensemble forecasting neighbourhood verification spatial spread-skill

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.2792

1. Introduction

Recent increases in computing power have allowed a shift towards higher resolution numerical weather prediction (NWP) models in which convection can be explicitly simulated. However, although these high resolution simulations produce realistic features (Mass *et al.* 2002; Lean *et al.* 2008), errors grow rapidly (Hohenegger and Schär 2007; Melhauser and Zhang 2012; Radhakrishna *et al.* 2012), and small-scale predictability is maintained for only a few hours. Hence, to fully benefit from convection permitting NWP it is necessary to understand and quantify the forecast uncertainty. Ensembles have been successfully used for this purpose in larger scale NWP (e.g. Palmer 2000, and references therein), and are now being run at convection permitting resolutions. In particular, convection permitting ensembles have been investigated for a range of case studies (Hanley *et al.* 2011; Leoncini *et al.* 2013; Clark *et al.* 2013; Hanley *et al.* 2013), nowcasting applications (Migliorini *et al.* 2011), and are now run, or about to be run, operationally at several forecasting centres (Baldauf *et al.* 2011; Gebhardt *et al.* 2011; Bouttier *et al.* 2012; Golding *et al.* 2014).

However, questions remain about the best methods for interpreting and evaluating convection permitting ensembles. In particular the ensemble mean, successfully used for smoothly varying, large-scale fields, may not be physically appropriate at the convective scale (e.g. Ancell 2013). This is particularly true for quantities with high spatial variability, such as precipitation forecasts; for these fields the ensemble mean field does not retain the physical structures of the individual member forecasts. Other open questions relate to the interpretation of forecast uncertainty, given the tiny fraction of realisations covered by the ensemble members, and to methods of forecast verification. Standard verification measures, such as the Root Mean Squared Error (RMSE e.g. Wilks 2011) are unsuited to the convective scale as they overly penalise spatial differences. Several more suitable methods have been proposed for verifying deterministic forecasts (e.g. Ebert 2008; Gilleland *et al.* 2009; Johnson and Wang 2012) that can now be developed for convection-permitting ensembles.

It has been shown that the skill of convective-scale forecasts is scale dependent, with skill increasing as a function of spatial scale (Roberts and Lean 2008; Roberts 2008; Ben Bouallègue and Theis 2014; Mittermaier *et al.* 2013; Mittermaier 2014). Clark *et al.* (2011) showed this was also true for ensemble forecasts, with ensemble skill increasing with both spatial scale and ensemble size. Given this dependence on spatial scale, methods have also been developed to evaluate the differences between ensemble member forecasts, a measure of the ensemble spread, at different spatial scales. In particular, Johnson *et al.* (2014) used wavelet decomposition to investigate perturbation growth, Surcel *et al.* (2014) used spectral decomposition to investigate the filtering properties of the ensemble mean and Rezacova *et al.* (2009); Zacharov and Rezacova (2009); Duc *et al.* (2013); Dey *et al.* (2014) used the Fractions Skill Score (FSS Roberts and Lean 2008; Roberts 2008) to develop a neighbourhood-based approach to calculate the ensemble spread and skill spatially.

Given the scale dependence of forecast errors, it is important to determine the scales over which forecasts should be considered to have skill. In Roberts and Lean (2008), the "skillful scale" was defined as the scale which gave an FSS value of $0.5 + f_0/2$, where f_0 is the total fraction of points in the domain exceeding the threshold. Using idealised and real examples, Roberts and Lean (2008) inferred that for small rainfall coverage (small f_0), the FSS equals this value when the neighbourhood size is equal to twice the separation of forecast rainfall features. More recently, work by Skok (2015) has shown analytically that, for simple idealised configurations in an infinite domain, the neighbourhood size is twice the spatial separation of precipitation objects when the FSS has a value of 0.5. Thus, using the FSS, the scale can be found at which a forecast is, on average, skillful across the model domain. Using the methodology of Dey *et al.* (2014), this reasoning can also be extended to the comparison of other forecast fields, for example in an ensemble. For this general situation the skillful scale generalises to a believable scale, the scale at which the fields from independent forecasts become sufficiently similar so that

the forecast forms useful, trustworthy guidance (assuming the ensemble is able to reproduce the range of possible scenarios).

The measures of skillful and believable scales of Roberts and Lean (2008); Dey *et al.* (2014) can provide a compact summary of both the domain-averaged spatial error and spread of an ensemble. However, as highlighted in Dey *et al.* (2014), by considering only one value to represent the whole domain, differences in spatial agreement across different parts of the domain are missed. These differences will arise because different meteorological phenomena, such as convective and frontal precipitation, may have an inherently different predictability and ensemble spread. Hence, it would be informative to examine the ensemble spatial characteristics in a manner that preserves location-dependent information.

Using similar principles to the FSS, this paper presents a new, location–dependent measure of the scales over which precipitation fields (either forecasts or observations) are acceptably similar (defined in Section 3.2). When calculated for ensemble members, these agreement scales, denoted as $S_{ij}^{A(\overline{mm})}$, indicate the area (surrounding each grid point) over which precipitation features in the individual member forecasts would be expected to occur. When ensemble members are compared with radar observations, the agreement scales, denoted $S_{ij}^{A(\overline{mo})}$, indicate the area (surrounding each grid point) over which precipitation features in the member forecasts agree with the radar observations. Note that, independently, the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ do not provide a measure of forecast accuracy. However, by comparing the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ the spatial spread-skill relationship of the ensemble can be investigated.

In Section 2 the neighbourhood approach is introduced, and spatial predictability is defined. The methods used to calculate the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are presented in Section 3, and compared to and contrasted with the FSS. For a new method to be of use for interpreting forecast performance, it is essential that it behaves in a sensible manner, and gives useful and robust information. To investigate these requirements for the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, an idealised ensemble was employed, with simple geometric forecast fields. By considering an idealised ensemble the method can be examined in detail for a large number of cases. The idealised experiments are described in Section 4. In Section 5 six convective case studies are presented using forecasts from the operational Met Office Global and Regional Ensemble Prediction System UK ensemble (MOGREPS-UK Mylne 2013; Golding *et al.* 2014). The aim is to understand how the agreement scales behave, what information they can provide about the ensemble spatial spread and error, and how this relates to physical processes. We do not aim to give a statistical verification of MOGREPS-UK. The overall conclusions are presented and discussed in Section 6.

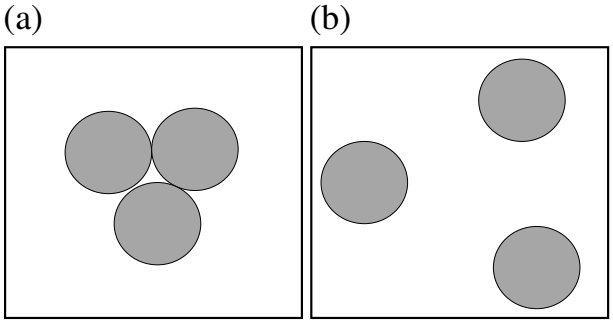

Figure 1. Schematic representing precipitation forecasts from three different ensemble members (grey circles, one per member). Each grey circle represents an area of forecast precipitation, say with a uniform rain rate of 0.1mm hr^{-1}. Events are shown with different levels of spatial predictability: (a) a spatially predictable event and (b) a less spatially predictable event.

2. Spatial predictability and the neighbourhood approach

Here, and for the remainder of this paper, the term "spatial predictability" refers to differences in the location of precipitation in the ensemble member forecasts. Cases where the member forecasts are in close agreement that precipitation will occur locally are termed spatially predictable, and cases where the location of precipitation is uncertain (i.e. when ensemble members produce rain at different places in the domain) are termed less spatially predictable. Examples with different spatial predictability are shown schematically in Fig. 1. Note that this definition of spatial predictability only refers to the positional differences between the ensemble member forecasts (i.e. amplitude errors are not included).

We use a neighbourhood based approach to quantify differences between precipitation forecasts. In the neighbourhood approach, forecasts are compared over differently sized areas

(neighbourhoods). Summary measures are then used to compare the forecasts over these areas. For example, the amount of precipitation exceeding a specified threshold, the maximum or average precipitation value of all raining points, or the average precipitation over all points in the neighbourhood, could be considered.

In this paper the average precipitation rate is taken from all points in the neighbourhood, including points with zero precipitation (without thresholding). Hence no distinction is made between amplitude, timing and spatial structural differences. This method was chosen to be as generally applicable as possible, giving an overview of the forecast differences, and keeping the number of parameters to a minimum. The aim is to provide a single summary measure of the location-dependent forecast differences. Of course, this comes at the cost of providing less detailed information about individual components such as timing errors, although some timing errors due to advection (rather than initiation or decay) are naturally included in the spatial approach.

It is informative to relate the neighbourhood approach used in this paper (which calculates the spatial agreement between fields; to be discussed in Section 3) to the spatial predictability as defined above. First consider the comparison of two binary fields, for example created by setting precipitation values to zero/one dependent on whether they are below/above a predetermined threshold. In this case, any differences in the neighbourhood averaged values of the two fields will relate to differences in the location of precipitation. Hence, in this situation, the forecasts will agree over a smaller/larger neighbourhood for cases with higher/lower spatial predictability. Thus, when binary fields are considered, the spatial predictability relates directly to the neighbourhood size. Next consider precipitation fields where no threshold has been applied. The spatial predictability will still influence the neighbourhood size over which the forecasts agree, but any difference in the magnitude of the two fields will also contribute. This is also true for other fields which, like precipitation, have high small scale variability and a background value of zero. Fields which vary smoothly over large scales (larger than the neighbourhood sizes being considered), with large scale gradients, require a different interpretation. In these instances there is no longer a direct link between neighbourhood size and spatial predictability. Hence, for large scale fields the neighbourhood approach compares only the fractional difference between the two fields.

3. Calculation of location-dependent agreement scales

3.1. *Overview of method*

First we focus on calculating location-dependent agreement scales for two different fields, say two ensemble member precipitation forecasts, denoted f_1 and f_2 . At each grid point in the domain, we search for the minimum neighbourhood size (hereafter the scale) over which suitable agreement between f_1 and f_2 is obtained. Here, and for the remainder of this paper, the scale is defined as the number of grid points from the centre to edge of the neighbourhood (excluding the central grid point). For example a 3 by 3 neighbourhood would have a scale of 1, and a 1 by 1 neighbourhood (a single grid point) would have a scale of zero. The scale at which suitable agreement is obtained between the forecasts f_1 and f_2 at this central point (x,y)=(i,j), will be referred to as the agreement scale $S_{ij}^{A(f_1 f_2)}$. Note that the $S_{ij}^{A(f_1 f_2)}$ provides a measure of the agreement between two fields, and is not a measure of forecast performance. For example, large/small values of $S_{ij}^{A(f_1 f_2)}$ indicate that large/small neighbourhoods are needed to obtain sufficient agreement between the fields, but this should not be interpreted as poor/good forecast performance.

The calculation of $S_{ij}^{A(f_1 f_2)}$ proceeds as follows:

1. One grid point in the domain is selected. Call this point P at i, j.
2. The precipitation values from the two forecasts are compared at point P, and their similarity assessed using the methods presented in Section 3.2.
3. If the forecasts are found to be suitably similar (defined in Section 3.2), then the agreement scale at point P, $S_{ij}^{A(f_1 f_2)}$, is the grid scale. If the fields are not suitably similar, then a square neighbourhood of scale 1 (3 by 3 grid points), centred upon the point P, is considered.

4. The spatial average precipitation amount over this neighbourhood is calculated separately for f_1 and f_2, as discussed in Section 2. Forecasts f_1 and f_2 are again compared, this time using the average precipitation amount over the neighbourhood, and their similarity assessed.
5. If, this time, the forecasts are found to be suitably similar, then a neighbourhood of size 1 is the agreement scale. If the fields are not suitably similar, then the scale is increased by 1 (i.e. to give a 5 by 5 grid point neighbourhood).
6. Steps 4 and 5 are repeated, for incrementally larger scales, until a scale has been found for which the forecasts are suitably similar around point P. Note that this is defined as the minimum agreement scale for comparing these forecasts: generally, it would be expected that the forecasts would also be in agreement over larger neighbourhoods.
7. Steps 1 to 6 are repeated for each grid point in the domain.

In point 6 it has been implicitly assumed that f_1 and f_2 always become increasingly similar as they are compared over increasingly large neighbourhoods. Although this has been shown to be true for precipitation fields on average (e.g. Roberts and Lean 2008; Clark *et al.* 2011; Mittermaier *et al.* 2013), there are situations when this will not be the case, for example when forecasts have reasonable agreement over a small neighbourhood (say they both predict a light shower), but as the neighbourhood increases, one field has no rain whereas the other has large amounts of rain. This situation will result in a noisy map of $S_{ij}^{A(f_1 f_2)}$, as neighbouring grid points could (depending on the exact field characteristics) have very different values. However, as the $S_{ij}^{A(f_1 f_2)}$ are only used after averaging over a number of field comparisons (to be discussed in Section 3.4), this is not found to be a problem in practice. Another instance when f_1 and f_2 will not become increasingly similar with increasing scale is when the fields have a large scale gradient. Although fields of this nature are unlikely to be seen for precipitation, the criterion for deciding whether the forecasts are suitably similar is designed to give a sensible outcome in the presence of such gradients as discussed in Section 3.2.

3.2. Criterion for assessing forecast similarity

It remains to define how the forecast similarity is assessed and how "suitably similar" is defined. Consider the comparison of two fields f_1 and f_2 for a given neighbourhood size (scale) S, and at grid point i, j. For both fields, the average over all points in the neighbourhood is taken: we denote these averages as f_{1ij}^S and f_{2ij}^S. The fields (assuming at least one average is non zero) are compared by taking the ratio of the squared difference between these averages and the sum of their squares:

$$D_{ij}^S = \begin{cases} \frac{(f_{1ij}^S - f_{2ij}^S)^2}{(f_{1ij}^S)^2 + (f_{2ij}^S)^2} & \text{if } f_{1ij}^S > 0 \text{ or } f_{2ij}^S > 0 \\ 1 & \text{if } f_{1ij}^S = 0 \text{ and } f_{2ij}^S = 0 \end{cases} \quad (1)$$

D_{ij}^S varies from zero to one. The numerator is a direct measure of the difference between the fields; the denominator a normalising factor selected such that comparison between a forecast which captures some precipitation ($f_{1ij}^S > 0$) and one with no precipitation ($f_{2ij}^S = 0$), gives a D_{ij}^S value of one. This is a convenient choice of normalisation: other normalisation factors are possible and would not change the overall method and conclusions presented here. Note that in the formulation of Eq. 1 positive fields have been assumed.

The fields are then deemed sufficiently similar (i.e. to be in agreement) at scale S if

$$D_{ij}^S \leq D_{\mathrm{crit},ij}^S \quad (2)$$

where

$$D_{\mathrm{crit},ij}^S = \alpha + (1-\alpha)\frac{S}{S_{\mathrm{lim}}}. \quad (3)$$

The agreement scale between forecasts f_1 and f_2 at point (i, j) is denoted $S_{ij}^{A(f_1 f_2)}$, and defined as the minimum scale S at which Eq. 2 is met. The minimum possible $S_{ij}^{A(f_1 f_2)}$ is zero (showing agreement between the forecasts at the grid scale) and the maximum possible $S_{ij}^{A(f_1 f_2)}$ is S_{lim} (showing no agreement between the forecasts, or no rain in

the neighbourhood for at least one of the forecasts). The interpretation of the agreement scales is further discussed in Section 4.2.

At the grid scale ($S = 0$) the second term on the right-hand side of Eq. 3 is zero and the constant α controls the acceptable fractional difference between f^S_{1ij} and f^S_{2ij}. Different values of α can be selected: $0 < \alpha \leq 1$ where $\alpha = 0$ corresponds to no bias being tolerated at the grid scale and for $\alpha = 1$ any bias is tolerated. $S_{\rm lim}$ is a predetermined, fixed maximum scale and, by construction, Eq. 2 is always satisfied at the scale $S_{\rm lim}$.

This maximum scale is important for both computational and scientific reasons. Computationally, it is more expensive to make the necessary calculations at increasingly larger scales, which is an important consideration in an operational context. Scientifically, there is a scale above which it is no longer appropriate to consider high resolution forecasts: for example, when there also exists a lower resolution forecast (e.g from a global model), better placed to assess large scale errors.

Additionally, it is necessary to separate cases where two forecasts predict the same event, but at a different location, from those where each forecast predicts essentially different events. Consider the comparison of two forecasts which both produce precipitation, but at a different location in the domain. In some situations the forecasts will be predicting a region of precipitation with the same physical characteristics. In this case, we could say that the same event is predicted by both forecasts, but with uncertainty in the location. This is the location uncertainty that can be quantified using the agreement scales, $S^{A(f_1 f_2)}_{ij}$. However, it is also possible that the forecasts are predicting different events entirely, such as convective showers due to low level convergence in one, and convection associated with a frontal system in another. In this second situation, the differences between the forecasts are not representative of their spatial uncertainty, and hence the values of $S^{A(f_1 f_2)}_{ij}$ could be misleading. Thus, when the agreement scales are calculated, it is an underlying assumption that the same events are being forecast by the two fields, but at different locations. As the scale increases this assumption is likely to be less valid and the forecasts are more likely to be representing different physical phenomena. Note that this assumption is needed because of the spatial neighbourhood approach where forecasts at different locations in the domain are considered together: it is not needed in traditional measures which compare fields at the same grid point only.

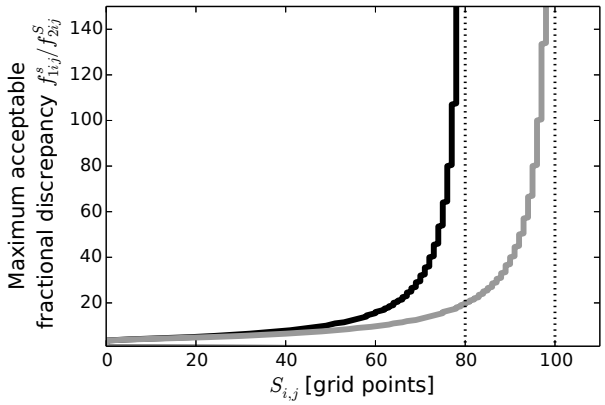

Figure 2. Maximum acceptable fractional discrepancy between f^S_{1ij} and f^S_{2ij} as a function of neighbourhood size S, for $\alpha = 0.5$ and $S_{\rm lim} = 80$ (black) and 100 (grey).

Equation 3 is formulated so that, as forecast differences increase, the scales of acceptable agreement tend smoothly towards $S_{\rm lim}$. Specifically, the fractional difference between the fields that is considered acceptable increases for increasing S until, at $S_{\rm lim}$ itself, any difference is accepted. The dependence of the acceptable fractional discrepancy between the fields as a function of spatial scale S is shown in Fig. 2 for $\alpha = 0.5$ and $S_{\rm lim}$=80 or 100. Thus, the agreement scales close to $S_{\rm lim}$ are highly dependent on this value. However, as long as $S_{\rm lim}$ is chosen to be sufficiently large that any useful information from the convective-scale forecasts has already been extracted, this will not effect the overall message from the agreement scales.

In the work presented here, values of $\alpha = 0.5$ and $S_{\rm lim} = 80$ have been used. For specific applications that require a more/less stringent match lower/higher values of α could be selected. For the forecasts analysed in Section 5, the maximum scale of 80 grid points corresponds to a square neighbourhood of 25921 grid points with total width 354.2 km (the model has a grid length of 2.2 km). Note that experiments were conducted with different values of both α and $S_{\rm lim}$ but, as these modifications did not affect the overall conclusions, they are not presented here.

3.3. Comparison with the Fractions Skill Score

It is informative at this stage to compare Eq. 2, defining the agreement scale at a particular location, with the FSS useful scale as defined in Roberts and Lean (2008) and given by the neighbourhood size at which the FSS=$0.5 + f_0/2$. Although there are some similarities between the agreement scales and the FSS useful scale, there are many fundamental differences. Hence these measures should not be confused.

As detailed in Roberts and Lean (2008), the FSS compares a forecast with gridded observations over different predetermined neighbourhood sizes. There are three steps to calculating the FSS between a forecast field and observations. First a threshold is selected, either as a fixed value (e.g 4 mm hr^{-1}) or as a percentile (e.g top 1% of precipitation field). The field is then converted to binary form with grid points set to 1 for values above the threshold and 0 otherwise. Next, a neighbourhood size is selected and, for each neighbourhood centred upon each grid point, the fraction of grid points with the value '1' within this square is computed. This step is completed for both fields to give two "fields of fractions", f and o. Finally, the FSS is calculated by comparing the mean squared error (MSE) of the fields of fractions with a reference MSE, MSE_{ref}, the largest possible MSE that can be obtained from the fields of fractions. For a predetermined neighbourhood size and domain size N_x by N_y grid points the FSS is then given by:

$$\mathrm{FSS} = \frac{\mathrm{MSE}}{\mathrm{MSE}_{ref}} = \frac{\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}[f_{i,j} - o_{i,j}]^2}{\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}[f_{i,j}^2 + o_{i,j}^2]}. \qquad (4)$$

where the sums are over all grid points in the domain.

There are some similarities between the method of calculating the agreement scales (Eq. 2) and the FSS (Eq. 4). For example, in both calculations, the difference of quantities squared is divided by the sum of their squares. However, there are also some important differences.

- The FSS is a score which can be used directly for forecast verification. In contrast, the agreement scales here provide a general measure of the agreement between different fields and do not directly measure forecast performance.
- The FSS gives a single domain-wide value for the spatial agreement, whereas the agreement scales provide a location-dependent map of the spatial agreement. Therefore, in the FSS, the squared difference between fields, and sum of the squares of the two fields, are further summed over all points in the domain. This is not the case for the agreement scales (Eq. 2), where each location is considered separately. The denominator of the FSS equation (Eq. 4) is the maximum possible difference that can be obtained from two fields of fractions, whereas in Eq. 2 the denominator is a convenient normalisation factor.
- Scales of interest are obtained for $S_{ij}^{A(f_1 f_2)}$ and the FSS when a criterion exceeds a value of 0.5 plus an extra term. It should be stressed that these criterion do not have the same meaning. For the FSS, the value "0.5" relates directly to the spatial separation of precipitation features (Roberts and Lean 2008; Skok 2015), whereas in Eq. 2 the value α (equal to 0.5 here) controls the bias considered acceptable. The additional terms in the criteria also have different functions in each of the two measures: that used for the FSS relates to the coverage of precipitation in the domain, whereas that in Eq. 2 ensures that the search algorithm always returns a meaningful scale.
- Although both equations consider errors both in precipitation location and precipitation amount, these are treated differently. In particular, the FSS is applied to precipitation fields that have undergone thresholding to produce binary fields. In contrast, the agreement scales compare the precipitation amounts themselves (Eq. 2). This is a more general approach which does not require a threshold to be defined, and directly considers the scale-dependent bias between the fields.

3.4. Calculations for an ensemble

Dey *et al.* (2014) used the FSS to estimate the domain-averaged spatial ensemble spread and skill by comparing all

independent pairs of ensemble members, and all ensemble member-radar pairs. Here a similar approach is applied to the agreement scales $S_{ij}^{A(f_1 f_2)}$ to calculate how the spatial agreement between ensemble members, and the spatial agreement between ensemble members and radar observations, varies with location across the domain.

To give a measure of the location-dependent agreement between ensemble members, the agreement scales $S_{ij}^{A(f_1 f_2)}$ are calculated separately for each independent pair of ensemble member forecasts. This gives

$$N_p = \frac{N(N-1)}{2} \tag{5}$$

fields of agreement scales for an ensemble of N members. It is necessary to provide a summary value of all these fields to quantify the overall spatial uncertainty of the ensemble at each point in the domain. Here, to get an agreement scale representative of the ensemble, the mean is taken, at each grid point in the domain, over the N_p values of $S_{ij}^{A(f_1 f_2)}$. Hence, for an ensemble of twelve members, 66 agreement scales would be separately calculated ($N_p = 66$), and the mean of these 66 fields would be taken at each grid point in the domain. As the distribution of the N_p agreement scales was found to be uni-modal, the mean is an appropriate value to characterise the distribution of individual scales. This mean field indicates the average agreement scale between the ensemble members at each grid point, and is denoted $S_{ij}^{A(\overline{mm})}$. It represents the scales over which the ensemble should be evaluated (believable scales), and the area over which individual features seen in the member forecasts should be expected to occur. Mathematically, the $S_{ij}^{A(\overline{mm})}$ are given by

$$\mathrm{S}_{ij}^{A(\overline{mm})} \equiv \frac{1}{N_p} \sum_{f_1=1}^{N-1} \sum_{f_2=f_1+1}^{N} \mathrm{S}_{ij}^{A(f_1 f_2)}. \tag{6}$$

In a similar manner to $S_{ij}^{A(\overline{mm})}$, we can also characterise the average spatial differences between ensemble members and radar observations, denoted $S_{ij}^{A(\overline{mo})}$. It is necessary to use radar observations for this comparison due to their high spatial coverage. To calculate the $S_{ij}^{A(\overline{mo})}$, the mean is taken, at each grid point, over the fields of agreement scales calculated from comparing all N member-radar pairs:

$$\mathrm{S}_{ij}^{A(\overline{mo})} \equiv \frac{1}{N} \sum_{f=1}^{N} \mathrm{S}_{ij}^{A(fo)}. \tag{7}$$

Therefore, for an ensemble of twelve members, there are 66 pairs contributing to the $S_{ij}^{A(\overline{mm})}$, but only twelve pairs contributing to the $S_{ij}^{A(\overline{mo})}$.

The $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are consistently defined, and measure respectively the average agreement between ensemble members, and the average agreement between ensemble members and radar observations. As the average agreement between ensemble members should, for a well spread ensemble system, be representative of the average difference between ensemble members and observations, the ensemble performance can be verified through comparing the $S_{ij}^{A(\overline{mm})}$ with the $S_{ij}^{A(\overline{mo})}$. In Section 4, an idealised system is used to show this comparison does indeed give useful information about the ensemble system.

It is informative to relate the comparison of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ to the traditional ensemble spread-skill relationship, which has proved useful for the analysis of synoptic scale ensembles (e.g. Buizza 1997; Leutbecher and Palmer 2008, and references therein). In the present context we relate the $S_{ij}^{A(\overline{mm})}$ to the spatial ensemble spread and the $S_{ij}^{A(\overline{mo})}$ to the spatial ensemble skill (skillful scales). Here, and for the remainder of this paper, the comparison of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ will be referred to as the spatial spread-skill relationship.

The spatial spread-skill relationship defined above differs in several key ways from the traditional spread-skill measures of ensemble standard deviation and RMSE as used, for example, by Buizza *et al.* (2005); Kong *et al.* (2007); Bouttier *et al.* (2012); Baker *et al.* (2014). In particular, the RMSE compares the ensemble mean to observations, and hence a minimum possible RMSE of zero can be obtained when the observations equal the ensemble mean. In contrast, the $S_{ij}^{A(\overline{mo})}$ compares the observations directly to each ensemble member. For any situation where the ensemble members differ spatially or in magnitude, the $S_{ij}^{A(\overline{mm})}$ will be non-zero. Hence the

minimum $S_{ij}^{A(\overline{mo})}$ will also be non-zero: this is limited by, and related to, the ensemble spread. Note that this is a general feature of spatial analysis and would also be true of other spatial comparison methods; for example, any method which considers the differences in location of forecast features between observations and individual ensemble members.

4. Idealised experiment

To investigate the properties of the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, an idealised experiment was performed. This allows links between the precipitation distribution and agreement scales to be explored for configurations with known properties. Additionally, by using this simple setup, the method's interpretation could be tested using many runs and configurations. Synthetic ensembles were created that were defined to be either spatially well spread, over spread or under spread, allowing the validity of the spatial spread-skill comparison between $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ to be tested.

4.1. Overall setup

To mirror the analysis of the real cases (to be discussed in Section 5) a domain of 193 by 242 grid points was created. Initially all points in the domain were set to zero, representing zero rain everywhere. To simulate precipitation, approximately circular areas ('rain blobs') within the domain were each set to an arbitrary value (> 0). To represent an ensemble of N forecasts at a given time t_1, the centres of N rain blobs were randomly positioned within a square 'rain area' of side L and lower left corner at point (X, Y). Similarly, to represent the radar observation at t_1, one rain blob was positioned within a square 'observation rain area' of length L_o and lower left corner at point (X_o, Y_o). To represent different draws from the ensemble distribution, or equivalently different forecasts of the event, multiple random draws were made for the ensemble and radar positions.

The standard ensemble configuration considered 13 different draws of a 12 member ensemble in order to mirror the number of times and members considered for the real cases (analysed in Section 5). In the standard setup the ensemble member and radar rain areas were set to $L = L_o = 50$ grid points positioned towards the centre of the domain with the lower left corner at $(60, 60)$. The standard rain blob radius was 8 grid lengths. An example of the ensemble member positions, from one random draw of the standard configuration, is given in Fig. 3.

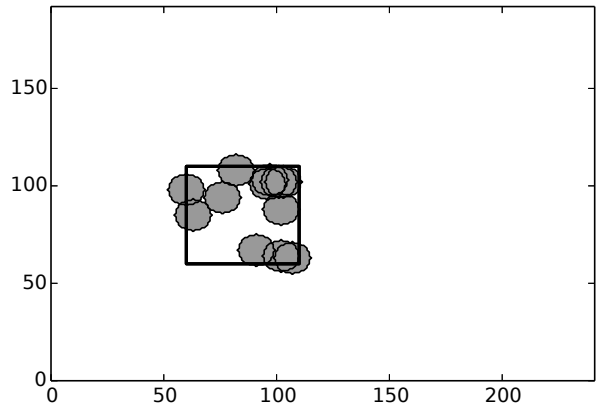

Figure 3. Example of ensemble member rain blobs (grey circles, one per member) positioned within a rain area (black square) for one random draw of the standard idealised setup.

4.2. Agreement scale maps

An example map of $S_{ij}^{A(\overline{mm})}$ from the standard ensemble configuration at one time is shown in Fig. 4. Near the centre of the rain area the scales are smallest, around 10 grid points. This scale is representative of the average separation of the rain blobs. Moving away from the precipitation area the $S_{ij}^{A(\overline{mm})}$ increases as the distance from the rain area dominates the agreement scales. This is an important feature of the $S_{ij}^{A(\overline{mm})}$: outside the rain area the scales are increasingly representative of the distance from the precipitation. This makes sense when considering the $S_{ij}^{A(\overline{mm})}$ to be the scales over which the precipitation fields should be evaluated: we are spatially comparing the precipitation (not the dry regions). We should emphasise that, as discussed in Section 3.1, the large values of $S_{ij}^{A(\overline{mm})}$ obtained at locations far from precipitation do not indicate a poor forecast (forecast quality can be measured through comparing $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$; Sections 3.4 and 4.4). For example, in the case of no rain anywhere in the domain for both forecast and observations we obtain $S_{ij}^{A(\overline{mm})} = S_{ij}^{A(\overline{mo})} = S_{\text{lim}}$ at every point in the domain indicating a perfect spatial spread-skill relationship.

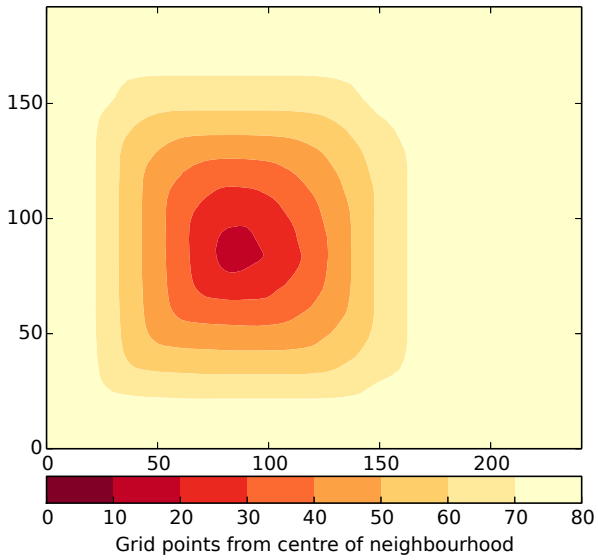

Figure 4. $S_{ij}^{A(\overline{mm})}$ from the idealised experiment standard configuration at one time. All points in the idealised domain are included.

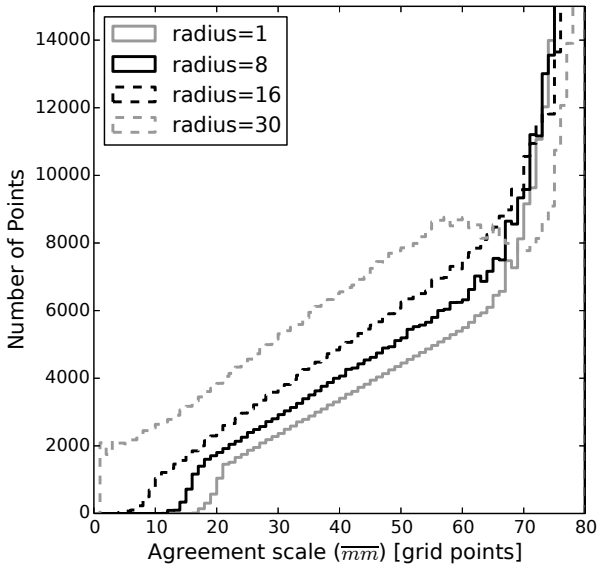

Figure 5. Histogram of $S_{ij}^{A(\overline{mm})}$ for all points in the domain. Idealised experiments are shown with different blob radii (r): grey solid line, $r = 1$; black solid line, $r = 8$ (standard radius); black dashed line, $r = 16$; grey dashed line, $r = 30$. Other parameters were unchanged from the standard configuration.

4.3. Different configurations

Maps of agreement scales are useful for understanding spatial predictability differences across the domain. To compare different configurations, histograms of all points from the $S_{ij}^{A(\overline{mm})}$ maps are considered. One important difference to investigate is the effect of considering different blob radii. If the $S_{ij}^{A(\overline{mm})}$ are behaving as expected, then larger/smaller rain blobs should have more/fewer locations with small $S_{ij}^{A(\overline{mm})}$ as they represent situations that are more/less spatially predictable. The histogram for configurations with different rain blob radii is given in Fig. 5, the other parameters were unchanged from the standard configuration. From Fig. 5 it can be seen that the $S_{ij}^{A(\overline{mm})}$ are behaving as expected: the experiment with a radius of 30 grid points has a minimum spatial scale 18 times smaller than that seen for the experiment with a radius of 1 (a single point). The experiments with larger radii have more points at all scales below 65. Above 65 this behaviour changes and the experiments with smaller radii have more points. Note that, as all experiments have the same total number of points, those experiments with more points at small scales must have fewer points at the largest scales: the fact that this crossover happens around 65 is due to the relative sizes of the rain area and the domain.

Experiments with varying numbers of ensemble members (from 4–20 members) gave very similar $S_{ij}^{A(\overline{mm})}$. This suggests that, at least for this simple idealised setup in which the ensemble spread is predefined, the $S_{ij}^{A(\overline{mm})}$ are not overly sensitive to the number of ensemble members. The investigation of the effect of ensemble size for real case studies would require the consideration of a large number of cases and weather regimes and is beyond the scope of this paper.

4.4. Different spatial spread-skill relationships

In this subsection the relationship between $S_{ij}^{A(\overline{mm})}$ (representing the believable scales, a measure of spatial ensemble spread) and $S_{ij}^{A(\overline{mo})}$ (representing the spatial ensemble skill) is illustrated using the idealised setup. If these measures are to provide useful information, they must differentiate between ensembles that are spatially well spread, over spread, and under spread. Here we use a general meaning of the term "under spread": ensembles are labelled under spread when they fail to capture the observed event. Hence, ensembles are under spread when there is not enough variety in the ensemble member forecasts, and also when all ensemble members forecast precipitation in the wrong place. This is consistent with the use of the term in the traditional spread-skill relationship (i.e. when comparing the ensemble standard deviation with the RMSE of the ensemble mean when comparing with observations). Of course, all members forecasting precipitation at the wrong location, and the lack of variety in ensemble member solutions, are

Table 1. Idealised ensemble settings for ensembles with different spread-skill relationships.

Spread-skill	L	L_o	X, Y	X_o, Y_o
Well spread ('close')	50	50	60,60	60,60
Over spread ('over')	50	10	60,60	80,80
Under spread ('under')	50	90	60,60	40,40
Displaced precipitation ('Miss')	50	50	60,60	110,60

two very different sources of poor ensemble performance. Although it will highlight that the error is there, the spatial spread-skill relationship obtained by comparing $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ cannot distinguish between these two possible error mechanisms.

The $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are compared for idealised ensembles with known spatial spread-skill properties: well spread, over spread, under spread due to not enough variation between members, and under spread due to wrongly-located precipitation. To generate a spatially well spread ensemble both members and radar were selected from the same area (i.e. $L = L_o$ and $(X, Y) = (X_o, Y_o)$). To generate an over/under spread ensemble the radar rain area was defined to be smaller/larger than the member rain area. An additional case, where the ensemble was under spread due to a spatial displacement between the ensemble and observations was also considered with $L = L_o$ but $(X, Y) \neq (X_o, Y_o)$. The radar and member rain areas for these different ensemble configurations are shown in Fig. 6, and the settings for these idealised setups are given in Table 1.

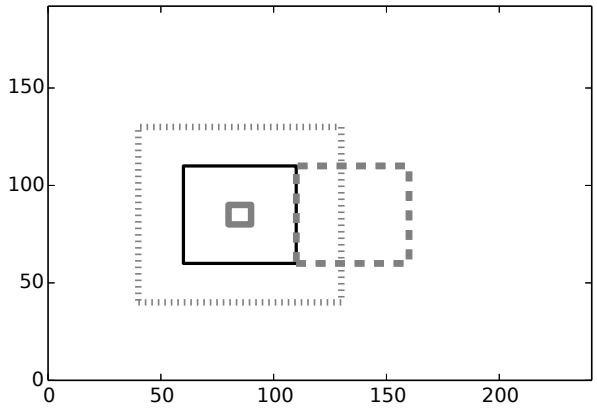

Figure 6. Positions of radar rain areas for cases with different spread-skill: well spread (black solid), over spread (grey solid), under spread (grey dotted) and under spread due to misplaced precipitation (grey dashed). For all experiments the ensemble members were selected from the black square.

4.5. Methods of comparing $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$

Although histograms of all points from the $S_{ij}^{A(\overline{mm})}$ maps allow the differences between configurations to be visualised (e.g Fig. 5), in order to fully compare $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ it is necessary to choose a method that enables a scale selective comparison, whilst preserving the location-dependent point-to-point relationship between the S_{ij}^{A} fields. One way to do this would be a simple scatter plot of the $S_{ij}^{A(\overline{mo})}$ against the $S_{ij}^{A(\overline{mm})}$. However, this would give a noisy result. To enable simpler comparison, we bin the scatter plot based on the $S_{ij}^{A(\overline{mm})}$ value.

First, a bin is selected, say from 0 to 9 grid points. The points for which the $S_{ij}^{A(\overline{mm})}$ value lies within the bin are then considered and the mean $S_{ij}^{A(\overline{mm})}$ over such points is calculated. By definition this mean value will lie within the selected bin. Next, the $S_{ij}^{A(\overline{mo})}$ mean value *over the same spatial points* is considered. If the ensemble is well spread this will equal the $S_{ij}^{A(\overline{mm})}$ mean value; if the ensemble is over/under spread then the $S_{ij}^{A(\overline{mo})}$ mean value will be smaller/larger than that of the $S_{ij}^{A(\overline{mm})}$. Hence, on the binned scatter plot, a well spread ensemble should lie on the diagonal, and under/over spread ensembles should lie above/below the diagonal.

We have checked these interpretations using various idealised ensembles with pre-defined spread-skill characteristics, such as those specified in Fig. 6 and Table 1. For example, binned scatter plots are shown in Fig. 7 for a bin size of 10 grid points. The 'close' experiment is shown in black and lies on the diagonal as expected: the average of all $S_{ij}^{A(\overline{mm})}$ points within a given bin is equal to the $S_{ij}^{A(\overline{mo})}$ averaged over the same points. The two experiments with under spread ensembles ('under' and 'miss') both lie above the diagonal, with $S_{ij}^{A(\overline{mo})}$ larger than $S_{ij}^{A(\overline{mm})}$ for a given bin. Similarly, as expected, the over spread case lies below the diagonal with $S_{ij}^{A(\overline{mo})}$ smaller than $S_{ij}^{A(\overline{mm})}$ for a given bin. This confirms that differences between $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are reflecting the different ensembles and provide useful information about the spatial spread-skill. Notice that, for the 'close' experiment, there is some departure from the diagonal at scales 10-18: the

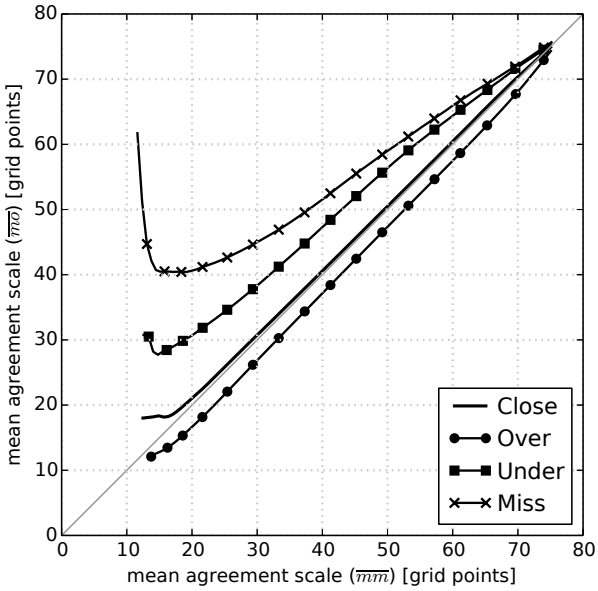

Figure 7. Binned scatter plot for idealised ensembles with different spread-skill characteristics: over spread (black with circles), well spread (black with no markers), under spread (black with squares), and, missed precipitation (black with crosses). A bin size of 10 grid points was used. Note that line markers are for illustration only and do not represent specific plotted points.

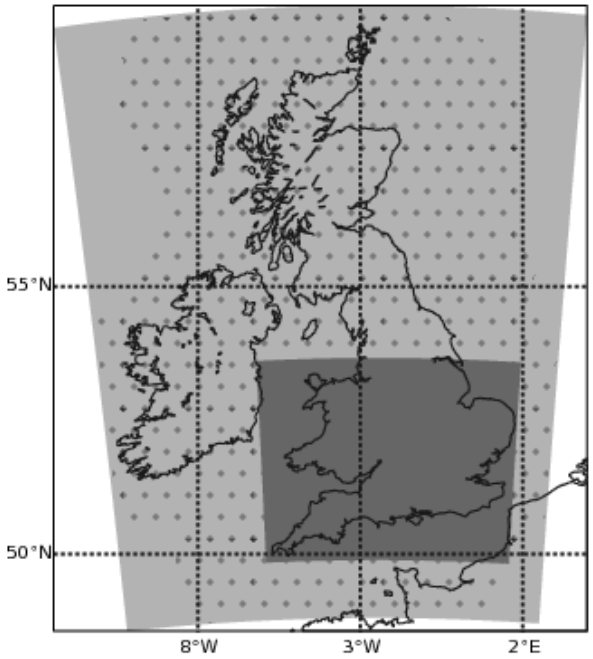

Figure 8. Domains for the UK 2.2 km model (light grey), the area of radar coverage (dotted), and the region used for forecast evaluation (dark grey).

average $S_{ij}^{A(\overline{mo})}$ over these points is larger than the average $S_{ij}^{A(\overline{mm})}$. This is due to our simple method of defining the idealised ensemble: randomly selecting a modest number of ensemble members within a given area results in a non-uniform member distribution over that area, which would for an ideal ensemble represent an uneven radar spatial probability distribution across the area. However, the radar distribution was assumed to be uniform. This interpretation was confirmed by experiments in which the rain blobs for the ensemble members were positioned not randomly but at fixed, uniformly-distributed, locations.

The results from this section show that the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ can successfully be used to determine the spatial spread-skill characteristics of an ensemble system, and that the binned scatter plot provides a particularly clear method of viewing these results. In section 5 these methods will be applied to real convective cases.

5. Convective cases from MOGREPS-UK

5.1. Model set up

In this study forecasts are evaluated from the Met Office Global and Regional Ensemble Prediction System UK ensemble, MOGREPS-UK, which has been run operationally since June 2013 (Mylne 2013; Golding *et al.* 2014). MOGREPS-UK runs with 12 members and a constant resolution 2.2 km grid over the UK. MOGREPS-UK is one way nested inside the global ensemble MOGREPS-G and, to reduce the jump in resolution between the two models, the edges of the MOGREPS-UK grid are stretched up to 4 km. The constant-resolution part of the MOGREPS-UK domain is shown in light grey in Fig. 8. To speed up processing times the analysis was performed over a smaller domain covering south/central England and Wales. This domain is shown in dark grey in Fig. 8. This study used radar-derived rain rates from the Radarnet system which provides a rain rate composite at 1 km resolution and includes calibration against rain gauge data (Golding 1998; Harrison *et al.* 2000, 2012). The region of radar coverage is shown by the dotted area in Fig. 8 and fully includes the analysis region. To make a fair comparison with the model, the Radarnet radar-derived rain rates were interpolated onto the 2.2 km resolution MOGREPS-UK grid before any comparisons were carried out.

At the time of writing, MOGREPS-UK members are downscaled inside MOGREPS-G perturbations, generated using an ensemble transform Kalman filter (ETKF), and then added to the Met Office 4D-Var analysis as described by Bowler *et al.* (2008, 2009). This perturbation strategy includes

a stochastic kinetic energy backscatter scheme and localisation in the ETKF. Model error is addressed in MOGREPS-G using the random parameters scheme to account for sub-grid process uncertainty. MOGREPS-G is run with 23 perturbed members and an unperturbed control. The MOGREPS-UK ensemble is started 3 hours after MOGREPS-G with initial and boundary conditions taken directly from the control and 11 perturbed members. The 0300 UTC start time was used for all cases presented in this paper and allows the model time to spin up before the times of interest for each case.

The version of the Met Office United Model (MetUM) operational in summer 2013, version 8.2, was used for this work. This version of the MetUM has a non-hydrostatic dynamical core with semi-Lagrangian advection (Davies *et al.* 2005). A comprehensive set of parametrizations are used in the MetUM including: surface exchange (Essery *et al.* 2001), boundary layer mixing (Lock *et al.* 2000), radiation (Edwards and Slingo 1996) and mixed phase cloud microphysics based on Wilson and Ballard (1999).

5.2. Introduction to cases

Six cases were selected from summer 2013 during the period of the Convective Orographic Precipitation Experiment (COPE Blyth *et al.* 2015; Leon *et al.* 2015). The COPE field campaign concentrated on the English southwest peninsula (SW peninsula, 50.0°N, 5.5°W – 51.5°N, 2.0°W) to investigate the processes controlling precipitation intensity. Five of the cases used here are from the COPE IOPs (intensive observing periods); the exception being Case A. The cases were subjectively selected to represent a variety of convective situations and differing predictability.

To illustrate the meteorology for each case, radar-derived instantaneous rain rates are shown in Fig. 9 for all cases, at selected times when convection occurred. The first three cases (A-C; Fig. 9a-c) exhibit deep convection associated with various features of the large scale flow. This is common for convection over the UK, which can develop within a variety of flow regimes (e.g. Browning and Roberts 1994, 1995; Morcrette *et al.* 2007; Russell *et al.* 2008, 2009). The remaining three case studies (Cases D-F; Fig. 9d-f) showed organisation of precipitation along the SW peninsula. This is a common meteorological situation for this region, particularly in southwesterly flow, and happens as a result of topographically induced convergence (e.g. Burt 2005; Golding *et al.* 2005; Leoncini *et al.* 2013; Warren *et al.* 2014).

Case A (17/07, Fig. 9a) occurred during an extended period of high pressure over the UK and exhibits a line of localised thunderstorms from 1600–1900 UTC. In Case B (23/07) convection developed along two troughs that were positioned over the UK, associated with a mature cyclone over the Atlantic. Two bands of precipitation occurred throughout the morning as shown in Fig. 9b at 0600 UTC. Case study C (27/07) was affected by a Mesoscale Convective System (MCS) which moved north from France throughout the day. The widespread precipitation associated with the MCS is shown at 2100 UTC in Fig. 9c. Although MCSs only occur twice a year on average in the UK (Gray and Marshall 1998; Lewis and Gray 2010), they are often high-impact events and so an important situation for assessing spatial uncertainty. In Case D (29/07) scattered convection was seen over England from 0800 UTC onwards with some organisation along the SW and South Wales (51.5°N, 5.0°W – 53.0°N, 2.0°W) peninsulas as shown in the radar data at 1500 UTC (Fig. 9d). This indicates that peninsula convergence played a role in convective initiation for this case but was not the dominating mechanism. Case E (02/08) featured a line of precipitation along the north coast of the SW peninsula extending north through Wales as shown in Fig. 9e at 1800 UTC. This precipitation was aligned with a cold front which extended southwest to northeast across the UK, suggesting that both large-scale forcing and peninsula convergence were important mechanisms for this case. Early in Case E convection was also seen further east ahead of the cold front. These deep convective storms resulted in heavy precipitation and many lightning strikes. In contrast to the other peninsula convergence cases (Cases D and E), convection in Case F (03/08, Fig. 9f) was predominantly linked to a convergence line along the SW peninsula.

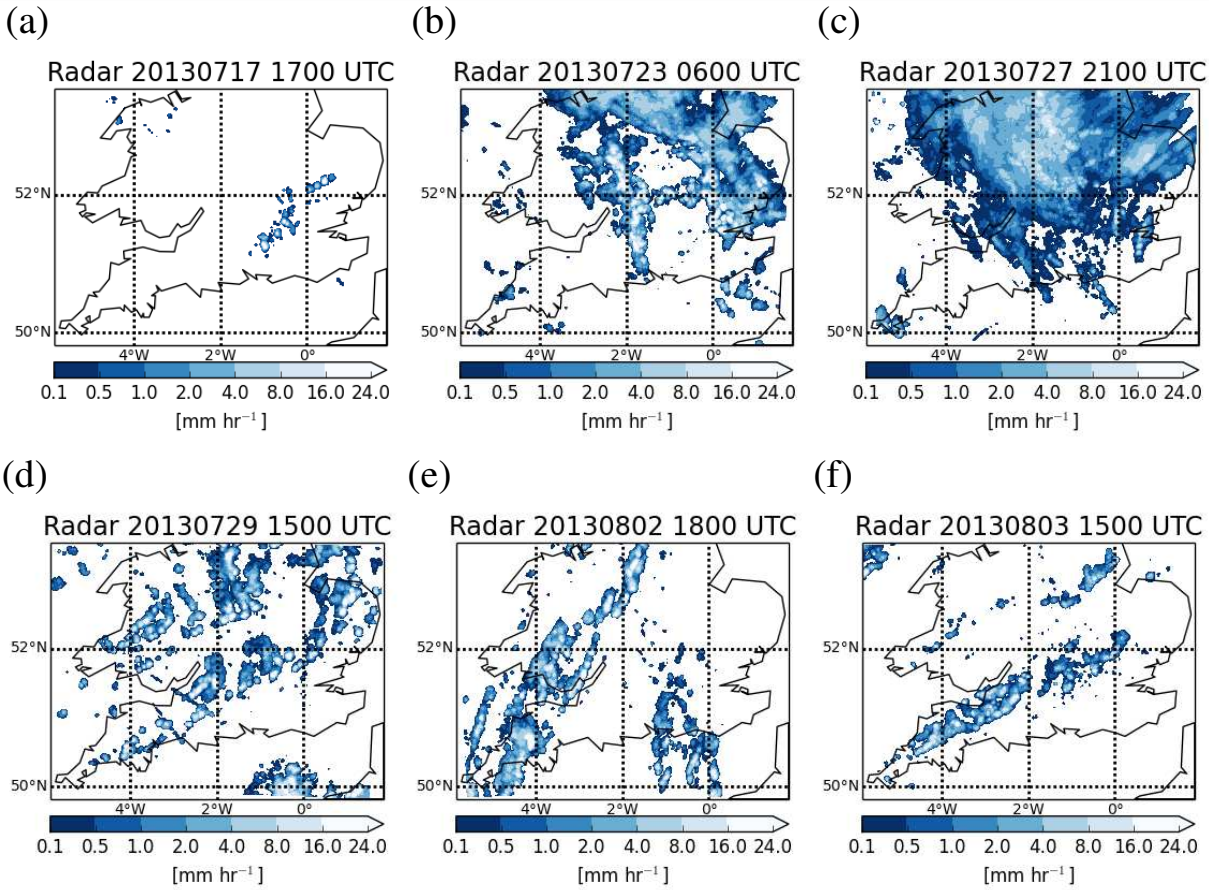

Figure 9. Instantaneous rain rates for the cases considered. For each case a time is shown that illustrates the main features of the rain on that day. (a) Case A at 1700 UTC, (b) Case B at 0600 UTC, (c) Case C at 2100 UTC, (d) Case D at 1500 UTC, (e) Case E at 1800 UTC and (f) Case F at 1500 UTC.

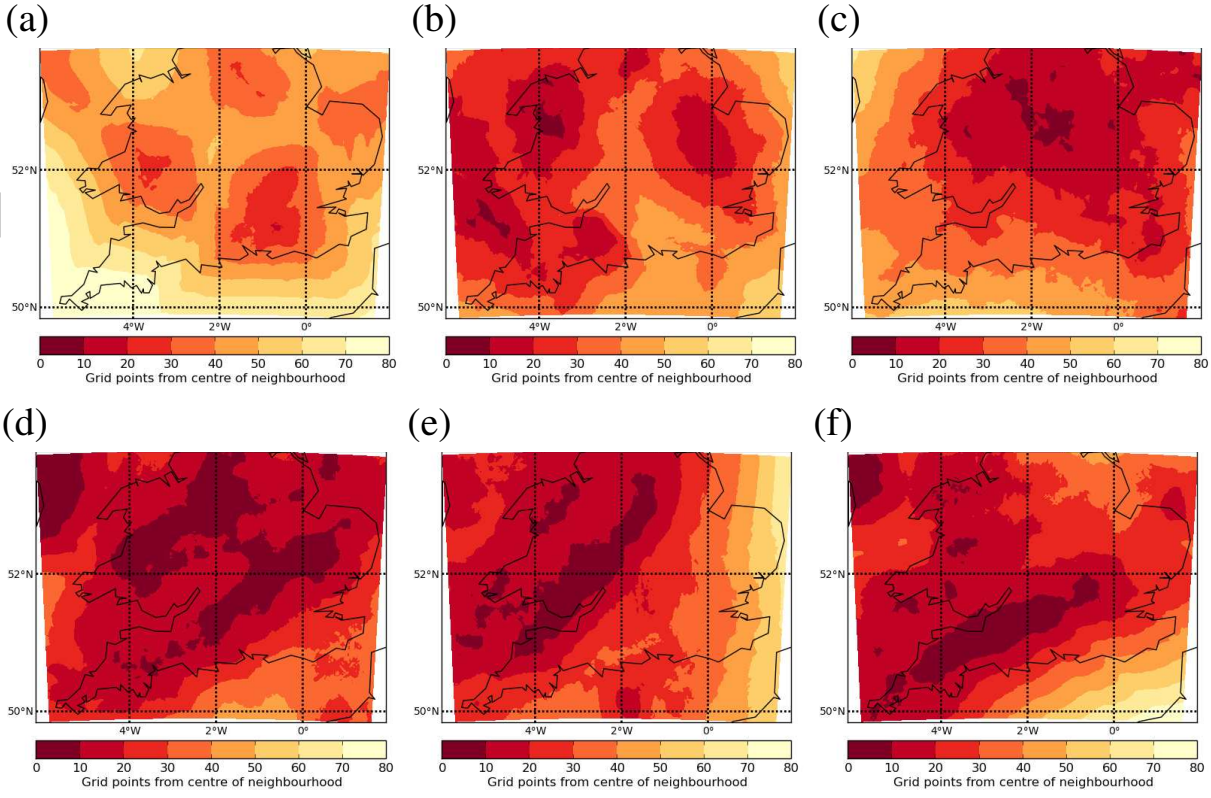

Figure 10. As Fig. 9, but showing the corresponding maps of member-member agreement scales $S_{ij}^{A(\overline{mm})}$ from the MOGREPS-UK ensemble.

5.3. *Results: Spatial maps*

To investigate the spatial agreement for these six cases we examine the spatial difference between ensemble members. To do this, $S_{ij}^{A(\overline{mm})}$ were calculated hourly for each case using instantaneous rain rates. Example $S_{ij}^{A(\overline{mm})}$ maps are given in

Fig. 10 at the same times as shown in Fig. 9 for the radar rain rate data. Comparison of Figs. 9 and 10 shows that in general the smaller $S_{ij}^{A(\overline{mm})}$ tend to be linked to areas of precipitation. This indicates that the method is behaving as expected: in areas of precipitation, the spatial differences in the placement of precipitation between members are smallest, giving smaller $S_{ij}^{A(\overline{mm})}$. There are additional aspects of the pattern of $S_{ij}^{A(\overline{mm})}$ that are highly case dependent.

For the cases showing peninsula convergence (Cases D, E and F, Fig. 10d-f) small scales are seen along the peninsula where the precipitation is highly spatially predictable. In this meteorological situation, the location of precipitation is tied to the local topography providing a constraint on the possible precipitation locations. Hence, higher spatial predictability and higher spatial agreement are expected for these cases. In contrast, in Case A (Fig. 10a) the precipitation has low spatial agreement with a minimum $S_{ij}^{A(\overline{mm})}$ of around 20 grid points. This is due to large spatial differences in the placement of precipitation between ensemble members for this case, possibly caused by subtle differences in the larger scale forcing: small variations in the large scale led to large variations in triggering locations for convection. For this case the model consistently predicts localised thunderstorms, but their location is uncertain. The $S_{ij}^{A(\overline{mm})}$ allows this valuable information to be easily extracted. The same conclusions can, of course, be drawn from close inspection of the individual member rain rate fields but that is a more cumbersome and qualitative process.

A further example of precipitation with lower spatial agreement is seen in Case E to the east of the domain (50.0°N, 1.5°W – 52.0°N, 1.0°E; Fig. 10e). Here the observations show convective storms moving north from France (e.g Fig. 9e) throughout the day. Similar behaviour is captured by a small number of ensemble members (the particular members, and the number of members is time dependent). In this region the $S_{ij}^{A(\overline{mm})}$ vary from 30 to 60 grid points, suggesting that precipitation could occur within a broad region. This information, in conjunction with a single reference ensemble member, or a deterministic forecast, would help assess the local spatial uncertainty for heavy rain.

The two cases with the most widespread precipitation, Case B and Case C (Figs 10b and 10c respectively), both have $S_{ij}^{A(\overline{mm})}$ of less than 20 grid points over the regions where precipitation occurred. For these cases the spatial uncertainty in the location of precipitation was much smaller than the size of the precipitation area, and hence there was a high degree of agreement and overlap between the individual member forecasts.

The results from Fig. 10 provide a summary of the spatial uncertainty within the ensemble for each case, in one single picture. This is useful for model interpretation and evaluation, and would be valuable in an operational forecasting context. However, it is also important to consider whether these scales are representative of the true spatial uncertainty for each case. To assess the 'spatial spread-skill relationship' the $S_{ij}^{A(\overline{mo})}$ was also calculated hourly for all cases as described in Sections 3.2 and 4. Example $S_{ij}^{A(\overline{mo})}$ maps for Case A at 1700 UTC and Case D at 1500 UTC are given in Fig. 11a and 11b respectively. Comparing Fig. 11a and 11b with Fig. 10a and 10d respectively, the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ look qualitatively similar. There are however some differences. In particular, the $S_{ij}^{A(\overline{mo})}$ have larger areas of both the smallest and largest scales, and are more noisy. These differences will be quantified in the following subsections. It is also interesting to compare the $S_{ij}^{A(\overline{mo})}$ with the radar observations for these cases, shown in Fig. 9a and Fig. 9d respectively. Similarly to the $S_{ij}^{A(\overline{mm})}$, the smallest $S_{ij}^{A(\overline{mo})}$ are seen in areas of precipitation, confirming that the method is behaving as expected. The minimum $S_{ij}^{A(\overline{mo})}$ values for Case A seen around 52.3°N, 3.0°W and 53.8°N, 1.5°W, are associated with low magnitude precipitation which does not show in Fig. 9a.

5.4. Results: Domain average

To summarise the overall spatial agreement scales, spatial spread-skill relationship, and time evolution of spatial agreement, we now consider the domain average $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$. The domain average value represents the scale that we would use to characterise the forecasts at all points in the domain if a single scale had to be chosen. Thus it is necessary to include all points in the average (i.e. the scales at points

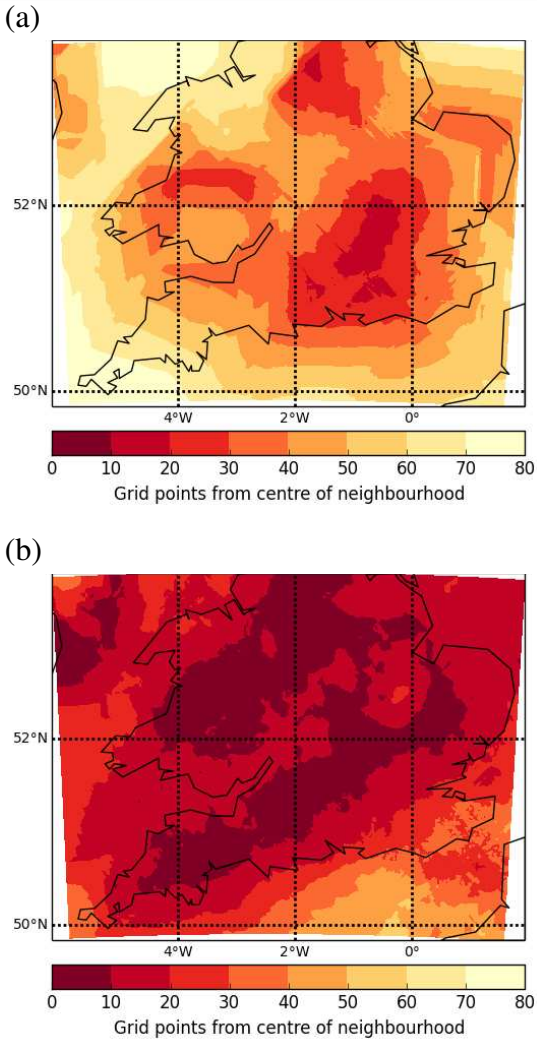

Figure 11. Member-radar agreement scales $S_{ij}^{A(\overline{mo})}$ for the (a) Case A at 1700 UTC and (b) Case D at 1500 UTC.

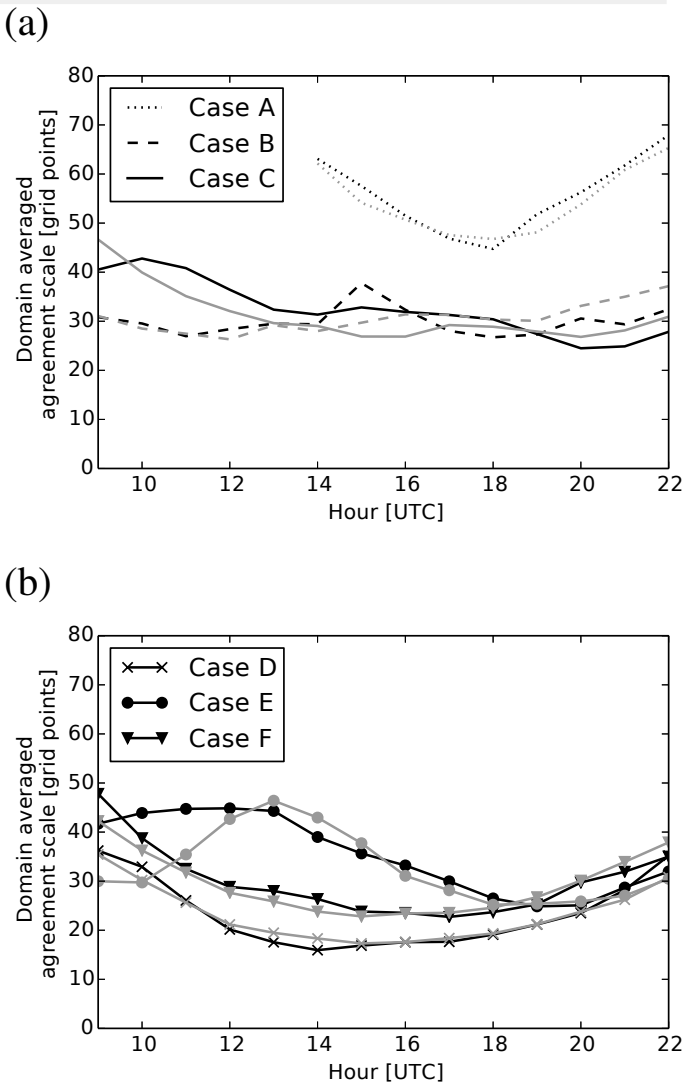

Figure 12. Time series of domain averaged $S_{ij}^{A(\overline{mm})}$ (grey) and $S_{ij}^{A(\overline{mo})}$ (black). (a) Case A, dotted; Case B, dashed; Case C, solid; (b) Case D, solid with crosses; Case E, solid with circles; Case F, solid with triangles.

where precipitation is forecast/observed, and also at points away from the precipitation).

The domain average $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are shown in Fig. 12a for Cases A-C (top row of Figs 9 and 10) and Fig. 12b for Cases D-F, (bottom row of Figs 9 and 10). The cases are shown from 0900 UTC to 2200 UTC (forecast lead time 06 hrs to 19 hrs), a period which covers the convective events of interest. Case A is only shown from 1400 UTC onwards when convection occurred: before this time there was no simulated precipitation over the domain and, additionally, problems with the radar data.

Case A has the largest average spatial agreement scales with a minimum domain average $S_{ij}^{A(\overline{mm})}$ of around 50 grid points. This agrees with the qualitative analysis of the agreement scale maps (Figs 10a and 11a). The model has captured the spatial uncertainty well on this day: the domain averaged $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are similar with the black and grey lines lying close to each other. This is perhaps unexpected given the large-scale uncertainties seen in the location of precipitation that day (shown by minimum $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ of around 20 grid points in Figs 10a and 11a).

The two cases with widespread precipitation, Cases B and C, both have a domain-averaged $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ in the range of 30-40 grid points. These cases have higher domain-averaged spatial agreement than Case A, due to higher spatial predictability (seen from the agreement scale maps for these cases; Figs 10b and 10c) and larger areas of precipitation. For Case C the domain averaged $S_{ij}^{A(\overline{mo})}$ are larger than the domain averaged $S_{ij}^{A(\overline{mm})}$ from 1000 UTC to 1800 UTC: the ensemble members are closer to each other than they are to observations and the ensemble forecast of the MCS is spatially under spread.

Out of all the cases, Case D has the smallest domain-averaged spatial agreement scales, with the domain average $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ dropping below 20 grid points from 1200 UTC to 1800 UTC. Again, this agrees with the qualitative analysis (Figs 10d and 11b). The other cases with peninsula convergence (Cases E and F) behave similarly to the Case D with domain average agreement scales below 30 grid points, and similar values of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$. In Case E, before 1500 UTC, the domain-average of the agreement scales is dominated by the spatially unpredictable precipitation to the east of the domain, and larger $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are seen. At these times the ensemble is, at least in a domain-averaged sense, under spread with a difference of over 10 grid points between the domain averaged $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$.

5.5. *Results: location-dependent comparison*

Although the domain average spatial agreement is a guide to the overall spatial predictability for a given case, it is more helpful to compare the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ in a scale selective manner that preserves the local information. For this purpose the binned scatter plot is employed, as applied to the ideal ensemble in Section 4. Results are shown for all cases except the Case A at 0900 UTC and 1300 UTC (Fig. 13a,b) and for all cases at 1700 UTC (Fig. 13c). A bin size of 10 grid points has been used for these plots. This bin size was chosen as it allows low agreement scales to be represented, whilst still considering enough grid points in each bin to give robust results. Similar conclusions are obtained from bin sizes in the range of 4 to 20 grid points, and are not presented here.

At 0900 UTC and 1300 UTC (Fig. 13a and b respectively) the spatial spread-skill relationship is highly case and time dependent. This can be related to the different physical processes occurring for each case and time, and also to biases between the forecast and observations. For Case B (dashed line) $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ were similar at 0900 UTC and the ensemble captured the spatial variability well. At 1300 UTC the small scale uncertainty was still captured well for this case, but larger scales ($S_{ij}^{A(\overline{mm})}$ above 40 grid points) showed $S_{ij}^{A(\overline{mo})}$ greater than $S_{ij}^{A(\overline{mm})}$. This is related to 8 out of the 12 ensemble members showing precipitation in south-central England at this time, although there was little observed

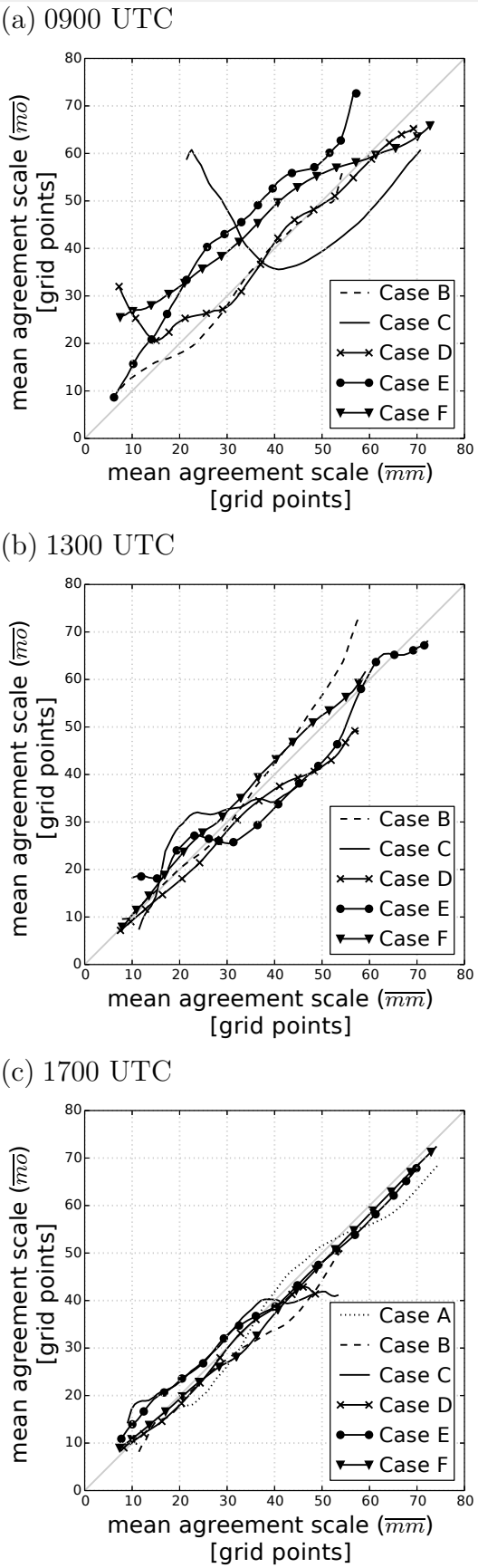

Figure 13. Binned scatter plot for the Cases B-F at (a) 0900 UTC, (b) 1300 UTC, and for all cases at (c) 1700 UTC. Individual traces are plotted for each case at the specified time: Case A, dotted; Case B, dashed; Case C, solid; Case D, solid with crosses; Case E, solid with circles; and, Case F, solid with triangles. Note that line markers are for illustration only and do not represent specific plotted points.

precipitation in this region. For Case C (solid line) a timing error at 0900 UTC results in all members predicting an MCS over England whereas, in reality, the MCS was still over the Channel. At this time the ensemble is over confident in the location of the MCS. By 1300 UTC the model MCS was still seen over southern England and the real MCS had 'caught up' due to a faster propagation speed. Hence, for Case C at 1300 UTC, there was a large overlap between the predicted and observed precipitation fields and an improved spatial spread-skill relationship. Note also that, at this time, both $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are less than 50 grid points at all points in the domain (there is no trace on the binned scatter plot above 50 grid points): there is high spatial agreement between the ensemble and observations at all points, because the rain is so widespread.

In Case D (solid line with crosses) the ensemble was spatially well spread at 0900 UTC for scales above 30 grid points, but under spread below these scales. This was due to differences in the placement of precipitation over north Wales resulting in the smallest values of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ occurring at different locations. At 1300 UTC the ensemble members are closer, on average, to observations than to each other and $S_{ij}^{A(\overline{mo})}$ less than $S_{ij}^{A(\overline{mm})}$. Case E (solid line with circles) is spatially under spread at 0900 UTC but spatially over spread by 1300 UTC. The spatial predictability for this case varied throughout the day as discussed for the domain-averaged values (Fig. 12). These results agree with those from the domain average: the ensemble is under spread at 0900 UTC suggesting that the uncertainty in the convection moving in from France was difficult for the ensemble to capture. This is possibly due to the convection initiating outside the MOGREPS-UK domain and so relying on the global model's convective parametrisation. Later in the day, when precipitation was mainly in an organised spatially-predictable line over the SW peninsula (as discussed with reference to Figs. 9 and 10), $S_{ij}^{A(\overline{mo})}$ values were smaller than $S_{ij}^{A(\overline{mm})}$: the radar fell within the ensemble distribution. This could indicate that the ensemble was too pessimistic about spatial accuracy at this time. In Case F (solid line with triangles) the ensemble was spatially under spread at both 0900 UTC and 1300 UTC with $S_{ij}^{A(\overline{mo})}$ greater than $S_{ij}^{A(\overline{mm})}$. This is particularly noticeable at the earlier time, and is related to the ensemble members producing showers in a different area of the domain to where they were seen in reality. Later in the day both model and observations produced precipitation associated with convergence lines from the SW and Welsh peninsulas and the spread-skill relationship improved.

At 1700 UTC (Fig. 13c) the case-to-case differences in spread-skill, seen at 0900 UTC and 1300 UTC, are much reduced. By this time the values of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are reflecting the fact that convection has developed and evolved over the course of the day. Earlier in the day initiation errors degrade the spatial spread-skill. However, as the precipitation remains for a number of hours, once initiation has occurred in both model and observations, there is a large degree of overlap. This highlights the link between spatial and temporal errors: a timing error will also result in a spatial error between fields. Note, however, that this result may also be linked to the choice of only a limited number of convective cases, in which convection was reasonably captured by the model.

6. Discussion and conclusions

This paper has presented a new spatial method for the characterisation and evaluation of the local spatial agreement between members in convective-scale ensembles. Based on a neighbourhood approach, the scales over which ensemble members reach a specified level of agreement ($S_{ij}^{A(\overline{mm})}$) were calculated, at each grid point in the domain, to give a measure of location-dependent believable scales for an ensemble forecast, i.e. the scales at which the ensemble members become sufficiently similar so that the forecast forms useful, trustworthy guidance. A method was also presented to verify the $S_{ij}^{A(\overline{mm})}$ by comparing with the scales at which ensemble members reached a required level of agreement with radar observations, denoted $S_{ij}^{A(\overline{mo})}$. The interpretation assumes that differences between fields over this neighbourhood represent the spatial uncertainties (or errors) and local biases in the forecast. This assumption is good for small neighbourhoods, but becomes less valid as the neighbourhood size increases: events far apart in two different

forecasts become more likely to represent different events rather than large displacement errors. This should be kept in mind when interpreting the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$.

To calculate the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, ensemble members were compared, either pairwise against each other, or against observations. At each grid point in the domain, the agreement scale between the fields was defined as the minimum neighbourhood size over which the fields were deemed to be acceptably similar. To decide whether the forecasts were acceptably similar, a criterion was defined based on two predetermined parameters. The first, α, controls the acceptable fractional difference between the fields, and the second, S_{lim}, is a fixed maximum scale at which the forecasts are always deemed to be sufficiently similar. For the examples presented in this paper the values $\alpha = 0.5$ and $S_{\text{lim}} = 80$ were used: other values could also be chosen to give a more, or less, stringent criterion. Thus, the required level of agreement is not fixed, and may be determined from the user's requirements.

In formulating the agreement scales, $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, the aim was to present a simple, generally applicable, method of quantifying forecast differences. These measures are not designed to distinguish between temporal, amplitude, and structural components of forecast uncertainty and error. Other methods (such as those discussed in Gilleland *et al.* (2009)) do attempt to provide such information for the verification of high resolution deterministic forecasts, and could be developed for application to ensemble systems. This information would be complementary to that obtained using the methods presented in this paper.

A simple idealised system was created to investigate the properties of the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$. Each individual ensemble member, and the observations, were represented by a circular blob of rain, randomly positioned within a square region. Using this simple setup, it was shown that the $S_{ij}^{A(\overline{mm})}$ successfully represent spatial differences with larger spatial differences leading to larger $S_{ij}^{A(\overline{mm})}$. The method was found to be robust to changes in the number of ensemble members and to the position of the square rain region within the domain. The idealised ensemble was further used to asses the utility of comparing the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ to investigate the performance of the ensemble forecasts for these cases. This comparison can be related to the traditional spread-skill relationship for ensemble evaluation, with the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ representing the ensemble spread and ensemble skill components respectively. Consistent with this, the comparison of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ was denoted as the the "spatial spread-skill relationship". Through comparing the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ it was possible to differentiate between pre-determined scenarios in which the synthetic precipitation is set up to be either well spread, over spread, or under spread spatially. The spatial spread-skill relationship was visualised through histograms of all agreement scale data, and using binned scatter plots. It was found that binned scatter plots provide a particularly useful method for assessing the spatial spread-skill properties because the location-dependent character of convective-scale uncertainty is respected.

To demonstrate the utility of these techniques as an investigation tool for operational ensemble systems, the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ were calculated for hourly instantaneous rain rates for six convective case studies run with the 2.2 km grid length, 12 member, operational MOGREPS-UK ensemble. These cases were selected to represent UK convection in a variety of regimes including: upper-level and large-scale forcing, topographical convergence and scattered convection. Maps of the $S_{ij}^{A(\overline{mm})}$ depicted the different levels of spatial agreement across the cases which was related to different levels of spatial predictability. For example, cases where precipitation was strongly linked to convergence along the SW peninsula showed high levels of spatial predictability, and high spatial agreement, with local $S_{ij}^{A(\overline{mm})}$ of less than 10 grid points. This high spatial predictability is expected from the topographic influence for these cases. In contrast, other cases, such as Case E, showed that precipitation could also be highly spatially unpredictable. It should be reiterated that, independently, the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ can not be used to measure forecast quality.

Used in conjunction with a single ensemble member, or deterministic forecast, the $S_{ij}^{A(\overline{mm})}$ provide a useful visualisation for forecasting. The rainfall structures themselves can be viewed from an individual model run (perhaps the

control) and the $S_{ij}^{A(\overline{mm})}$ map can be used to view the spatial uncertainty in that rainfall given by the ensemble. This provides a method of quickly assessing the spatial predictability obtained from the ensemble. It gives a more physically meaningful view of ensemble-member differences than using grid point measures, for example, the variance at each grid point.

To demonstrate how the location-dependent agreement scales can be used to diagnose ensemble performance, the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ were compared for the six summer convective cases. Note that the aim was to provide concrete examples of how these techniques can be applied and interpreted, not to provide a statistical verification of the operational ensemble system. It was found that, as well as having different levels of spatial agreement, the different cases showed different spatial spread-skill relationships. Poor spatial spread-skill consistency, measured by larger differences between the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, could be linked to differences between the model and observations, such as a timing error or precipitation incorrectly forecast by the model. For these six convective cases, the spatial spread-skill relationship improved in the afternoon, suggesting that it was the spatial characteristics during precipitation initiation that were most difficult for the model to handle in these instances. Once established, precipitation occurred for a number of hours and the spatial spread-skill improved. Through comparing the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, these features of the ensemble performance were easily identified. This suggests that the agreement scales would provide a valuable diagnostic for verifying the spatial ensemble performance. Future work will conduct such an investigation for the MOGREPS-UK ensemble. Additionally, these methods could be used to assess the impact of changes to the forecasting system, for example the use of stochastic increments to model systematic initiation uncertainties (e.g. Leoncini *et al.* 2010).

This paper has focused on calculating the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ for forecasts of instantaneous rain rates. Rain rates were selected for this study to avoid any temporal smoothing from using precipitation accumulations, and hence to focus on the spatial features. Of course, the methods presented here could also be used to evaluate precipitation accumulations. More generally, although precipitation forecasts are a key application of these methods (due to their high spatial uncertainty and the availability of radar observations for verification), the comparison of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ is equally applicable to other positive meteorological fields where gridded observations are available, for example from satellite imagery or for comparison against analysis. Work by the authors has found that the $S_{ij}^{A(\overline{mm})}$ calculated for other fields, particularly those with variability on small scales such as cloud fraction, positive divergence and humidity, yields useful information. It is also possible to calculate the $S_{ij}^{A(\overline{mm})}$ at different vertical levels in order to probe the vertical structure of horizontal spatial differences. This is the subject of ongoing work.

It should be emphasised that for fields other than precipitation, the link between the agreement scales and spatial predictability may be lost. In particular, for fields which vary on large scales, such as those with large scale gradients, the agreement scales will reflect only the bias between the fields over the area in question. The link between the agreement scales and spatial predictability could be reestablished by converting the field to binary (i.e setting points to one or zero dependent on their position above/below a predefined threshold) before calculating the agreement scales. Using a threshold would remove any bias (or background gradient) and hence only relate the agreement scales to positional differences between the fields, but would also make the agreement scales less general: a threshold must be selected and the bias between fields is no longer considered. Additionally, a value of S_{lim} appropriate to the large scale differences between these fields (i.e. larger than that used for precipitation) would have to be selected.

There are some limitations to this study. In particular it has been assumed here that the radar data is 'truth' and observational errors have been neglected. Although the Radarnet radar data has been quality checked (Golding 1998; Harrison *et al.* 2000, 2012) and the rain rate composite is used operationally in the Met Office nowcasting (Bowler *et al.* 2006) and latent heat nudging assimilation (Simonin *et al.* 2014) systems, there are still likely to be unaccounted-for errors.

Accounting for these errors in the $S_{ij}^{A(\overline{mo})}$ is an important avenue of future investigation.

Despite these limitations there are some important conclusions from this work. A simple method has been demonstrated to calculate the spatial differences between pairs of ensemble members ($S_{ij}^{A(\overline{mm})}$) and also between ensemble members and observations ($S_{ij}^{A(\overline{mo})}$). The method is easily applied to other ensemble systems, and to fields other than precipitation (e.g. satellite imagery). For idealised simulations, and six case studies with an operational ensemble system, these measures were found to give a location-dependent and physically meaningful summary of information from the ensemble. This suggests that these measures could be used to better understand forecasting systems, and hence to highlight areas needing improvement. Additionally, these methods could be used in a forecasting context to visualise the spatial uncertainty forecast by the ensemble.

Acknowledgement

S.Dey is supported by a NERC PhD studentship with CASE support from the Met Office. S.Migliorini acknowledges support from the NERC National Centre for Earth Observation. We also acknowledge use of the MONSooN system, a collaborative facility supplied under the Joint Weather and Climate Research Programme, which is a strategic partnership between the Met Office and the Natural Environment Research Council.

References

Ancell BC. 2013. Nonlinear characteristics of ensemble perturbation evolution and their application to forecasting high-impact events. *Weather and Forecasting* **28**(6): 1353–1365.

Baker L, Rudd A, Migliorini S, Bannister R. 2014. Representation of model error in a convective-scale ensemble prediction system. *Nonlinear Processes in Geophysics* **21**(1): 19–39.

Baldauf M, Seifert A, Förstner J, Majewski D, Raschendorfer M, Reinhardt T. 2011. Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. *Monthly Weather Review* **139**(12): 3887–3905.

Ben Bouallègue Z, Theis SE. 2014. Spatial techniques applied to precipitation ensemble forecasts: from verification results to probabilistic products. *Meteorological Applications* **21**(4): 922–929.

Blyth AM, Bennett LJ, Collier CG. 2015. High-resolution observations of precipitation from cumulonimbus clouds. *Meteorological Applications* **22**(1): 75–89.

Bouttier F, Vié B, Nuissier O, Raynaud L. 2012. Impact of stochastic physics in a convection-permitting ensemble. *Monthly Weather Review* **140**(11): 3706–3721.

Bowler NE, Arribas A, Beare SE, Mylne KR, Shutts GJ. 2009. The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **135**(640): 767–776.

Bowler NE, Arribas A, Mylne KR, Robertson KB, Beare SE. 2008. The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **134**(632): 703–722.

Bowler NE, Pierce CE, Seed AW. 2006. STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society* **132**(620): 2127–2155.

Browning KA, Roberts NM. 1994. Use of satellite imagery to diagnose events leading to frontal thunderstorms: Part I of a case study. *Meteorological Applications* **1**(4): 303–310.

Browning KA, Roberts NM. 1995. Use of satellite imagery to diagnose events leading to frontal thunderstorms: Part II of a case study. *Meteorological Applications* **2**(1): 3–9.

Buizza R. 1997. Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Monthly Weather Review* **125**: 99–119.

Buizza R, Houtekamer P, Pellerin G, Toth Z, Zhu Y, Wei M. 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review* **133**(5): 1076–1097.

Burt S. 2005. Cloudburst upon Hendraburnick down: the Boscastle storm of 16 August 2004. *Weather* **60**(8): 219–227.

Clark AJ, Gao J, Marsh PT, Smith T, Kain JS, Correia Jr J, Xue M, Kong F. 2013. Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Weather and Forecasting* **28**(2): 387–407.

Clark AJ, Kain JS, Stensrud DJ, Xue M, Kong F, Coniglio MC, Thomas KW, Wang Y, Brewster K, Gao J, *et al.* 2011. Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Monthly Weather Review* **139**(5): 1410–1418.

Davies T, Cullen MJP, Malcolm AJ, Mawson MH, Staniforth A, White AA, Wood N. 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Quarterly Journal of the Royal Meteorological Society* **131**(608):

1759–1782.

Dey SR, Leoncini G, Roberts NM, Plant RS, Migliorini S. 2014. A spatial view of ensemble spread in convection permitting ensembles. *Monthly Weather Review* **142**(11): 4091–4107.

Duc L, Saito K, Seko H. 2013. Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus A* **65**(0).

Ebert EE. 2008. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications* **15**(1): 51–64.

Edwards JM, Slingo A. 1996. Studies with a flexible new radiation code. I: Choosing a configuration for a large-scale model. *Quarterly Journal of the Royal Meteorological Society* **122**(531): 689–719.

Essery R, Best M, Cox P. 2001. MOSES 2.2 technical documentation. Technical report, Hadley Centre Technical Note.

Gebhardt C, Theis S, Paulat M, Ben Bouallègue Z. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research* **100**(2): 168–177.

Gilleland E, Ahijevych D, Brown BG, Casati B, Ebert EE. 2009. Intercomparison of spatial forecast verification methods. *Weather and Forecasting* **24**(5): 1416–1430.

Golding B, Ballard S, Mylne K, Roberts N, Saulter A, Wilson C, Agnew P, Davis L, Trice J, Jones C, *et al.* 2014. Forecasting capabilities for the London 2012 olympics. *Bulletin of the American Meteorological Society* **95**(6): 883–896.

Golding B, Clark P, May B. 2005. The Boscastle flood: Meteorological analysis of the conditions leading to flooding on 16 August 2004. *Weather* **60**(8): 230–235.

Golding BW. 1998. Nimrod: a system for generating automated very short range forecasts. *Meteorological Applications* **5**(1): 1–16.

Gray M, Marshall C. 1998. Mesoscale convective systems over the UK, 1981–97. *Weather* **53**(11): 388–396.

Hanley K, Kirshbaum D, Roberts N, Leoncini G. 2013. Sensitivities of a squall line over central Europe in a convective-scale ensemble. *Monthly Weather Review* **141**(1): 112–133.

Hanley KE, Kirshbaum DJ, Belcher SE, Roberts NM, Leoncini G. 2011. Ensemble predictability of an isolated mountain thunderstorm in a high-resolution model. *Quarterly Journal of the Royal Meteorological Society* **137**(661): 2124–2137.

Harrison DL, Driscoll SJ, Kitchen M. 2000. Improving precipitation estimates from weather radar using quality control and correction techniques. *Meteorological Applications* **7**(2): 135–144.

Harrison DL, Norman K, Pierce C, Gaussiat N. 2012. Radar products for hydrological applications in the UK. *Proceedings of the ICE - Water Management* **165**: 89–103(14).

Hohenegger C, Schär C. 2007. Atmospheric predictability at synoptic versus cloud-resolving scales. *Bulletin of the American Meteorological Society* **88**(7): 1783–1793.

Johnson A, Wang X. 2012. Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Monthly Weather Review* **140**(9): 3054–3077.

Johnson A, Wang X, Xue M, Kong F, Zhao G, Wang Y, Thomas KW, Brewster KA, Gao J. 2014. Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Monthly Weather Review* **142**(3): 1053–1073.

Kong F, Droegemeier KK, Hickmon NL. 2007. Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part II: Storm-scale experiments. *Monthly Weather Review* **135**(3): 759–782.

Lean HW, Clark PA, Dixon M, Roberts NM, Fitch A, Forbes R, Halliwell C. 2008. Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Monthly weather review* **136**(9): 3408 – 3424.

Leon DC, French JR, Lasher-Trapp S, Blyth AM, Abel SJ, Ballard S, Barrett A, Bennett LJ, Bower K, Brooks B, *et al.* 2015. The COnvective Precipitation Experiment (COPE): Investigating the origins of heavy precipitation in the southwestern UK. *Bulletin of the American Meteorological Society* doi:10.1175/BAMS-D-14-00157.1.

Leoncini G, Plant R, Gray S, Clark P. 2013. Ensemble forecasts of a flood-producing storm: comparison of the influence of model-state perturbations and parameter modifications. *Quarterly Journal of the Royal Meteorological Society* **139**(670): 198–211.

Leoncini G, Plant RS, Gray SL, Clark PA. 2010. Perturbation growth at the convective scale for CSIP IOP18. *Quarterly Journal of the Royal Meteorological Society* **136**(648): 653–670.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *Journal of Computational Physics* **227**: 3515–3539.

Lewis MW, Gray SL. 2010. Categorisation of synoptic environments associated with mesoscale convective systems over the UK. *Atmospheric Research* **97**(1): 194–213.

Lock A, Brown A, Bush M, Martin G, Smith R. 2000. A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests. *Monthly Weather Review* **128**(9): 3187–3199.

Mass CF, Ovens D, Westrick K, Colle BA. 2002. Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society* **83**(3): 407–430.

Melhauser C, Zhang F. 2012. Practical and intrinsic predictability of severe and convective weather at the mesoscales. *Journal of the Atmospheric Sciences* **69**(11): 3350–3371.

Migliorini S, Dixon M, Bannister R, Ballard S. 2011. Ensemble prediction for nowcasting with a convection-permitting model-I: description of the system and the impact of radar-derived surface precipitation rates. *Tellus A* **63**(3): 468–496.

Mittermaier M, Roberts N, Thompson SA. 2013. A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorological Applications* **20**(2): 176–186.

Mittermaier MP. 2014. A strategy for verifying near-convection-resolving model forecasts at observing sites. *Weather and Forecasting* **29**(2): 185–204.

Morcrette C, Lean H, Browning K, Nicol J, Roberts N, Clark P, Russell A, Blyth A. 2007. Combination of mesoscale and synoptic mechanisms for triggering an isolated thunderstorm: Observational case study of CSIP IOP 1. *Monthly Weather Review* **135**(11): 3728–3749.

Mylne K. 2013. Scientific framework for the ensemble prediction system for the UKV. MOSAC PAPER 18.6, UK Meteorological Office, URL `http://www.metoffice.gov.uk/media/pdf/q/0/MOSAC_18.6_Mylne.pdf`.

Palmer TN. 2000. Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics* **63**(2): 71.

Radhakrishna B, Zawadzki I, Fabry F. 2012. Predictability of precipitation from continental radar images. Part V: Growth and decay. *Journal of the Atmospheric Sciences* **69**(11): 3336–3349.

Rezacova D, Zacharov P, Sokol Z. 2009. Uncertainty in the area-related QPF for heavy convective precipitation. *Atmospheric Research* **93**(1): 238–246.

Roberts N. 2008. Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications* **15**(1): 163–169.

Roberts NM, Lean HW. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review* **136**(1): 78–97.

Russell A, Vaughan G, Norton E, Morcrette C, Browning K, Blyth A. 2008. Convective inhibition beneath an upper-level PV anomaly. *Quarterly Journal of the Royal Meteorological Society* **134**(631): 371–383.

Russell A, Vaughan G, Norton EG, Ricketts H, Morcrette CJ, Hewison TJ, Browning K, Blyth AM, *et al.* 2009. Convection forced by a descending dry layer and low-level moist convergence. *Tellus A* **61**(2): 250–263.

Simonin D, Ballard S, Li Z. 2014. Doppler radar radial wind assimilation using an hourly cycling 3D-Var with a 1.5 km resolution version of the Met Office Unified Model for nowcasting. *Quarterly Journal of the Royal Meteorological Society* **140**(684): 2298–2314.

Skok G. 2015. Analysis of fraction skill score properties for a displaced rainband in a rectangular domain. *Meteorological Applications* **22**(3): 477–484.

Surcel M, Zawadzki I, Yau MK. 2014. On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Monthly Weather Review* **142**(3): 1093–1105.

Warren RA, Kirshbaum DJ, Plant RS, Lean HW. 2014. A Boscastle-type quasi-stationary convective system over the UK southwest peninsula. *Quarterly Journal of the Royal Meteorological Society* **140**(678): 240–257.

Wilks DS. 2011. *Statistical methods in the atmospheric sciences*, vol. 100. Academic press.

Wilson DR, Ballard SP. 1999. A microphysically based precipitation scheme for the UK meteorological Office Mnified Model. *Quarterly Journal of the Royal Meteorological Society* **125**(557): 1607–1636.

Zacharov P, Rezacova D. 2009. Using the fractions skill score to assess the relationship between an ensemble QPF spread and skill. *Atmospheric Research* **94**(4): 684–693.