Assessing spatial precipitation uncertainties in a convective-scale ensemble

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

[a] *Department of Meteorology, University of Reading, Reading, UK*

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

[c] *Met Office, Exeter, UK*

[*] Correspondence to: Seonaid Dey, Centre for Ecology and Hydrology, Maclean Building, Crowmarsh Gifford, Wallingford, OX10 8BB. E-mail: seodey@ceh.ac.uk

New techniques have recently been developed to quantify the location-dependent spatial agreement between ensemble members, and the spatial spread-skill relationship. In this paper a summer of convection permitting ensemble forecasts are analysed to better understand the factors influencing location-dependent spatial agreement of precipitation fields and the spatial spread-skill relationship over the UK. The aim is to further investigate the agreement scale method, and to highlight the information that could be extracted for a more long-term routine model evaluation. Overall, for summer 2013, the UK 2.2km-resolution ensemble system was found to be reasonably well spread spatially, although there was a tendency for the ensemble to be over confident in the location of precipitation. For the forecast lead times considered (up to 36 hrs) a diurnal cycle was seen in the spatial agreement and in the spatial spread-skill relationship: the forecast spread and error did not increase noticeably with forecast lead time. Both the spatial agreement, and the spatial spread-skill, were dependent on the fractional coverage and average intensity of precipitation. A poor spread-skill relationship was associated with a low fractional coverage of rain and low average rain rates. The times with a smaller fractional coverage, or lower intensity, of precipitation were found to have lower spatial agreement. The spatial agreement was found to be location dependant, with higher confidence in the location of precipitation to the northwest of the UK.

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

1. Introduction

One of the challenges for weather forecasting is how to produce accurate and informative precipitation forecasts. Recent advances in computer power have allowed convective precipitation to be explicitly predicted using 'convection permitting' models with grid spacings of order 1km. These deterministic simulations produce realistic precipitation structures (e.g. Mass *et al.* 2002; Lean *et al.* 2008). However, due to the rapid error growth observed on the convective scale (of order hours: Hohenegger and Schär 2007; Melhauser and Zhang 2012; Radhakrishna *et al.* 2012), the location of convective-scale precipitation cannot be accurately predicted deterministically (e.g. Ben Bouallègue and Theis 2014; Surcel *et al.* 2016). Thus, in order to forecast convective scale precipitation, it is necessary to use an ensemble approach where the uncertainty in precipitation location can be quantified. Convective scale ensembles are now operational at several forecasting centres (Baldauf *et al.* 2011; Gebhardt *et al.* 2011; Bouttier *et al.* 2012; Golding *et al.* 2014).

Using a convective-scale ensemble system, it should be possible to give useful probabilistic forecasts of local precipitation, taking into account uncertainties in the precipitation location. Of course, this discussion assumes that the ensemble is well calibrated and unbiased; that the ensemble dispersion at a given time is representative of the true uncertainties in the forecast. How best to measure this convective scale spread-skill relationship is an open question. Other questions remain about the best method for obtaining information from convective scale ensembles; in particular how to quantify the uncertainty in precipitation location.

Conventional metrics of assessing ensemble characteristics, such as the ensemble standard deviation and Root Mean Square Error of the ensemble mean (RMSE, e.g. Wilks 2011) are inappropriate for use at the convective scale due to the double penalty problem where (even small) spatial differences are overly penalised. Additionally, due to the fast error growth observed at the convective scale, the ensemble mean is not a physical representation of the individual member forecasts (e.g. Ancell 2013). To address the double penalty problem in the verification of deterministic precipitation forecasts, a number of new forecast performance metrics have been developed (e.g. Roberts and Lean 2008; Ebert 2008; Gilleland *et al.* 2009; Johnson and Wang 2012). More recently, new methods have been explored for characterising both the skill, and dispersion, of convective-scale ensemble forecasts (Clark *et al.* 2011; Johnson *et al.* 2014; Surcel *et al.* 2014; Dey *et al.* 2014).

The methods of Clark *et al.* (2011); Surcel *et al.* (2014); Dey *et al.* (2014) provide a summary of the ensemble performance over the whole domain, which is useful to characterise the overall ensemble performance. In addition to this summary information, it is also important to investigate how the dispersion and skill of convection permitting ensembles vary with location in the domain. This is particularly true when considering fields such as precipitation, where different locations in the domain can sit within very different physical regimes (for example frontal precipitation compared with scattered convection). Using wavelet decomposition, Johnson *et al.* (2014) show the scale dependence of differences between the control forecast and observations, and how this varies across the domain. To consider the scale dependence of the ensemble spread-skill in a location-dependent manner, we use the agreement scales of Dey *et al.* (2016).

The agreement scale method calculates the length of the square area (labelled the agreement scale) surrounding each grid point over which pairs of precipitation fields meet a predefined similarity criterion. The agreement scale indicates the area over which forecast precipitation features should be expected to occur. The method provides an overview of the spatial ensemble characteristics while retaining location-dependent information (i.e. allowing the investigation of how uncertainty varies across the domain). Using the methods of Dey *et al.* (2016) both the spatial ensemble spread and the spatial spread-skill relationship can be computed.

The aims of this paper are twofold:

1. To use the agreement scales to investigate the spatial characteristics of summer UK precipitation, as obtained from model data and observations, for one particular season (June, July and August 2013).

2. To highlight areas/issues that might be of interest as a focus point for more routine, longer-term model evaluation and verification. As spatial neighbourhood methods can be computationally intense and data heavy it is useful to do this for an initial one-season study to allow informed choices to be made for longer assessments.

Note that, to enable a detailed investigation using the agreement scales, we do not compare with other methods. Such a comparison, in both theoretical and practical terms, is an important area of future investigation.

This paper examines hourly forecasts of UK rain rates from one particular operational ensemble, the Met Office Global and Regional Ensemble Prediction System UK ensemble (MOGREPS-UK Golding *et al.* 2014). MOGREPS-UK is introduced in Section 2 along with the radar data used for this study. To provide a context for the proceeding sections, an overview is given of the precipitation over the 2013 summer season. Section 3 details the analysis methods used, including details of the agreement scale method and its interpretation.

Results focus first on the ensemble information (spatial ensemble spread) to investigate the behaviour of, and information obtained from, the agreement scales over the UK for summer 2013 (addressing the first paper aim). In Section 4.1 agreement scale results averaged over the whole summer period are presented. In Section 4.2, the effect of different precipitation characteristics (fractional coverage of precipitation across the domain, and average intensity of raining points) on the agreement scales is investigated. Section 4.3 discusses the dependence of the agreement scales on the precipitation threshold used in the analysis, and Section 4.4 presents results for different times of day. To address the second aim of this paper, in Section 5 results are presented for the average spatial spread-skill relationship for the Summer 2013 season. The precipitation characteristics discussed in Section 4.2 are also considered in the context of the spatial spread-skill relationship. Finally, the overall conclusions from this work are presented and discussed in Section 6.

2. Data and model

2.1. Ensemble data

The MOGREPS-UK ensemble consists of 12 members one way nested inside members of the global ensemble MOGREPS-G (33 km grid spacing in the mid-latitudes). The lateral boundary conditions from MOGREPS-G are applied over a 5 point rim zone and blended with the MOGREPS-UK values over an additional 3 points as described in Davies (2014). MOGREPS-UK is run on variable resolution grid covering the UK and Ireland. The inner region of this grid, shown in light grey in Figure 1, is constantly spaced at 2.2km. Outside this constant resolution region, the grid spacing is gradually increased up to 4km to reduce the jump in resolution from MOGREPS-G. A full description of the variable resolution grid can be found in Tang *et al.* (2013). For this study, to speed up processing, two smaller subdomains were considered, covering the regions shown in mid-grey and dark grey in Figure 1. The subdomains were selected to fall within the area of radar data coverage (to be discussed in Section 2.2). As the same overall conclusions were drawn from both domains, this paper focuses on the northern domain to maintain brevity. Results for the southern domain can be found in Dey (2016).

MOGREPS-G perturbations are 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 11 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 MOGREPS-G forecasts. A 0300 UTC MOGREPS-UK start time was used for all forecasts presented in this paper.

For this study, both MOGREPS-UK and MOGREPS-G were run using version 8.2 of the Met Office Unified Model (MetUM), the version operational in summer 2013. Version 8.2 has a non-hydrostatic dynamical core with semi-Lagrangian advection (Davies *et al.* 2005) and a comprehensive set of parametrizations 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). Where possible, parameters are consistent across MOGREPS-UK and MOGREPS-G. The main difference is the explicit representation of convection (no convection scheme) in MOGREPS-UK, compared to MOGREPS-G where a convection scheme based on Gregory and Rowntree (1990) is used.

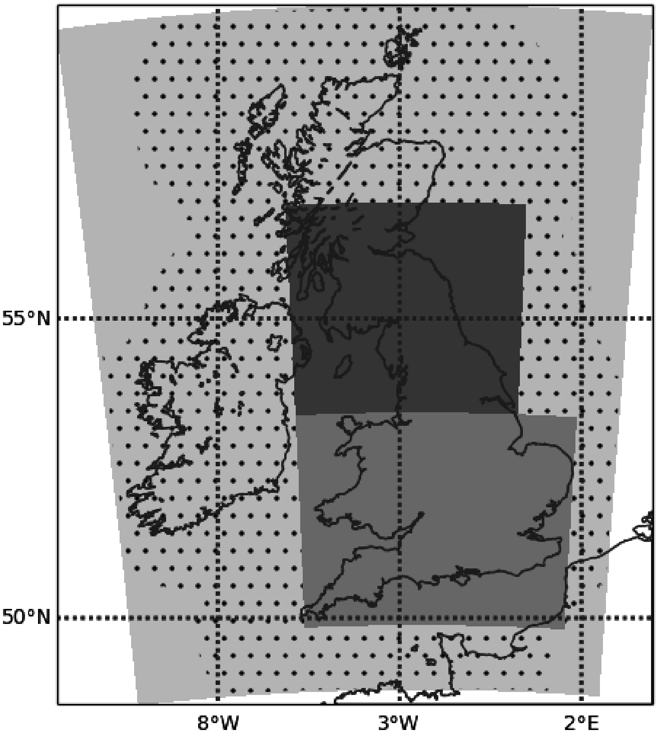

Figure 1. Domains considered: 2.2km MOGREPS-UK domain (light grey), radar coverage (dotted), northern domain (dark grey) and southern domain (middle grey).

2.2. *Radar data*

This study uses radar data from the Radarnet system (Golding 1998; Harrison *et al.* 2000, 2012), which provides a 1 km grid spacing rain rate composite over the UK, covering the dotted area shown in Figure 1. The Radarnet rain rates were bi-linearly interpolated onto the 2.2 km MOGREPS-UK grid before any comparisons were carried out. The results were not found to be sensitive to the re-gridding method: similar results were obtained when re-griding by averaging onto the 2.2km grid.

The Radarnet system includes many quality control measures, such as the subtraction of mean noise, application of a speckle filter and fuzzy logic to the reflectivity fields, identification of spurious echos, and corrections for radar-beam attenuation and topography (Harrison *et al.* 2012). Gauge data is also used to remove any systematic bias. However, despite these measures some unaccounted-for systematic errors remain. Hence, in this paper, additional checks were made on the radar composites. In particular, the data were not analysed at times when rain rates were apparently unphysical (defined to be greater than $300 mm\,hr^{-1}$), and times when several radars were offline (June 11^{th}2300 UTC, 12^{th} 0000 UTC, July 2^{nd} 0800 UTC and 18^{th} 0700 UTC to 1300 UTC). Occasionally, there were single points in the radar composite with missing rain rate data. As these points usually occurred within dry regions, their rain rates were set to zero. The radar data were also checked visually.

Note that, once these additional checks had been imposed, no further account was taken of errors in the Radarnet data: i.e. the Radarnet data was taken as ‘truth’. The automatic inclusion of observation errors in the methods of Dey *et al.* (2016) is an important avenue of future investigation which should be considered for an operational product.

Model data were obtained from the Met Office operational archive. From June 19^{th} 0300 UTC to June 20^{th} 1500 UTC no MOGREPS-UK data were available: these times have been removed from the analysis. The archived data did not contain any rain rates below $0.01 mm\,hr^{-1}$. For consistency, all points in the Radarnet data with rain rates below $0.01 mm\,hr^{-1}$ were also set to zero.

2.3. *Season overview*

Summer 2013 was slightly dryer and sunnier than average, with a dry warm period at the start of July, and a wet period from the end of July into the start of August (Met Office 2013). The season-averaged spatial distribution of precipitation agrees with previously published UK precipitation climatologies (e.g. Warren 2014; Fairman *et al.* 2015).

Figure 2 shows rain rate averages over all dates in summer 2013 for forecast lead times T+6 to T+29 (0900 UTC on the forecast start day to 0800 UTC the following day). A 24 hour averaging period was chosen to ensure that only one diurnal cycle was considered for each forecast, with the start time selected to be sufficiently far into the forecast to avoid spin up effects. Here, and for the remainder of this paper, this averaging method will be referred to as averaging over the "whole summer 2013 period".

Figures 2a and 2b show rain rate averages over the whole summer 2013 period for the north domain, for an ensemble member (here the control; other members give similar results) and the Radarnet data respectively. Only times with Radarnet data are included. The average precipitation is similar in the ensemble member and the Radarnet data. There are slight differences: for example the radar data has less precipitation over the North Sea and to the east of the UK. Differences between the ensemble and Radarnet precipitation fields will be quantified in Section 5. Figure 2c shows the average precipitation for one ensemble member over the whole of the MOGREPS-UK domain. All data from summer 2013 was included in Figure 2c (i.e. including times with no Radarnet data). Comparing Figure 2a with the north-domain region in Figure 2c (outlined in thick black), we see that the results are not overly impacted by neglecting times with missing Radarnet data.

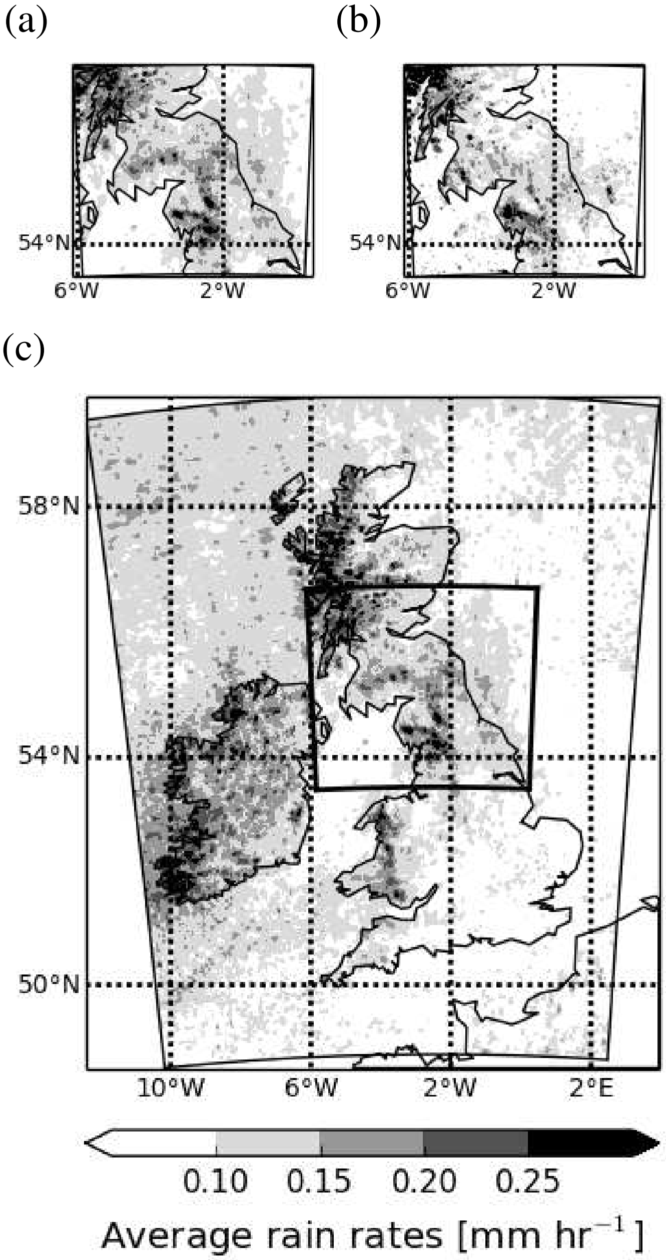

Figure 2. Rain rates averaged over all dates in June, July and August 2013, and forecast lead times from T+6 (0900 UTC) to T+29 (0800 UTC the following day) inclusive. A threshold of $0.01mm\,hr^{-1}$ was applied to the rain rate fields before averaging, with all rain rates below the threshold set to zero. (a) An ensemble member (the control; other members lead to the same conclusions) for the North domain only including times with Radarnet data available, (b) Radarnet data for the North domain and (c) an ensemble member (the control) for the UK domain with all data included.

3. Analysis methods

This paper measures the local spatial agreement between ensemble members, and between ensemble members and radar observations, using the methods of Dey *et al.* (2016). In particular, we use the average agreement scale between member-member pairs, denoted $S_{ij}^{A(\overline{mm})}$, and the average agreement scale between member-radar pairs, denoted $S_{ij}^{A(\overline{mo})}$. For ease of reference, we maintain the notation of Dey *et al.* (2016). Thus, in $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, we have

- "S_{ij}": A scale defining a square area (neighbourhood) centred upon grid point ij. S is the distance (in grid length units) from the centre to the edge of this area, not including the central grid point (for example, a 5 by 5 grid point area would have S=2, a 3 by 3 area would have S=1, and a single grid point would have S=0).
- $\overline{mm}$ or $\overline{mo}$ indicate the quantities being compared: all ensemble member pairs, or ensemble members and radar observations respectively.
- "A" indicates that S is the scale at which a specified level of agreement (to be discussed in Section 3.1) is obtained, at grid point ij, between pairs of ensemble

members ($\overline{mm}$) or between ensemble members and radar observations ($\overline{mo}$).

For ease of reading, the methods of calculating $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are summarised in Section 3.1. Key features of the agreement scales, and their interpretation, are then discussed in Section 3.2.

3.1. Calculation of agreement scales

To calculate $S_{ij}^{A(\overline{mm})}$ or $S_{ij}^{A(\overline{mo})}$ we must first focus on single pairs of fields, that is a pair of ensemble members, or an ensemble member and radar observations. The aim is to calculate the agreement scales $S_{ij}^{A(f_1 f_2)}$ between these two fields f_1 and f_2. Note that $S_{ij}^{A(f_1 f_2)}$ is calculated separately at each grid point in the domain. Hence, for simplicity, this discussion will focus on one particular point, labelled point P.

First the rain rate values of f_1 and f_2 at grid point P (f_{1ij}^0 and f_{2ij}^0, where the superscript "0" indicates that we are comparing f_{1ij} and f_{2ij} at a single grid point, that is at a scale of 0) are compared by calculating the quantity

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

f_{1ij} and f_{2ij} are considered to be suitably similar at this single grid point (a scale of zero) if $D_{ij}^0 \leq D_{\mathrm{crit},ij}^0$, where $D_{\mathrm{crit},ij}^0 = \alpha$, a pre-defined constant. Consistent with Dey *et al.* (2016) a value of $\alpha = 0.5$ is used in this paper. This choice means that a ratio f_{1ij}^0/f_{2ij}^0 in the range 2-$\sqrt{3}$ to 2+$\sqrt{3}$ is considered suitably similar at the grid scale, so that the criterion is primarily dictated by whether rainfall occurs in both fields at the given location, and differences in relative magnitude of up to 3.73 are tolerated.

If f_1 and f_2 are found to be suitably similar at a scale of zero (the single grid point P), then the agreement scale at point P, $S_{ij}^{A(f_1 f_2)}$ is zero. If f_1 and f_2 are not found to be suitably similar, then we consider instead an area of scale =1 (3 by 3 grid points) centred upon point P. The average rain rate values of f_1 and f_2 over this area (f_{1ij}^1 and f_{2ij}^1) are calculated, and compared in a similar manner to Equation 1, which generalises for any scale S to give:

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

f_{1ij} and f_{2ij} are considered to be suitably similar at a scale of S if

$$D_{ij}^S \leq D_{\mathrm{crit},ij}^S \tag{3}$$

where

$$D_{\mathrm{crit},ij}^S = \alpha + (1-\alpha)\frac{S}{S_{\mathrm{lim}}}. \tag{4}$$

Note that $D_{\mathrm{crit},ij}^S$ depends on S: larger forecast differences are considered acceptable for larger scales. S_{lim} is a predetermined, fixed maximum scale and, by construction, Eq. 3 is always satisfied at the scale S_{lim}. Further discussion regarding the reasoning behind Equation 4 can be found in Dey *et al.* (2016). In this paper we use a value of $S_{\mathrm{lim}} = 80$, consistent with Dey *et al.* (2016), and suitable for the domains considered here.

If f_{1ij}^1 and f_{2ij}^1 are found to be suitably similar, then the agreement scale at point P, $S_{ij}^{A(f_1 f_2)}$ is one. If f_{1ij}^1 and f_{2ij}^1 are not found to be suitably similar, then the process described above is repeated for incrementally larger scales ($S = 2, 3, ..., S_{\mathrm{lim}}$) until an agreement scale is found.

By calculating the agreement scales at each grid point in the domain, we obtain a map of agreement between the fields f_1 and f_2. However, as discussed in Dey *et al.* (2016) these maps can be noisy, due to the differences between f_1 and f_2 not always decreasing uniformly with increasing neighbourhood size (precipitation fields have been shown to become increasingly similar with increasing neighbourhood size on average (Roberts and Lean 2008; Clark *et al.* 2011; Mittermaier *et al.* 2013) but not necessarily for individual comparisons). To obtain smooth agreement scale maps it is necessary to average over a number of field comparisons. This is done for the calculation of $S_{ij}^{A(\overline{mm})}$ by taking the mean, at each grid point, over the $S_{ij}^{A(f_1 f_2)}$ calculated separately for each independent pair of ensemble members. Thus, for an ensemble of N members, we have $N_p = \frac{N(N-1)}{2}$ independent member pairs, N_p values of $S_{ij}^{A(f_1 f_2)}$, and so N_p values contributing to $S_{ij}^{A(\overline{mm})}$. Similarly, for the calculation of $S_{ij}^{A(\overline{mo})}$ we have N ensemble member–radar pairs, N fields of $S_{ij}^{A(f_1 f_2)}$, and an average of these N values produces $S_{ij}^{A(\overline{mo})}$. Although $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are calculated by averaging over a different number of pairs, Dey *et al.* (2016) showed (using an idealised experiment) that they can be compared to diagnose the spatial spread-skill relationship of the ensemble.

3.2. *Key features of the agreement scales*

The $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ provide measures of the agreement between precipitation fields at each grid point in the domain. In particular, they are calculated by considering differences in the amount of precipitation between two fields, when averaging over a given neighbourhood size. This is important for the meaning and interpretation of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$.

Consider the comparison of two ensemble members over a neighbourhood centred within a region of precipitation. The difference between the average precipitation amounts over this neighbourhood will be influenced by differences in the placement of precipitation between the members (in this paper this is referred to as the spatial predictability) and also differences in the intensity of precipitation. Next consider a neighbourhood centred on a point away from the region of precipitation. In this situation the agreement scale will be determined by the distance of the central point from the precipitation: Equations 2 and 4 compare only precipitation differences between the fields so, when there is no precipitation, the criterion of Equation 3 is not met and a larger neighbourhood is sought. The combination of these effects, as measured by the agreement scales, will be referred to as the “spatial agreement” between the fields. These features of the analysis methods have two key implications for interpreting the results in this paper:

1. Larger values of $S_{ij}^{A(\overline{mo})}$ do not indicate a poorer performance of the ensemble. Instead, they show that a large neighbourhood size is needed at this point to find consistency in the precipitation fields. Hence, when considered independently of $S_{ij}^{A(\overline{mm})}$, the $S_{ij}^{A(\overline{mo})}$ can not be used to verify the ensemble performance. However, as the $S_{ij}^{A(\overline{mm})}$, and the $S_{ij}^{A(\overline{mo})}$ are consistently defined, a comparison of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ *can* be used to verify the ensemble performance, and to diagnose the spatial spread-skill relationship of the ensemble.
2. As the $S_{ij}^{A(\overline{mm})}$ and the $S_{ij}^{A(\overline{mo})}$ are influenced by the spatial predictability, bias in precipitation intensity and distance from the precipitation, care must be taken in their interpretation. For example, a systematic bias between the ensemble and radar (such as the high bias in the model often seen in convection permitting forecasts, e.g. Lean *et al.* (2008)) may result in an an overestimation of the $S_{ij}^{A(\overline{mo})}$ at grid points where the ensemble predicts heavier precipitation than is seen in the radar, and an underestimation of $S_{ij}^{A(\overline{mo})}$ at grid points where the ensemble predicts lighter precipitation than is seen in the radar. Thus, the effect of a systematic bias will depend on the distribution (across the domain) of differences between ensemble and radar rain rates. Hence, it is necessary to test the effect of bias before drawing conclusions from a comparison of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$.

If a single forecast (for a specific time) is considered we can compare the $S_{ij}^{A(\overline{mm})}$ to a precipitation field, say of one ensemble member, and ascertain where the scales represent spatial predictability (i.e. where the location is in the vicinity of precipitation). However, this direct comparison is not possible if we consider an average over a number of lead times or cases. Thus, when the $S_{ij}^{A(\overline{mm})}$ are averaged over a number of cases, the scales will (by design of the method) have a dependence on the coverage of precipitation: this is examined in Section 4.2. The dependence of agreement scales on precipitation coverage makes physical sense: we expect to be more confident in the location of precipitation when the precipitation covers a larger area. Surcel *et al.* (2016) also demonstrate that precipitation is less predictable in situations with a lower precipitation coverage.

3.3. *Thresholding*

As discussed in Section 3.1, $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are calculated from the precipitation fields themselves: it is not necessary to use a precipitation threshold on the fields as done in other methods, such as the Fractions Skill Score (Roberts and Lean 2008), which use a threshold to define binary fields. However, there are situations (such as when producing probability forecasts) where it is useful to consider different ranges of precipitation intensity. This can be done for the agreement scales by applying a lower precipitation threshold to the fields before calculating $S_{ij}^{A(\overline{mm})}$ or $S_{ij}^{A(\overline{mo})}$. Thresholds of 0.1, 1.0, and 4.0 $mm\,hr^{-1}$ are considered here, with all rain rate values below the precipitation threshold set to zero before any calculations are carried out (rain rate values above the threshold are unchanged). When temporal averages are taken of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$, times where rain rates do not exceed the threshold at any point in the domain for any ensemble member or for the radar data (i.e. times which are totally dry), are not included in the average. As these times would have $S_{ij}^{A(\overline{mm})} = S_{ij}^{A(\overline{mo})} = S_{\mathrm{lim}}$ at all grid points in the domain, including them would introduce a high bias on $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$.

In Sections 4.2 and 5 the effect of precipitation characteristics (fractional coverage of precipitation across the domain, or the average rain rate of raining points across the domain) on the agreement scales is considered. To define the fractional coverage of precipitation, or the average over raining points, a threshold must be selected to define the points which are considered to be precipitating or not. Unless otherwise specified, a threshold of 0.01 $mm\,hr^{-1}$ is used to make this distinction.

3.4. *Notation*

For ease of reference, this subsection summarises the notation used. All of the quantities refer to a specific forecast time.

- $S_{ij}^{A(\overline{mm})}$ or $S_{ij}^{A(\overline{mo})}$ denote location-dependent agreement scales between ensemble member–member pairs or ensemble member–radar pairs respectively.
- $S^{A(\overline{mm})}$ denotes the $S_{ij}^{A(\overline{mm})}$ averaged over all grid points in the domain ("domain averaged agreement scale").
- $S_{0.1}^{A(\overline{mm})}$ denotes a domain averaged agreement scale calculated for a specified precipitation threshold (here $0.1mm\,hr^{-1}$).
- $\mathrm{Cover}_{0.01}$ denotes the fraction of the domain covered by precipitation with rain rates exceeding a specified threshold (here $0.01mm\,hr^{-1}$).
- $\mathrm{Intensity}_{0.01}$ denotes the rain rate average of points in the domain with rain rates exceeding a specified threshold (here $0.01mm\,hr^{-1}$).

4. Results: agreement between ensemble members

This section uses the $S_{ij}^{A(\overline{mm})}$ to investigate spatial characteristics of precipitation over summer 2013 as forecast by the MOGREPS-UK ensemble. Through linking the $S_{ij}^{A(\overline{mm})}$ to properties of the precipitation, the $S_{ij}^{A(\overline{mm})}$ methodology is also investigated.

4.1. *Season averaged results*

Results are first presented for $S_{ij}^{A(\overline{mm})}$ averaged over the whole summer 2013 period. The aim is to give an overview of the spatial agreement over this particular summer season, highlighting regions of the domain where the ensemble is more confident about the location of precipitation (due to higher spatial predictability, or larger precipitation coverage). These average agreement scales indicate the typical areas (neighbourhood sizes) over which precipitation in the ensemble should be considered accurate, if a single fixed scale had to be chosen at each grid point in the domain. Of course, this interpretation only holds if the ensemble is well spread spatially; if the $S_{ij}^{A(\overline{mm})}$ is representative of the $S_{ij}^{A(\overline{mo})}$. The $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ are compared in Section 5. At individual times, the scales can differ considerably from the average values. Results showing how the $S_{ij}^{A(\overline{mm})}$ depend on precipitation characteristics will be presented in Section 4.2.

Figure 3 shows the $S_{ij}^{A(\overline{mm})}$ averaged over the whole summer 2013 period for the MOGREPS-UK domain. The $S_{ij}^{A(\overline{mm})}$ are smaller in the northwest, over mountainous regions, and along the western coasts of both the UK and Ireland: in these regions the ensemble is more confident about the location of precipitation. The ensemble is not confident about the location of precipitation close to the grid scale, with a minimum time-mean $S_{ij}^{A(\overline{mm})}$ (i.e. the minimum value in Figure 3) of 12 grid points (a total neighbourhood length of 55km). This reinforces the need to use neighbourhood methods in the interpretation of precipitation forecasts.

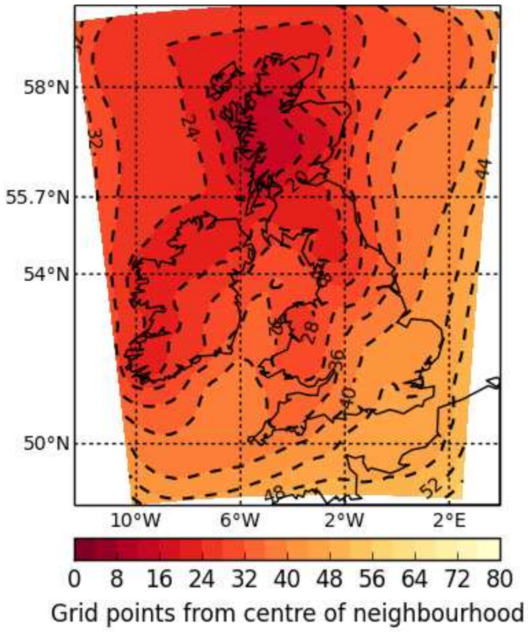

Figure 3. Map of the member-member agreement scales $S_{ij}^{A(\overline{mm})}$ averaged over forecasts from T+6 (0900 UTC) to T+29 (0800 UTC the following day) for all dates in June, July and August 2013. All MOGREPS-UK data within these times have been included (i.e. including times with no Radarnet data).

As expected from the method of calculating agreement scales, the distribution of $S_{ij}^{A(\overline{mm})}$ in Figure 3 resembles the distribution of average rain rates across the same period, shown for an ensemble member in Figure 2. To test whether the variation in $S_{ij}^{A(\overline{mm})}$ is explained fully by the variations in average rain rate across the domain, histograms were produced for the summer 2013 average $S_{ij}^{A(\overline{mm})}$ (as shown in Figure 3) conditioned on rain rates in Figure 2c. Different parts of the domain were separately considered to highlight variations in the $S_{ij}^{A(\overline{mm})}$ distribution.

Figure 4 shows histograms for mean rain rate ranges 0.1 to 0.2$mm\ hr^{-1}$, 0.2 to 0.3$mm\ hr^{-1}$, and above 0.3$mm\ hr^{-1}$. Figure 4a includes points north of 55.7°N, while Figure 4b includes points south of 55.7°N. Both panels show that points with heavier seasonal average rain rates have a narrower distribution of season-averaged $S_{ij}^{A(\overline{mm})}$, with smaller mean $S_{ij}^{A(\overline{mm})}$. This confirms that variations in the amount of rain do relate to variations in the $S_{ij}^{A(\overline{mm})}$. However, although the distributions in Figures 4a and 4b have similar shapes there are also differences that are not accounted for by considering different rain rate ranges. In particular, the distributions in Figure 4a have smaller mean agreement scales than those in Figure 4b. This shows that the smaller agreement scales seen to the north of Figure 3 are not explained fully by this region being wetter on average; there are other factors, possibly related to the higher and steeper orography in this region giving higher spatial predictability of precipitation. Although the clearest differences in $S_{ij}^{A(\overline{mm})}$ distributions were seen when splitting distributions at 55.7°N (as shown in Figure 4; approximately located at the Scottish lowlands), similar conclusions were also drawn from comparing other regions (not shown). For example, the Cumbrian hills (around 55°N,-4.5°E) showed higher season average $S_{ij}^{A(\overline{mm})}$ than the Pennines (around 54.5°N,-3°E), in a way that is not explained by the rainfall amounts.

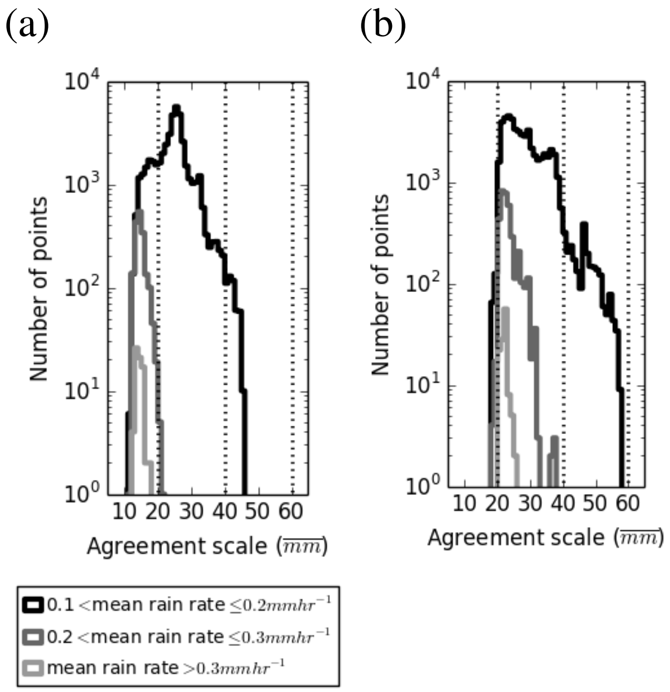

Figure 4. Histograms showing the distributions of summer 2013 average member-member agreement scale $S_{ij}^{A(\overline{mm})}$ (as shown in Figure 3) using only those points from Figure 2 with mean rain rates falling within specific ranges. (a) considering only points in Figure 3 north of 55.7°N and (b) considering only points in Figure 3 south of 55.7°N. Results are shown for three rain rate ranges: 0.1 to 0.2$mm\ hr^{-1}$, 0.2 to 0.3$mm\ hr^{-1}$, and above 0.3$mm\ hr^{-1}$.

4.2. Dependence of spatial agreement on precipitation characteristics

It is expected that the $S_{ij}^{A(\overline{mm})}$ will depend on the specific characteristics of the precipitation field. In this subsection links are made between $S^{A(\overline{mm})}$ and two domain-wide measures of the precipitation characteristics, the fraction of the domain covered by precipitation (fractional coverage) and the average rain rate of points in the domain with precipitation (intensity of precipitation). The fractional coverage and intensity of precipitation were calculated as explained in Section 3.3, for one ensemble member (here the control; using other ensemble members gave similar results). Here, and for the remainder of this paper, results are presented for the north domain (shown in dark grey in Figure 1), and only times with Radarnet data are included in the analysis.

As the $S_{ij}^{A(\overline{mm})}$ measure the overlap between precipitation fields (a larger overlap giving smaller agreement scales), and distance from the precipitation (which will always be smaller when there is more precipitation), it is expected that smaller values of $S_{ij}^{A(\overline{mm})}$ will be found when there is a larger coverage of precipitation. This makes physical sense, and agrees with the results found by Surcel *et al.* (2016) using the decorrelation scale of Surcel *et al.* (2014). Here we ask how much of the variation in the $S^{A(\overline{mm})}$ is explained by variations in the fractional coverage? Figure 5a shows a scatter plot of $\text{Cover}_{0.01}$ against $S^{A(\overline{mm})}$, with each point corresponding to a forecast time in summer 2013 (hourly data from T+6 to T+29 were considered). A negative correlation is found between these variables: as expected, smaller scales are generally seen at times with higher precipitation coverage. However, there is still a spread of values, giving a correlation magnitude of 0.6. This suggests that, even though the coverage of precipitation does influence the $S^{A(\overline{mm})}$, the agreement scales also contain additional information.

Figure 5b shows a scatter plot, in the same format as 5a, but this time with $\text{Intensity}_{0.01}$ plotted on the y-axis. Similarly to the $\text{Cover}_{0.01}$ results, a negative correlation is seen between $\text{Intensity}_{0.01}$ and $S^{A(\overline{mm})}$, but with a lower correlation magnitude of 0.43. Thus, cases with higher domain

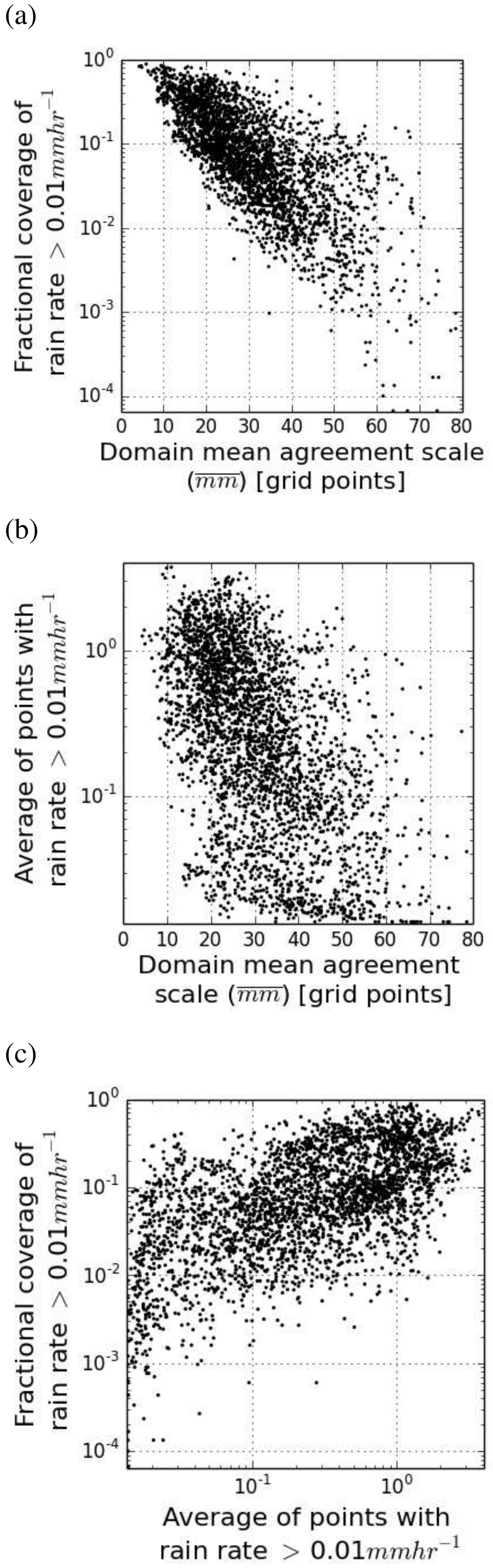

Figure 5. Scatter plots of (a) $\text{Cover}_{0.01}$ against $S^{A(\overline{mm})}$, (b) $\text{Intensity}_{0.01}$ against $S^{A(\overline{mm})}$ and (c) $\text{Cover}_{0.01}$ against $\text{Intensity}_{0.01}$. Each point on the scatter plot corresponds to a forecast time (hourly from T+6; 0900 UTC on forecast start day to T+29; 0800 UTC the following day). Correlations of -0.6, -0.43 and 0.45 were obtained for sub-figures (a), (b) and (c) respectively.

averaged rain rates have, in general, smaller $S^{A(\overline{mm})}$. However, as can be seen from Figure 5b, there is a large range of possible values of Intensity$_{0.01}$ for a given value of $S^{A(\overline{mm})}$. As shown in 5c, Cover$_{0.01}$ and Intensity$_{0.01}$ correlate positively with each other, a correlation of 0.45 being obtained. Thus we find that cases with higher average rain rates often also have a higher coverage of precipitation; the higher precipitation values tend to be embedded inside larger precipitation structures.

4.3. Varying precipitation threshold

Section 4.2 related the $S^{A(\overline{mm})}$ to precipitation characteristics. Now we investigate how the spatial agreement depends on the range of precipitation values considered, by applying thresholds to the precipitation fields before calculating the $S_{ij}^{A(\overline{mm})}$. The method of applying thresholds was presented in Section 3.3.

Figure 6a-c show maps of the $S_{ij}^{A(\overline{mm})}$ averaged over the whole summer 2013 period, calculated for precipitation thresholds 0.01, 0.1 and 1.0 $mm\,hr^{-1}$ respectively. All three thresholds have a similar spatial pattern of $S_{ij}^{A(\overline{mm})}$, with smaller scales to the northwest and over land, agreeing with the MOGREPS-UK domain results (Figure 3). The consistency of the location-dependence of the season-averaged $S_{ij}^{A(\overline{mm})}$ shows that different rain rate ranges have, on average, similar influences on their spatial predictability, for example the topography and orography.

A change in the magnitude of $S_{ij}^{A(\overline{mm})}$ with increasing threshold may be expected as higher precipitation thresholds will result in lower values of Cover, and higher values of Intensity. However, as lower values of Cover are associated with *larger* $S_{ij}^{A(\overline{mm})}$, and higher values of Intensity are associated with *smaller* $S_{ij}^{A(\overline{mm})}$, the sign of the threshold dependence is not easily predicted. Figure 6 shows that higher thresholds result in larger season-average agreement scales, suggesting that it is the difference in Cover between the thresholds that has the most impact. It should be noted that the results of Section 4.2 hold when considering the different thresholds separately.

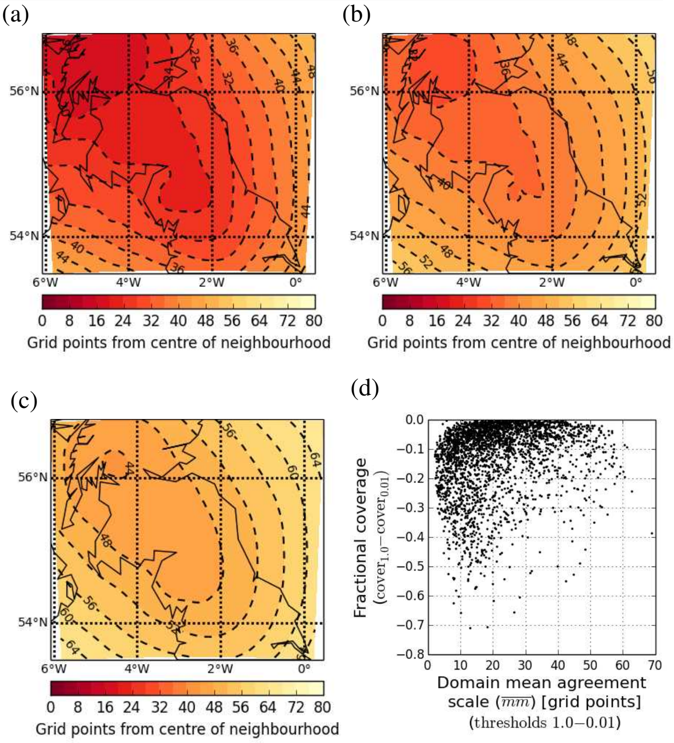

Figure 6. (a)-(c) Maps of $S_{ij}^{A(\overline{mm})}$ for different precipitation thresholds averaged over forecast lead times T+6 (0900 UTC) to T+29 (1800 UTC the following day) where precipitation occurred over the specified threshold (at at least one grid point in the domain). Results are shown for rain rates greater than (a) 0.01$mm\,hr^{-1}$, (b) 0.1$mm\,hr^{-1}$, and (c) 1.0$mm\,hr^{-1}$. (d) Scatter plot of the difference in fractional coverage (Cover$_{1.0}$−Cover$_{0.01}$) at each time included in the average for (a)-(c) against the corresponding difference between $S_{1.0}^{A(\overline{mm})}$ and $S_{0.01}^{A(\overline{mm})}$.

To investigate the extent to which the variation in $S_{ij}^{A(\overline{mm})}$ for different thresholds relates to differences in Cover we compare, at each time in summer 2013, the difference in fractional coverage of precipitation between two thresholds, to the difference in $S^{A(\overline{mm})}$ between the same thresholds. This is shown in Figure 6d for a comparison of the 1.0$mm\,hr^{-1}$ and 0.01$mm\,hr^{-1}$ threshold results. The fractional coverage difference at each time (Cover$_{1.0}$ − Cover$_{0.01}$) is plotted against the corresponding difference in domain averaged agreement scale ($S_{1.0}^{A(\overline{mm})} - S_{0.01}^{A(\overline{mm})}$). From Figure 6d a low positive correlation of 0.22 is obtained between the coverage and agreement scale differences. As a negative correlation would have been expected from the relationship between *Cover* and $S_{ij}^{A(\overline{mm})}$, this suggests that other factors, such as perhaps differences in the spatial structure of precipitation or differences in the inherent predictability of different precipitation intensities, contribute noticeably to the threshold dependence of the $S_{ij}^{A(\overline{mm})}$.

4.4. Diurnal effects

Sections 4.1 to 4.3 have included together all forecast lead times from T+6 (0900 UTC) to T+29 (0800 UTC the following day). In this section the temporal evolution of the $S_{ij}^{A(\overline{mm})}$ throughout the forecast is investigated. In particular we focus on the season average $S_{ij}^{A(\overline{mm})}$, separated by forecast lead-time. Unlike the findings relating $S^{A(\overline{mm})}$ to precipitation characteristics (Section 4.2) the temporal evolution of the $S^{A(\overline{mm})}$ is found to depend on the precipitation threshold applied to the fields. In this subsection results are presented using thresholds of 0.01 and 1.0 $mm\,hr^{-1}$ which summarise the range of observed behaviour.

Figure 7 shows the season average $S_{ij}^{A(\overline{mm})}$ at forecast lead times T+12 (1500 UTC), T+24 (0300 UTC) and T+36 (1500 UTC) for thresholds $0.01mm\,hr^{-1}$ (left) and $1.0mm\,hr^{-1}$ (right). The $S_{ij}^{A(\overline{mm})}$ vary with a diurnal cycle, with similar $S_{ij}^{A(\overline{mm})}$ seen at T+12 and T+36 (1500 UTC on the forecast start day, and 1500 UTC on the following day). This similarity is also found when comparing other forecast lead times separated by 24 hours (e.g. T+6 with T+30), as seen from time series of the $S^{A(\overline{mm})}$ (Figure 8c). Thus, neither threshold shows a clear trend of $S^{A(\overline{mm})}$ increasing with forecast lead time, which might have been expected (on average) early on in the forecast from the growth of forecast errors with lead time (e.g. Hohenegger and Schär 2007; Melhauser and Zhang 2012). This suggests that, for the rain rate fields considered here, which show high variability over small spatial distances, small scale processes dominate over the large scale growth of forecast errors.

The $0.01mm\,hr^{-1}$ and $1.0mm\,hr^{-1}$ thresholds show a very different diurnal evolution of the $S_{ij}^{A(\overline{mm})}$. Agreement scales for the $0.01mm\,hr^{-1}$ threshold are much less variable with time than for higher thresholds. This will be discussed again in Section 5 in the context of the spread-skill relationship. For higher precipitation thresholds (exemplified here by the $1.0mm\,hr^{-1}$ threshold results), a marked diurnal cycle is seen in the agreement scales, with larger $S_{ij}^{A(\overline{mm})}$ (lower spatial agreement) seen at night (Figure 7d), and smaller $S_{ij}^{A(\overline{mm})}$ (higher spatial agreement) seen in the day (Figures 7b,f).

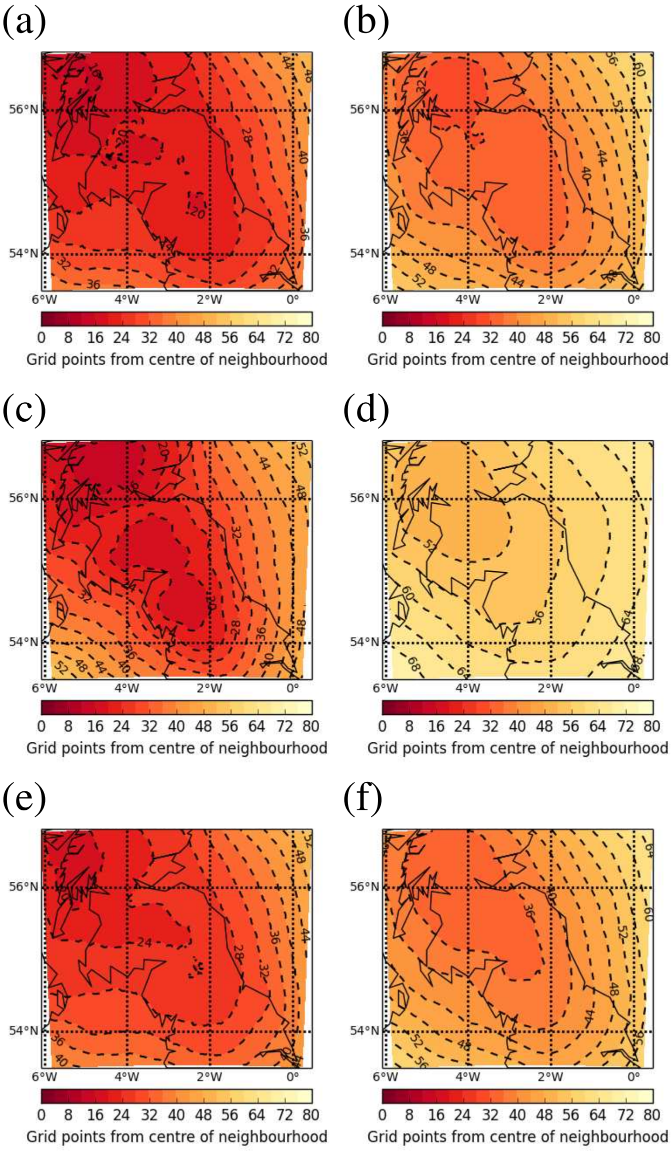

Figure 7. Maps of the member-member agreement scales $S_{ij}^{A(\overline{mm})}$ split by time of day (forecast lead time) for two thresholds: $0.01mm\,hr^{-1}$ (left) and $1.0mm\,hr^{-1}$ (right). Three forecast lead times are shown: (a) and (b) T+12; 1500 UTC, (c) and (d) T+24; 0300 UTC, (e) and (f) T+36; 1500UTC. Data were averaged over dates in summer 2013 where precipitation occurred over the specified threshold (at at-least one grid point in the domain).

As the $S_{ij}^{A(\overline{mm})}$ are related to Cover and Intensity, we investigate whether the diurnal cycle in $S_{ij}^{A(\overline{mm})}$ is related to the diurnal cycle in these precipitation characteristics. Figure 8 shows time series (from forecast lead times T+1 to T+36, averaged over all dates in Summer 2013) of Cover and Intensity for precipitation thresholds 0.01, 0.1 and $1.0mm\,hr^{-1}$. By construction, smaller values of Cover and higher values of Intensity are seen for higher thresholds. The Intensity time series show a clear diurnal cycle, with the heaviest average rain rates seen in the afternoon (T+12 and T+36): daytime convective activity is influencing the Intensity values. The values of Cover show less temporal variation. Correlations, calculated between the time series shown in Figure 8a and Figure 8b and the corresponding time series of $S^{A(\overline{mm})}$ (shown in Figure 8c), are given in Table 1. Low correlations (not significant, as defined by a 2-tailed p-value

Table 1. Correlations between time series of $S^{A(\overline{mm})}$, Intensity and $S^{A(\overline{mm})}$,Cover for different precipitation thresholds. All forecast lead times (T+1 to T+36) were included in the time series (similar results were obtained when only including times from T+6 to avoid spin-up effects.)

Threshold $[mm\,hr^{-1}]$	0.01	0.1	1.0
Correlation with Cover	0.17	-0.2	-0.27
Correlation with Intensity	-0.71	-0.92	-0.91

of greater than 0.05) are found between $S^{A(\overline{mm})}$ and Cover. Higher, significant, correlations are found between $S^{A(\overline{mm})}$ and Intensity. This shows that the diurnal cycle in $S^{A(\overline{mm})}$ is more strongly anti-correlated to the diurnal cycle in Intensity, than to the diurnal cycle in Cover.

5. Comparing with observations

Section 4 presented results of the agreement scales calculated between ensemble member pairs, $S_{ij}^{A(\overline{mm})}$, to investigate the spatial precipitation characteristics for a UK summer season, and to examine the utility of the agreement scale method. It was shown that the $S_{ij}^{A(\overline{mm})}$ are useful for understanding the factors influencing spatial agreement of ensemble member precipitation fields. This is helpful in understanding the model behaviour. However, in order to provide useful forecasts of spatial agreement, it is necessary that the ensemble has a good spatial spread-skill relationship. Thus the spatial differences between pairs of ensemble members should be representative of the differences between ensemble members and truth (here given by radar observations). In this section, the $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ (introduced in Section 3.1) are compared over summer 2013 to quantify the ensemble performance over this period. The aim is to highlight influences on the spatial spread-skill relationship which could be used to inform longer term, routine model evaluation and verification (the second aim of this paper, see Section 1).

The spatial spread-skill results are presented in the form of a binned scatter plot, as introduced in Dey *et al.* (2016). The binned scatter plot allows the two fields of the $S_{ij}^{A(\overline{mm})}$ and the $S_{ij}^{A(\overline{mo})}$ to be compared (at a specified forecast time) while preserving location-dependent information. To produce a binned scatter plot, a bin-size is first selected; here a bin size (agreement scale range) of 10 grid points is used for

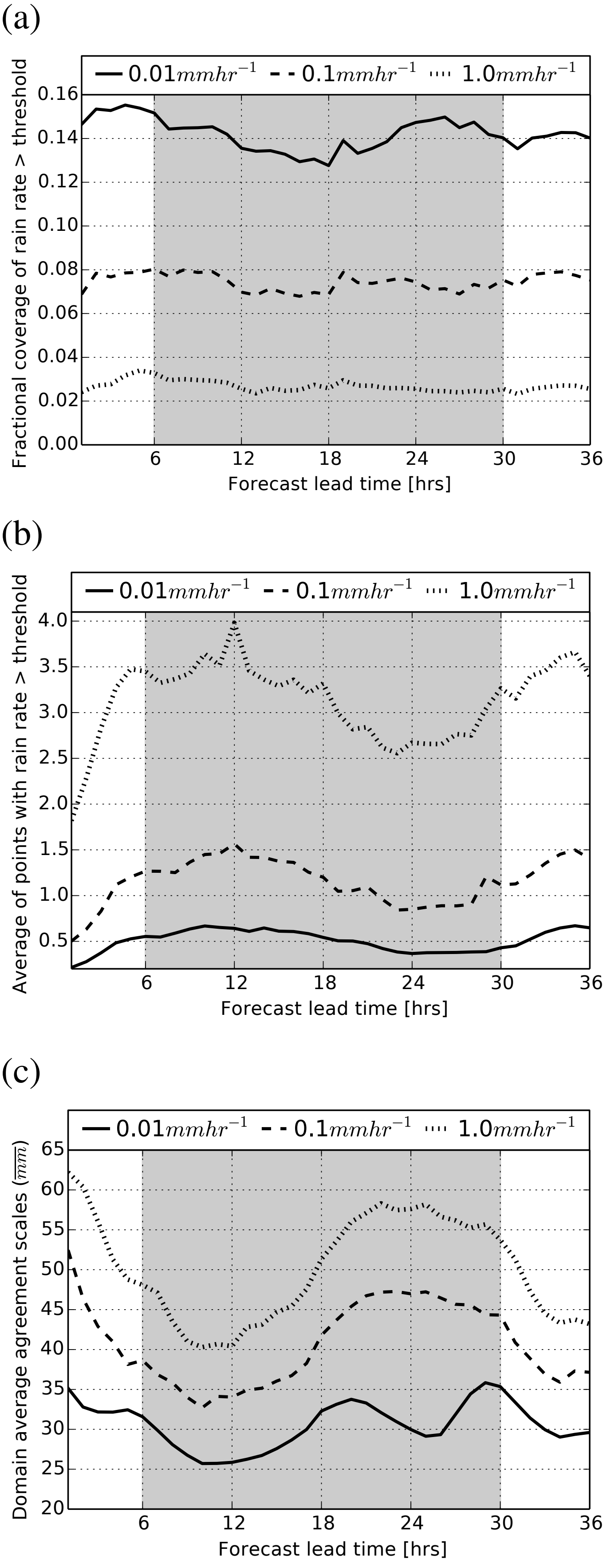

Figure 8. Time series averaged over all dates in summer 2013 for (a) Cover, (b) Intensity and (c) $S^{A(\overline{mm})}$. Each plot shows results for three thresholds: 0.01$mm\,hr^{-1}$ (solid), 0.1$mm\,hr^{-1}$ (dashed) and 1.0$mm\,hr^{-1}$ (dotted). The 24 hour averaging period used for plots of the whole summer 2013 period (0900 UTC, forecast lead time T+9 to 0800 UTC the following day, forecast lead time T+29) is shown in grey.

each bin. This bin-size was found to be sufficiently large to ensure enough points in each bin to give meaningful results, but sufficiently small to retain scale-dependent information. A running bin is used, with bins from 1 to 10, 2 to 11, 3 to 12, ..., 71 to 80 grid points. For each bin the $S_{ij}^{A(\overline{mm})}$ are first considered, and the average taken of the $S_{ij}^{A(\overline{mm})}$ over all grid points whose values fall into the specified bin-range. This value is plotted on the x-axis. Next, the average $S_{ij}^{A(\overline{mo})}$ value *over these same grid points* is calculated and plotted on the y-axis. Thus, after considering all bins, we produce a line of mean $S_{ij}^{A(\overline{mo})}$ (for each bin) against mean $S_{ij}^{A(\overline{mm})}$ (for each bin). If this line falls above the diagonal, then we have $S_{ij}^{A(\overline{mo})}$ greater than $S_{ij}^{A(\overline{mm})}$: the ensemble is spatially under spread. If the line falls below the diagonal we have $S_{ij}^{A(\overline{mo})}$ less than $S_{ij}^{A(\overline{mm})}$, and the ensemble is spatially over spread. By taking the average of binned scatter plot traces calculated over a large number of different times, the spatial spread-skill relationship of the ensemble can be quantified (Dey *et al.* 2016).

5.1. Season averaged results

First, to give an overview of the ensemble performance over the three month period, Figure 9 shows the average binned histograms over the whole summer 2013 season (all dates and forecast lead times T+6 to T+29). Results are shown for four different precipitation thresholds: 0.01, 0.1, 1.0 and 4.0 $mm\,hr^{-1}$. The different thresholds give similar results: all show lines slightly above the diagonal (for agreement scales below 50 grid points), suggesting that, for this particular summer period the ensemble was slightly under spread spatially. This is most noticeable for the smallest agreement-scale bins, which are located in areas of precipitation. Hence, the under estimation of these scales by the ensemble is linked to differences in the spatial predictability of the precipitation (as opposed to just the amount of precipitation in the domain). For agreement scales above 50 grid points, the lines on the binned scatter plot lie close-to or below the diagonal, showing that these scales tend to be slightly over estimated by the ensemble. It is thought that this is caused by the radar rain rates having a larger number of separate regions of precipitation in the domain, although it has not been possible to quantitatively prove this interpretation in this current study.

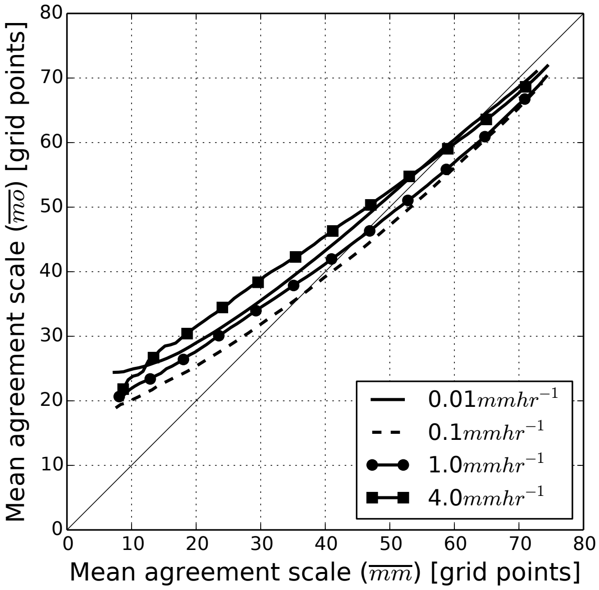

Figure 9. Binned scatter plots averaged over Summer 2013 and lead times T+6 to T+29 for thresholds 0.01 (solid), 0.1 (dashed), 1.0 (solid with circles) and 4.0 (solid with squares) $mm\,hr^{-1}$.

Of course, as discussed in Section 3.2, a systematic bias between the ensemble members and radar can effect the spatial spread-skill relationship. To investigate whether this is the case for MOGREPS-UK, all figures in this section were reproduced with an artificial bias applied to the radar data before calculating the $S_{ij}^{A(\overline{mo})}$. This was achieved by multiplying the radar data by 0.5 (to simulate the ensemble over-predicting precipitation) and 1.5 (to simulate the model under-predicting precipitation). These values were selected to be slightly larger than the bias in the ensemble members, which (when estimated from Intensity calculated from the radar divided by Intensity calculated for one ensemble member) varies between 0.8 (for the 0.1$mm\,hr^{-1}$ threshold) and 1.3 (for the 4$mm\,hr^{-1}$ threshold). The bias was applied after thresholding the data to ensure the same fractional coverage was considered. It was found that adding the artificial bias did not significantly change the results in Figures 9, 10 and 11, and did not alter the overall conclusions presented. This gives confidence in the interpretation that it is spatial predictability differences that lead to the ensembles appearing under spread.

Although the different precipitation threshold results shown in Figure 9 lead to similar results, there are some differences. For example, when a 0.1 $mm\,hr^{-1}$ threshold is applied the ensemble has a better spatial spread-skill relationship than for a 0.01 $mm\,hr^{-1}$ threshold (for scales below 50 grid points). In general, one might expect the less predictable precipitation associated with higher thresholds to be harder to quantify, and indeed this is seen in Figure 9 for the $1.0 mm\,hr^{-1}$ and $4.0 mm\,hr^{-1}$ thresholds. A detailed investigation of the individual ensemble member and radar fields showed that the improvement in spread-skill between the $0.01 mm\,hr^{-1}$ and $0.1 mm\,hr^{-1}$ thresholds was due to the forecasts having around twice the number of points with rain rates in this range (compared to the radar observations), and these points being located within precipitation regions for the model, but at the edge of precipitation regions for the radar data. Hence, removing points with rain rates from 0.01 to $0.1 mm\,hr^{-1}$ (by applying the $0.1 mm\,hr^{-1}$ threshold) resulted in a greater variation of precipitation structures between ensemble members, and an increase in $S_{ij}^{A(\overline{mm})}$ with respect to $S_{ij}^{A(\overline{mo})}$, leading to a better spread-skill relationship for scales less than 50 grid points. This behaviour emphasises how the spatial agreement, as measured by the agreement scales, is directly related to the precipitation structures themselves, and gives useful information about the ensemble performance. This information is not easily extracted from 'domain wide' summary measures of rainfall features, or from time average rainfall maps.

5.2. Dependence on precipitation structure

In Section 4.3 it was shown that the $S_{ij}^{A(\overline{mm})}$ were dependent on the fractional coverage of precipitation across the domain (higher precipitation coverage giving smaller $S_{ij}^{A(\overline{mm})}$), and the average intensity of precipitation across the domain (higher intensity giving smaller $S_{ij}^{A(\overline{mm})}$). Similar relationships were found for the $S_{ij}^{A(\overline{mo})}$ (not shown). Here we investigate whether the fractional coverage of precipitation ($\text{Cover}_{0.01}$) or the average intensity of raining points ($\text{Intensity}_{0.01}$), affects the spatial spread-skill relationship. In particular we ask whether there are situations for which the $S_{ij}^{A(\overline{mm})}$ provides a poorer indication of the $S_{ij}^{A(\overline{mo})}$, which may be of particular interest for future, more long-term ensemble verification studies. Note that the dependence of $S_{ij}^{A(\overline{mm})}$ and $S_{ij}^{A(\overline{mo})}$ on $\text{Cover}_{0.01}$ and $\text{Intensity}_{0.01}$ does not necessarily imply that the spatial spread-skill relationship will also depend on these measures.

To investigate how the spatial spread-skill relationship depends on $\text{Cover}_{0.01}$ and $\text{Intensity}_{0.01}$ we use binned scatter plots averaged over times in summer 2013 (using forecast lead times T+6 to T+29) where the $\text{Cover}_{0.01}$ or $\text{Intensity}_{0.01}$ (shown in Figure 5) fall within predefined ranges. Four ranges were selected for $\text{Cover}_{0.01}$ (0 to 1%, 1% to 10%, 10% to 20% and 20% to 100%), and three for $\text{Intensity}_{0.01}$ (0.01 to $0.1 mm\,hr^{-1}$, 0.1 to $1.0 mm\,hr^{-1}$, 1.0 to $4.0 mm\,hr^{-1}$). These ranges were chosen to allow the average to be taken over a sufficient number of times to reduce noise in the results (a minimum of 200 times is considered, for the $\text{Cover}_{0.01}$ range 0 to 1%).

Figure 10 shows binned scatter plots for data subset by the ranges discussed above for (a) $\text{Cover}_{0.01}$ and (b) $\text{Intensity}_{0.01}$. It can be seen that the spatial spread-skill relationship is highly dependent on both measures, with poorer spatial spread-skill seen for times with lower values of $\text{Cover}_{0.01}$, and times with lower values of $\text{Intensity}_{0.01}$ (note these are not necessarily the same times).

Thus, for summer 2013, the MOGREPS-UK ensemble was most under-spread at times with low rain rates and at times with a small fractional coverage of precipitation across the domain. It may be thought that these situations, which individually have less impact than heavier more widespread precipitation events, are of little importance, or that it is unreasonable to expect models to be able to predict such cases and that they should be excluded from the analysis (Nachamkin and Schmidt 2015). However, we argue that these situations are an important consideration if automated probability products are to be produced from the ensemble output. For example, if the ensemble were to regularly suggest a high chance of light precipitation within a small given region, and it rained somewhere else instead, this could degrade users' confidence.

5.3. Dependence of spatial spread-skill on diurnal effects

Finally we consider the effect of time of day (different forecast lead times) on the spatial spread-skill relationship. As the

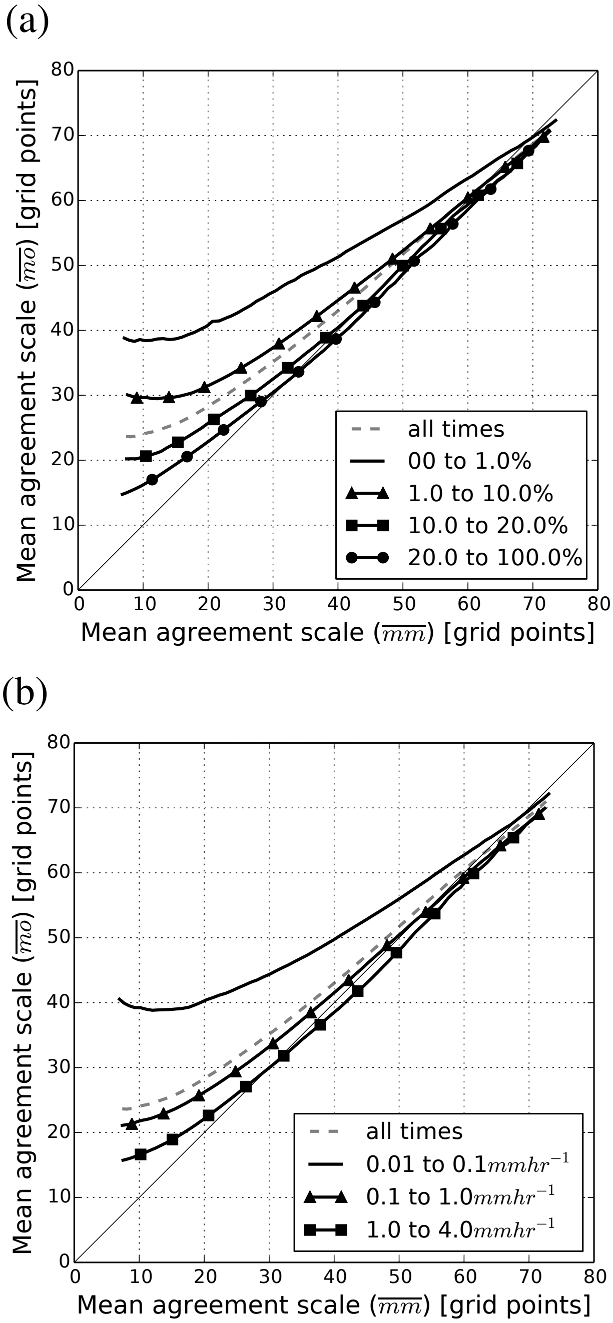

Figure 10. Binned scatter plots for a threshold of $0.01 mm\, hr^{-1}$ averaged over times in summer 2013 and forecast lead times T+6 to T+29 with predefined precipitation characteristics: (a) for varying ranges of $\text{Cover}_{0.01}$, and (b) for varying ranges of $\text{Intensity}_{0.01}$.

effect of forecast time on $S_{ij}^{A(\overline{mm})}$ was found to depend on the precipitation threshold considered (Section 4.4) results are presented for two precipitation thresholds, 0.01 and $1.0 mm\, hr^{-1}$.

Figure 11 shows binned scatter plots averaged over all dates in Summer 2013, for (a) $0.01 mm\, hr^{-1}$ and (b) $1.0 mm\, hr^{-1}$ precipitation thresholds. The same three forecast lead times (times of day) used in Figure 7 are shown here: T+12 (1500 UTC), T+24 (0300 UTC) and T+36 (1500 UTC). The average over forecast lead times T+6 to T+29 (0900 UTC to 0800 UTC the following day) is also included for reference. Figure 11 shows that splitting the data by time of day (i.e. the effect of the diurnal cycle) has less impact than splitting by fractional coverage or average rain amount (Figure 10): it is the precipitation characteristics that have most effect on the spatial spread-skill relationship.

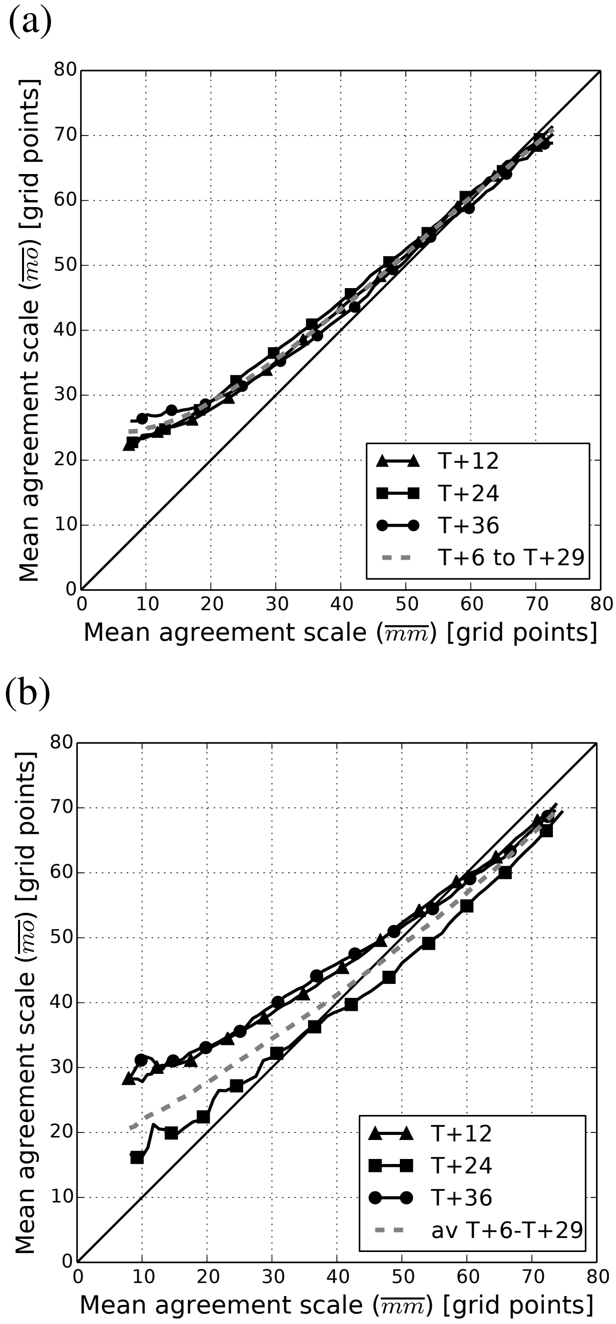

Figure 11. Binned scatter plots for a threshold of (a) $0.01 mm\, hr^{-1}$ and (b) 1.0 $mm\, hr^{-1}$ at selected forecast lead times.

The difference between the $S_{ij}^{A(\overline{mm})}$ results for 0.01 and $1.0 mm\, hr^{-1}$ precipitation thresholds (a stronger diurnal cycle was found for the higher threshold) is also seen in the spatial spread-skill results. Specifically, there is little temporal variation in the $0.01 mm\, hr^{-1}$ threshold results (Figure 11a) whereas the $1.0 mm\, hr^{-1}$ threshold results show a clear diurnal cycle (Figure 11b). For the 1.0$mm\,hr^{-1}$ threshold, the ensemble was more spatially under-spread during the day (T+12, T+36), and less spatially under-spread (or even spatially over-spread for larger $S^{A(\overline{mm})}_{ij}$) at night. Comparison with Figure 7 shows that the ensemble was most under spread (the $S^{A(\overline{mm})}_{ij}$ were too small) at the times when the smallest $S^{A(\overline{mm})}_{ij}$ were found, and slightly spatially over spread (for agreement scales above 50 grid points) when largest $S^{A(\overline{mm})}_{ij}$ were seen. This suggests that the ensemble is overestimating the diurnal range of spatial agreement scales.

Given the dependence of the spatial spread-skill on the fractional coverage and intensity of precipitation (Section 5.2), it is useful to relate the diurnal cycle in spatial spread-skill to the diurnal cycle of differences in Cover and Intensity between the ensemble and radar observations. Time series of Cover and Intensity (averaged over all dates in summer 2013) were calculated for both an ensemble member (as shown in Figure 8, hereafter labelled Cover$_{\text{Control}}$ and Intensity$_{\text{Control}}$) and for the radar data (hereafter Cover$_{\text{Radar}}$ or Intensity$_{\text{Radar}}$). Correlations calculated between time series of Cover$_{\text{Control}}$ – Cover$_{\text{Radar}}$, Intensity$_{\text{Control}}$ – Intensity$_{\text{Radar}}$, and $S^{A(\overline{mm})}$–$S^{A(\overline{mo})}$ are given in Table 2 for the thresholds 0.01, 0.1 and 1.0$mm\,hr^{-1}$. These suggest that differences in the diurnal cycle of Cover and Intensity (between the ensemble and radar data) do play a role in the diurnal cycle of spatial spread-skill, but do not fully explain it. Correlations with Cover$_{\text{Control}}$ – Cover$_{\text{Radar}}$ vary around -0.6, with no consistent threshold dependence. Correlations with Intensity$_{\text{Control}}$ – Intensity$_{\text{Radar}}$ are close to zero for the 0.01$mm\,hr^{-1}$ threshold and not significant (as defined by a 2-tailed p-value of greater than 0.05). For higher thresholds the correlations negative, and of larger magnitude. Thus, when the ensemble overestimates the average precipitation intensity it underestimates the $S^{A(\overline{mm})}$ (i.e. is too confident about the rainfall location).

Table 2. Correlations between time series of $S^{A(\overline{mm})}$–$S^{A(\overline{mo})}$ and Intensity$_{\text{Control}}$ –Intensity$_{\text{Radar}}$ or Cover$_{\text{Control}}$ – Cover$_{\text{Radar}}$ for different precipitation thresholds. All forecast lead times (T+1 to T+36) were included in the time series (similar results were obtained when only including times from T+6 to avoid spin-up effects).

Threshold $[mm\,hr^{-1}]$	0.01	0.1	1.0
Correlation with Cover$_{\text{Control}}$ - Cover$_{\text{Radar}}$	-0.74	-0.58	-0.63
Correlation with Intensity$_{\text{Control}}$ - Intensity$_{\text{Radar}}$	-0.03	-0.74	-0.82

6. Discussion and conclusions

This paper has investigated the spatial characteristics of Summer 2013 UK precipitation, using the MOGREPS-UK convective scale ensemble system operational at the time, and radar observations. To focus on the location-dependence of the spatial ensemble behaviour the methods of Dey *et al.* (2016) were employed. In order to understand relationships between the agreement scales and features of the precipitation fields, the spatial agreement between ensemble member pairs, $S^{A(\overline{mm})}_{ij}$, was considered. The ensemble spatial spread-skill relationship was also investigated by comparing the $S^{A(\overline{mm})}_{ij}$ with the spatial agreement between ensemble members and radar observations, $S^{A(\overline{mo})}_{ij}$. Different precipitation characteristics, different times of day, and different precipitation thresholds were investigated to highlight areas and issues that might form a focal point for more long-term routine forecast evaluation.

Overall, for summer 2013, smaller $S^{A(\overline{mm})}$ (indicating higher spatial agreement between ensemble member precipitation fields) were seen at times with a larger fractional coverage of precipitation across the domain, or higher average rain rates. This is expected from the method of calculating $S^{A(\overline{mm})}_{ij}$, which considers the spatial overlap between precipitation fields. However, correlations between the fractional coverage and $S^{A(\overline{mm})}$ (-0.6; Figure 5a) or between the average intensity of raining points and $S^{A(\overline{mm})}$ (-0.43; Figure 5b) are too low to fully explain the variations in $S^{A(\overline{mm})}$: other factors are also important. Thus, the $S^{A(\overline{mm})}$ contain information that cannot be simply obtained by considering only the fractional coverage or intensity of precipitation. This was confirmed by results of the $S^{A(\overline{mm})}_{ij}$ calculated over the whole of the UK and Ireland. Although smaller agreement scale values were obtained to the northwest, and over the west coast of both the UK and Ireland, which were on average wetter, the differences in $S^{A(\overline{mm})}_{ij}$ were not fully explained by the precipitation differences.

It was found that different precipitation thresholds did not influence the spatial variation of $S_{ij}^{A(\overline{mm})}$ across the domain, suggesting that different precipitation ranges have similar constraints on the spatial predictability of precipitation. The $S_{ij}^{A(\overline{mm})}$ magnitude was found to depend on the precipitation threshold used, with higher thresholds (selecting heavier precipitation) showing larger $S_{ij}^{A(\overline{mm})}$. This is expected as higher precipitation thresholds select a lower fractional coverage of precipitation, and agrees with the work of Dey *et al.* (2014).

When considering the spatial spread-skill relationship it was found that, overall for summer 2013, the MOGREPS-UK ensemble was slightly spatially under spread, particularly for small values of $S_{ij}^{A(\overline{mm})}$ (i.e when the ensemble was confident about the positioning of precipitation). These results agree with those of Tennant (2015) who found MOGREPS-UK to be under spread for the variables temperature and visibility, and agree also with the general perception that convection permitting ensembles are under spread (e.g. Clark *et al.* 2011; Bouttier *et al.* 2012; Duda *et al.* 2016). Note that, traditional spread-skill measures are inappropriate for the convective scale, and hence, for precipitation, give results which should be interpreted with caution. Consistent with this, comparing the Mean Squared Error and variance of the ensemble member forecasts considered in this paper produced noisy results, which could not be easily interpreted. These results can be found in Dey (2016) and have not been repeated in this paper.

Similar results were obtained for different precipitation thresholds. Note that these results are for one particular summer period, and the version of the ensemble operational at that time. It would be valuable to perform a similar analysis for other ensemble versions, and for a larger data sample, to verify the performance of MOGREPS-UK more generally.

Similarly to the $S_{ij}^{A(\overline{mm})}$, the spatial spread-skill relationship was found to be dependent on the fractional coverage, and intensity, of precipitation across the domain. In particular, for summer 2013, the ensemble was most spatially under-spread for times with low fractional coverage, or times with low average precipitation intensity. Although precipitation with such characteristics does not have the same direct impact of heavy or widespread precipitation, it is nonetheless an important consideration if the ensemble system is to be used to generate automatic products. Hence, it is recommended that a long-term location-dependent spatial verification of the ensemble system does include, and considers separately, times with low rain rates or low fractional coverage.

It is expected that, on average, differences between ensemble member forecasts increase with increasing forecast lead time due to the upscale growth of forecast errors. This was not seen for the convective scale ensemble data considered in this paper. In particular, the $S_{ij}^{A(\overline{mm})}$ were not found to increase (which would indicate increased spatial differences), and the ensemble spread-skill was not found to deteriorate with lead time. Possible reasons for this include the short 36 hour forecast used for this study (during 36 hours the large-scale errors will remain small) and the consideration of rain rates which vary over small scales and are influenced by very localised processes. The consideration of longer lead time convective scale ensemble forecasts would be a useful avenue of future investigation.

A diurnal cycle was seen for the $S_{ij}^{A(\overline{mm})}$ and for the spatial spread-skill relationship. The diurnal cycle is stronger for higher precipitation thresholds, with smaller values of $S_{ij}^{A(\overline{mm})}$, and a poorer spread-skill relationship (the ensemble is more spatially under-spread), seen during the afternoon. This suggests that, for summer 2013, the ensemble overestimated the diurnal variability in agreement scales. The diurnal cycle in $S_{ij}^{A(\overline{mm})}$, and in the spatial spread-skill relationship, were related to the diurnal cycle in fractional coverage and precipitation intensity. For both $S_{ij}^{A(\overline{mm})}$ and the spatial spread-skill relationship, higher magnitude correlations were found with the diurnal cycle in precipitation intensity, than with the diurnal cycle in fractional coverage (as shown in Tables 1 and 2). Further investigating the diurnal effects on ensemble spatial agreement, perhaps thorough detailed case studies, would allow the responsible processes in the model to be identified, and highlight areas for model improvement.

In summary, this paper demonstrates the useful information that can be gained about ensemble performance and characteristics by using the location-dependent spatial approach of Dey *et al.* (2016). Areas have also been identified for further detailed studies, and also the potential for longer term routine ensemble and model verification. For example, our results suggest that it would be useful to include several forecast lead times in a long term investigation of the spatial ensemble spread-skill relationship. This would allow the impact of forecast lead time on an ensembles' ability to provide spatial information to be accurately assessed. Other investigations should examine the possibility of including observation uncertainty in the agreement-scale method.

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.

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.

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.

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. 2014. Lateral boundary conditions for limited area models. *Quarterly Journal of the Royal Meteorological Society* **140**(678): 185–196.

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 SRA. 2016. A spatial approach to the analysis of convective scale ensemble systems. PhD thesis, Department of Meteorology, University of Reading, URL `http://centaur.reading.ac.uk/65945/`.

Dey SRA, 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.

Dey SRA, Roberts NM, Plant RS, Migliorini S. 2016. A new method for the characterisation and verification of local spatial predictability for convective scale ensembles. *Quarterly Journal of the Royal Meteorological Society* doi:10.1002/qj.2792. Early Online Release.

Duda JD, Wang X, Kong F, Xue M, Berner J. 2016. Impact of a stochastic kinetic energy backscatter scheme on warm season convection-allowing ensemble forecasts. *Monthly Weather Review* **144**(5): 1887–1908.

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.

Fairman JG, Schultz DM, Kirshbaum DJ, Gray SL, Barrett AI. 2015. A radar-based rainfall climatology of Great Britain and Ireland. *Weather* **70**(5): 153–158.

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

Gregory D, Rowntree PR. 1990. A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Monthly Weather Review* **118**(7): 1483–1506.

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.

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.

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.

Met Office. 2013. Met office weather summaries. URL `http://www.metoffice.gov.uk/climate/uk/summaries/2013/summer`. Accessed 28/04/2015.

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.

Nachamkin JE, Schmidt J. 2015. Applying a neighborhood fractions sampling approach as a diagnostic tool. *Monthly Weather Review* **143**(11): 4736– 4749.

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.

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.

Surcel M, Zawadzki I, Yau M. 2016. The case-to-case variability of the predictability of precipitation by a storm-scale ensemble forecasting system. *Monthly Weather Review* **144**(1): 193–212.

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.

Tang Y, Lean HW, Bornemann J. 2013. The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteorological Applications* **20**(4): 417–426.

Tennant W. 2015. Improving initial condition perturbations for MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society* doi:10.1002/qj.2524. Early Online Release.

Warren RA. 2014. Quasi-stationary convective systems in the uk. PhD thesis, Department of Meteorology, University of Reading.

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.