# A spatial view of ensemble spread in convection permitting ensembles

SEONAID R. A. DEY *

*Departement of Meteorology, University of Reading, Reading, UK*

GIOVANNI LEONCINI†, NIGEL M. ROBERTS

*MetOffice@Reading, Met Office, Reading, UK*

ROBERT S. PLANT

*Departement of Meteorology, University of Reading, Reading, UK*

STEFANO MIGLIORINI

*Departement of Meteorology, University of Reading, Reading, UK*

* *Corresponding author address:* Seonaid Dey, Department of Meteorology, University of Reading, Earley Gate, PO Box 243,Reading,RG6 6BB

E-mail: s.dey@pgr.reading.ac.uk

† Current affiliation: Aspen Re, Zürich, Switzerland

ABSTRACT

With movement towards kilometer scale ensembles, new techniques are needed for their characterization. We present a new methodology for detailed spatial ensemble characterization using the Fractions Skill Score (FSS). To evaluate spatial forecast differences the average and standard deviation are taken of the FSS calculated over all ensemble member-member pairs at different scales and lead times. These methods were found to give important information about the ensemble behavior allowing the identification of useful spatial scales, spin-up times for the model, and upscale growth of errors and forecast differences. The ensemble spread was found to be highly dependent on the spatial scales considered and the threshold applied to the field. High thresholds picked out localized and intense values that gave large temporal variability in ensemble spread: local processes and under sampling dominate for these thresholds. For lower thresholds the ensemble spread increases with time as differences between the ensemble members upscale. Two convective cases were investigated based on the UK Met Office United Model run at 2.2 km resolution. Different ensemble types were considered: ensembles produced using the Met Office Global and Regional Ensemble Prediction System (MOGREPS) and an ensemble produced using different model physics configurations. Comparison of the MOGREPS and multiphysics ensembles demonstrated the utility of spatial ensemble evaluation techniques for assessing the impact of different perturbation strategies and the need for assessing spread at different, believable, spatial scales.

# 1. Introduction

It has been long known that at small spatial scales forecast errors grow more rapidly (Lorenz (1969); Ehrendorfer (1997); Palmer (2000) and references therein) possibly resulting in rapid upscale error growth in high resolution models. In recent years these subjects have again come under discussion as increases in computer power allow models to be run at higher and higher resolutions (Mass et al. (2002) and references therein, Lean et al. (2008)). Hohenegger and Schär (2007a) compared the predictability at large (around 80 km) and convection-permitting (2.2 km) scales and found error doubling times around ten times shorter for the higher resolution simulations. Further work has investigated the links between mesoscale processes and error growth with a focus on moist dynamics (Zhang 2005; Hohenegger et al. 2006) and the separation of equilibrium and triggered convection to distinguish different modes of predictability in convective events (Keil and Craig 2011; Zimmer et al. 2011; Craig et al. 2012; Keil et al. 2013).

Ensemble prediction systems strive to represent the meteorological uncertainty present in a particular forecast and have been widely used to assess error growth in a variety of high-resolution situations (Walser et al. 2004; Walser and Schär 2004; Hohenegger and Schär 2007b; Hanley et al. 2011, 2013). Further investigations have been conducted into different ensemble perturbation strategies for high resolution ensembles including initial condition perturbations (Migliorini et al. 2011; Caron 2013; Kühnlein et al. 2013), physics perturbations (Stensrud et al. 2000; Hacker et al. 2011; Gebhardt et al. 2011; Vié et al. 2012; Baker et al. 2014), perturbation of boundary layer parameters (Martin and Xue 2006; Leoncini et al. 2010; Done et al. 2012) and the use of different physics schemes (Berner et al. 2011; Leoncini et al. 2012).

The aim of this paper is to provide a new methodology for evaluating, thoroughly, the differences between members of a convection permitting ensemble and the dependence of these differences on spatial scale. These methods are based on the Fractions Skill Score (FSS, Roberts and Lean (2008); Roberts (2008)). Various considerations are discussed including the forecast evolution through different lead times, the effect of considering different threshold values for the fields used to calculate the FSS, and the comparison of different forecast variables. For the demonstrative purposes of this paper two convective cases are considered using ensembles produced as part of the Met Office Global and Regional Ensemble Prediction System (MOGREPS, Bowler et al. (2008, 2009)). The spatial spread of the ensemble members is characterized and the realism of the ensemble spread is tested by comparing with the skill against radar derived precipitation accumulations. Radar data is necessary as a verification source because of its high spatial coverage.

The technique used to determine spatial differences between members can also be used for the comparison of different model formulations within the ensemble. To demonstrate this, different model physics configurations were considered in addition to the MOGREPS ensemble members for the second case study. This specific example is provided to demonstrate the utility of spatial evaluation techniques in the comparison of different ensemble formulations. Note, however, that a complete systematic evaluation comparing different types of physics configuration is outside the scope of this paper. To do this it would be necessary to consider a large number of cases with different convective forcing as detailed by, for example, Stensrud et al. (2000); Keil et al. (2013). The spatial ensemble spread produced by different physics configurations strategies is evaluated and compared to that of the MOGREPS ensemble. In operational frameworks, different physics configurations are often considered in addition to initial and boundary condition perturbations and so the spatial spread produced by an ensemble with different MOGREPS members combined with different physics configurations is also investigated.

To evaluate convection permitting ensembles in a sensible way it is necessary to choose a verification approach that considers multiple spatial scales and does not suffer from the double penalty problem where spatial errors are penalized twice: once for being a near miss, and again for begin a false positive. Many possible spatial verification approaches have been proposed in recent years; for an overview the reader is referred to the review papers of Ebert

(2008), Gilleland et al. (2009), and Johnson and Wang (2013). The spatial approach has also been applied to ensembles (Clark et al. 2011; Johnson et al. 2014; Surcel et al. 2014). Here we have chosen to focus on the Fractions Skill Score (FSS) of Roberts and Lean (2008); Roberts (2008). The FSS is a fuzzy verification measure used to compare two fields within a given square neighborhood.

Since its original formulation the FSS has been used for different applications and several further developments have been proposed. Schwartz et al. (2010) consider circular neighborhoods to calculate the field of fractions at each grid point and then produce probabilistic guidance using the field of fractions as a neighborhood probability. Duda and Gallus (2013) also use the circular neighborhood approach, verifying the precipitation of mesoscale convective systems. In this paper the FSS is considered over a square neighborhood as detailed in Roberts and Lean (2008); Roberts (2008). Duc et al. (2013) extend the FSS to include temporal and ensemble dimensions to give a single FSS value representative of the ensemble. A single field of fractions including spatial, temporal and ensemble information is then compared with observations. This is useful for providing an overview of model performance but does not provide information regarding the spread - skill relationship of the ensemble or the spatial differences between individual pairs of ensemble members.

Rezacova et al. (2009) use the FSS to calculate the ensemble spread-skill relationship with the ensemble skill calculated from the FSS between ensemble member- radar comparisons and the ensemble spread from the FSS between perturbed ensemble members and the ensemble control. Following on from this Zacharov and Rezacova (2009) determine a relationship between the FSS estimates of ensemble spread and skill and use this to predict the ensemble skill given the spread. Zacharov and Rezacova (2009) consider together FSS results from differently sized neighborhoods. This method was chosen because there is no fixed scale that can give a FSS skill value over different cases. However, as different physical behavior is apparent at different spatial scales (e.g as shown in Roberts (2008)) it is informative also to investigate how the ensemble spread varies with spatial scale which is the subject of this paper. Whereas Rezacova et al. (2009); Zacharov and Rezacova (2009) only consider comparisons between perturbed ensemble members and the control, in this paper the FSS between all independent member-member pairs is considered. Considering all members in this manner is the best representation of total spread as it includes fully the inter-member variability and does not rely on the ensemble mean which is known to lie outside the model manifold (Ancell 2013). Further work by the authors (in preparation, to be submitted shortly to Monthly Weather Review) considers other possible methods of member comparison.

Here we present the following: in Section 2 we introduce the two case studies that will provide examples throughout the paper. The model configuration is also discussed along with a justification for our method of using the FSS. Section 3 provides examples of our results for ensembles with different IC and LBC perturbations and results for different physics configurations are discussed in Section 4. Finally, in Section 5 we summarize the conclusions from this work and discuss areas of further investigation.

# 2. Method

## *a. Cases*

Two convective cases were chosen for the demonstrative purposes of this paper. In these cases convection occurs in different synoptic situations. The first case, 23 April 2011, was chosen as an example of organized spring convection over England and will be referred to as the 'organized Spring' case. This case has a low pressure system centered to the northwest of the UK and a high pressure system centered over Scandinavia. A frontal structure stretches down across the western UK. As the front moves eastward a convergence line forms across eastern England ahead of the front. This convergence line is shown in the UK Met Office analysis at 1800 UTC on the 23 April (Figure 1a). Convective storms developed in the vicinity of this convergence line with precipitation first seen at 1400 UTC on 23 April, and continuing until 0300 UTC on 24 April. At 1800 UTC a band of frontal precipitation enters the model domain from the NW preceding an occluded front which enters the domain at 0000 UTC on the 24 April.

The second case, 8 July 2011, features a number of convective storms that formed over the UK in an area of instability within the circulation of a decaying low pressure system. At 0600 UTC the low center was situated over Ireland as shown in Figure 1b. Throughout the day the low center then moved towards the northeast reaching the northeast of England by 1800 UTC. By 1400 UTC there were many heavy showers over Scotland as indicated by the Nimrod radar system (not shown). Convective clouds associated with these showers were also seen from visible satellite observations from the Meteosat geostationary satellite. For this case study we focus on one particular storm that formed over the Edinburgh area of eastern Scotland and remained stationary for around four hours producing large rainfall totals (0900 UTC – 2100 UTC radar derived precipitation totals of over 64 mm) and flooding. In future discussion this will be referred to as the 'flooding' case. Previous analysis of this case by Leoncini et al. (2011) showed that the Met Office 2.2 km ensemble on this occasion gave a 30% to 40% chance of a flood-producing storm within 25 km of Edinburgh; a level of significant risk.

### b. *Model Setup*

The UK Met Office Unified Model (MetUM) runs with a non-hydrostatic dynamical core with semi-Lagrangian advection (Davies et al. 2005). A comprehensive set of parametrizations are used 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). Version 7.7 of the global ensemble prediction system (MOGREPS-G) was run at a resolution of around 60km in the mid-latitude regions with 70 vertical levels. MOGREPS-G provided the initial conditions (ICs) and lateral boundary conditions (LBCs) for the North Atlantic and European (NAE) regional model run at 18km resolution with 70 vertical levels. Perturbations were generated using an ensemble transform Kalman filter 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 localization. Model error is addressed using the "random parameters" scheme for both ensembles to account for sub-grid processes uncertainty. Both the global and regional ensembles have 23 perturbed members and an unperturbed control.

For the case studies described here a high resolution ensemble, run over the Met Office variable resolution UK domain, was one way nested inside the NAE model. This domain has a constant resolution 2.2 km grid over the UK with the grid stretched up to 4 km around the domain edges to reduce the jump in resolution when downscaling from the NAE model. No further data assimilation was included when downscaling from the NAE to UK domain. The global and NAE models were run with a convection scheme based on Gregory and Rowntree (1990) but modified since (Derbyshire et al. 2011). The 2.2 km model has explicit convection only (no convection scheme). The 2.2 km UK domain is shown in Figure 2 in light gray and is approximately 920 km W-E by 1200 km N-S.

For the flooding case eleven perturbed members plus a control were run over the 2.2 m domain using LBCs and ICs taken from the first eleven members, and control, of the NAE regional ensemble (MOGREPS-R). Twelve simulations were run because this was the ensemble size being considered for an operational 2.2 m ensemble system (MOGREPS-UK, operational since 2013 (Mylne 2013)). To allow the flood producing storm over Edinburgh to be investigated, analysis for this case was also completed over a small 100 km domain surrounding this region. This subdomain is highlighted in Figure 2 in dark gray.

For the organized Spring case an ensemble of 8 MOGREPS simulations were run (seven perturbed members plus a control). This reduction in size allowed 5 different physics configurations to be considered for each MOGREPS simulation (giving a total of 40 simulations). The different model configurations were:

i. A control ensemble with the standard model settings labeled "standard".

ii. An ensemble with a restricted version of the convection scheme (Roberts 2003) as would be applied to the Met Office 4 km deterministic model (labeled "conv").

iii. An ensemble with the time step increased from 25 s to 50 s labeled "time". It is interesting to investigate the effects of a longer time step as increasing the time step reduces the computational cost of the simulation but may increase model error.

iv. An ensemble with increased time step and restricted convection scheme labeled "conv+time".

v. An ensemble with modifications to the graupel labeled "grp". The graupel modification allows the production of graupel through the capture of rain by snow and results in an increased graupel mass. This modification has become a standard option in Met UM versions 8.0 onwards (Wilkinson 2011).

It must be emphasized that these model configurations were chosen to demonstrate the methodology presented in this paper, not as possible implementations to the UK Met Office ensemble prediction system. Note also that the model variations are neither stochastic nor designed to represent the model error, although they do, nevertheless, represent plausible alternative formulations. The UK model for the organized Spring case was started at 0600 UTC on 23 April 2011, the flooding case at 1800 UTC on 7 July 2011. MOGREPS-G and MOGREPS-R were initiated 6 hrs and 3 hrs respectively before the UK model. For both cases the UK model was run up to lead times of 36 hours.

### *c. How the FSS is used*

The FSS is described in Roberts and Lean (2008) and summarized here for ease of reading. To calculate the FSS a threshold is first selected, say for precipitation, 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 converted to binary form with grid points set to 1 for values above the threshold and 0 otherwise. A neighborhood size is then selected and, for each neighborhood centered upon each grid point, the fraction of grid points with the value '1' within this square is computed. Two fields of fractions (denoted $A$ and $B$), say from a model and observations, are then compared using the mean squared error (MSE). For a neighborhood size $n$ and domain size $N_x$ by $N_y$ grid points this is given by:

$$MSE_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [A_{(n)i,j} - B_{(n)i,j}]^2. \quad (1)$$

The fractions skill score is computed by comparing $MSE_{(n)}$ with a reference MSE, $MSE_{(n)ref}$.

$$FSS_{(n)} = 1 - \frac{MSE_{(n)}}{MSE_{(n)ref}} \quad (2)$$

where $MSE_{(n)ref}$ is the largest possible MSE that can be obtained from fraction fields $A$ and $B$:

$$MSE_{(n)ref} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [A_{(n)i,j}^2 + B_{(n)i,j}^2]. \quad (3)$$

The FSS varies from 0 (complete mismatch between the fields) to one (perfect match between the fields).

Different neighborhood sizes are considered in order to evaluate the FSS at different spatial scales. Here we define the neighborhood size to be the total width of the square neighborhood in km. The smallest possible neighborhood is 2.2 km, set by the grid scale. No bias exists between the binary fields created using percentile thresholds as, by definition, the same number of points exceed the threshold for both fields. Hence, for percentile thresholds, the maximum possible spatial disagreement is found for two fields which place the points of interest at opposite edges of the domain. A perfect match is only obtained between fields with this maximum disagreement when they are compared over a neighborhood of twice the smallest dimension of the domain. In other words, the FSS will only equal 1 when the neighborhood size is equal to twice the smallest dimension of the domain. This sets the maximum neighborhood size for percentile thresholds. For value thresholds the fields may be biased and this argument does not hold. For the examples presented here only percentile thresholds are considered and the maximum neighborhood size is 1848 km for the UK domain and 200 km for the 100km subdomain.

The FSS can be calculated at a particular time between two different forecasts, or between a forecasts and observations, the former giving a measure of spatial spread, the latter of spatial skill. The ensemble spread is characterized by calculating the FSS for all independent member – member pairs ($N_p(N)$, for an ensemble of $N$ members) resulting in

$$N_p(N) = N \times (N-1)/2 \tag{4}$$

comparisons. Here, and for the remainder of this paper, the control is treated as an additional ensemble member. Hence, for the flooding case we have 12 MOGREPS members (the 11 perturbed members and unperturbed control) and for the organized Spring case we have 8 MOGREPS members for each physics configuration (the 7 perturbed members and unperturbed control). Justification for this method comes from our interest in the total spatial ensemble spread. In this situation the spatial location of a feature in the control forecast is not necessarily at the center of corresponding features in the perturbed members and therefore we do not wish to assign any special status to the control forecast. Figure 3 demonstrates the advantages of our method: when considering the control as an additional ensemble member one can distinguish the different spatial spread in Figures 3a and 3b, whereas when only comparing against the control the spread in Figures 3a and 3b is indistinguishable.

The ensemble skill is assessed by comparing the model hourly precipitation accumulations with those derived from the UK Met Office Nimrod radar system. The Nimrod system includes calibration against rain gauge data and aims to remove common sources of error (Golding 1998; Harrison et al. 2000). For the summer case 1 km Nimrod radar derived hourly precipitation accumulations are interpolated onto the 2.2 km model grid. Nimrod data at 1 km resolution was not available for analysis of the organized Spring case so 5 km data was instead used. The area of Nimrod coverage differs slightly from the UK 2.2 km domain over which the model is run and is indicated by the dotted region in Figure 2. All analysis involving radar data, or the comparison of model and radar data, only considers the area with radar coverage. We assume the radar data is representative of the precipitation that occurred and ignore observational errors, which would have to be considered within a routine verification framework. Visual examination of the radar fields found no obvious errors.

To assess ensemble skill each model simulation is separately compared with radar observations, whilst to assess ensemble spread we compare all possible pairings of the model runs. Again consider Figure 3, but this time take the filled black circles to represent the location of precipitation in the radar data. As a measure of ensemble skill we are only considering the spatial differences associated with the solid arrows. These measures of 'spread' and 'skill' consider different numbers of member-member or member-radar pairs, raising questions about a direct comparison of these metrics. However, answering these questions is not the subject of this paper which focuses on the characterization of spatial ensemble spread, with spatial ensemble skill considered only to put the spread into context. Further work by the authors (in preparation, to be submitted shortly to Monthly Weather Review) focuses in more detail on these metrics in the context of the spread-skill relationship.

Three different comparison strategies were used for the organized Spring case to characterize the differences between spatial spread in the MOGREPS ensemble and that produced through considering different physics configurations. 8 MOGREPS ensemble members ($N = 8$), and 5 different physics configurations ($N = 5$), were considered. Additionally results were produced using a subset of two physics configurations ($N = 2$) to allow spatial differences resulting from individual configurations to be investigated.

i. All independent comparisons were made between the MOGREPS members for a given physics configuration, with each physics configuration treated separately. Considering all 5 physics configurations in this manner gives $N_p(8) \times 5 = 140$ comparisons, a strategy denoted as MOGREPS5. Considering 2 physics configurations in this manner gives $N_p(8) \times 2 = 56$ comparisons, denoted as MOGREPS2.

ii. All independent comparisons between the different physics configurations for a given MOGREPS member, with each MOGREPS member treated separately. Considering all 5 physics configurations gives $8 \times N_p(5) = 80$ comparisons for this strategy denoted as Physics5. Considering 2 physics configurations gives $8 \times N_p(2) = 16$ comparisons (Physics2).

iii. Comparisons between different MOGREPS members which additionally have different physics configurations. For example, MOGREPS member 2 with the standard physics configuration might be compared with MOGREPS members 1,3,4,...,12 with the physics configurations conv, conv+time, time and grp. Considering all 5 physics configurations with this comparison strategy, referred to as MOGREPS5+Physics5, gives $N_p(8) \times N_p(5) = 280$ comparisons. Considering 2 physics configurations (MOGREPS2+Physics2), gives $N_p(8) \times N_p(2) = 28$ comparisons.

Given the large number of FSS values $FSS_i$ (one calculated for each comparison) it is necessary to consolidate this information to provide an overview of spatial ensemble behavior. In this paper the mean is taken over the relevant set of $FSS_i$. When calculated over member-member pairs this is referred to as dFSSmean where 'd' indicates that this is a measure of ensemble dispersion. When calculated over member-radar pairs this is referred to as eFSSmean where 'e' indicates that this is a measure of ensemble error. dFSSmean gives an indication of the average spatial agreement within the ensemble for a given neighborhood size. In other words, we can select a level of spatial agreement for the ensemble, represented by the value of dFSSmean, and ask at what neighborhood size this agreement is obtained.

As the ensemble members do not necessarily have an even spatial distribution, a range of $FSS_i$ will be obtained from the different ensemble member-member pairs. For example, if the majority of ensemble members place rain at the same spatial location but a small number of members place the rain far away this may produce a similar value of dFSSmean as a situation in which all ensemble members place the rain at slightly different spatial locations. Hence it is also important to investigate the range of FSS values surrounding dFSSmean. To do this the standard deviation of FSS values, dFSSstdev, is used. dFSSstdev is closely linked to the standard error in dFSSmean, $\frac{dFSSstdev}{\sqrt{N_{FSS}}}$ where $N_{FSS}$ is the number of $FSS_i$ samples used to calculate dFSSmean. As the purpose of this paper is to focus on the spatial distribution of ensemble members, we consider dFSSstdev and avoid the $\frac{1}{\sqrt{N_{FSS}}}$ dependence on ensemble size. This allows the spatial distribution of differently sized ensembles to be compared.

In order to make a spatial comparison between different ensembles it is necessary to find scales which are believable and have a reasonable level of spatial agreement. For the purposes of this paper, 'believable' scales for the intercomparison of ensemble members are derived in an equivalent manner to those scales that would be considered skillful if the comparison was instead against observations (assuming that the ensemble is well spread). This scale is quantified using the methodology of Roberts and Lean (2008) where a neighborhood size is considered believable ('skillful') if a FSS value of

$$FSS \geq 0.5 + \frac{f_0}{2} \tag{5}$$

is obtained for that neighborhood. $f_0$ is equal to fraction of the field considered in the FSS calculation (for example, considering the top $99^{th}$ percentile threshold would give $f_0 = 0.01$) and Equation 5 simplifies to an equality when the neighborhood is twice the spatial difference between two binary fields (Roberts and Lean 2008; Roberts 2008). As $f_0$ is small Equation 5 can be approximated as $FSS \geq 0.5$.

### *d. Thresholding*

The FSS can be calculated using either fixed value or percentile thresholds. Following on from the work of Roberts (2008); Mittermaier and Roberts (2010) this paper focuses on the use of percentile thresholds to allow the spatial distribution of phenomena to be investigated. Higher percentile thresholds are associated with smaller, more extreme forecast features, and lower percentile thresholds are associated with larger-scale smoother features (Roberts 2008). Note that here, and in all future discussion, the percentile threshold is applied over the whole domain, including areas both with and without precipitation.

To understand the effect of applying percentile thresholds it is informative to investigate the values corresponding to each threshold. Examples for hourly precipitation values corresponding to the $90^{th}$ and $99^{th}$ percentile thresholds are given in Figure 4. These percentile thresholds are used as examples throughout this paper. All ensemble members (gray solid lines) and radar (black lines) are shown for the organized Spring case (top) and Summer flooding case (bottom). From both cases and thresholds it can be seen that the radar percentile thresholds generally correspond to lower precipitation values than the model. This bias in the model compared to radar is an important consideration for model evaluation. However, it is also important to investigate the spatial distribution of precipitation: using percentile thresholds allows us to focus on this despite the model bias.

For the Spring case at the $90^{th}$ percentile threshold (Figure 4a) the radar values drop to zero after 16 hours. After this time radar derived precipitation covers less than 10% of the domain. This demonstrates that the $90^{th}$ percentile, and other percentile thresholds below the $90^{th}$, are not a suitable threshold for radar precipitation accumulations for this case. For all cases (apart from the unlikely event of 100% coverage) there will be a limited area covered by precipitation in both the model and observations, and a corresponding minimum suitable percentile threshold. In an operational situation this minimum threshold could easily be calculated from the fraction of precipitation coverage. All FSS results presented in this paper have been calculated using percentile thresholds above this minimum value.

For the Spring case the 8 MOGREPS members from the standard physics configuration are shown in dark gray in Figure 4a and Figure 4b and, although differing by up to 2.5 mm in accumulation values (for the $99^{th}$ percentile threshold), follow the same overall trend throughout the day. This suggests that the ensemble members produce precipitation features, such as that associated with frontal passage, at similar times. The simulations for all MOGREPS members and the other 4 physics configurations are shown in light gray with the different physics configurations clustering around the corresponding MOGREPS member. In these experiments the different physics configurations have little effect on the precipitation value corresponding to a given percentile threshold. Interestingly, Figure 4a and Figure 4b show peaks in precipitation values at different times: Figure 4a ($90^{th}$ percentile) at a lead time of 20 hrs and Figure 4b ($99^{th}$ percentile) at a lead time of 12 hrs. The higher threshold considers only the areas of convective precipitation, giving a corresponding value that peaks when these storms are strongest whereas the lower threshold includes frontal precipitation and peaks where this is heaviest.

The 12 members for the Summer flooding case are shown for thresholds calculated over the full UK domain (dark gray) and limited area domain (light gray). Beyond a lead time of 15 hours, when convection occurred over Edinburgh, values for the limited domain are up to 5 times larger than those over the UK domain. Considering this area separately using percentile thresholds allows the flood producing storm to be investigated. It should be noted that using high value thresholds over the UK domain would also select the Edinburgh area. However, for this highly variable case some ensemble members missed the convection over Edinburgh, and do not produce sufficiently high precipitation values. It is not possible to choose a value threshold that is high enough to select only the area of convection, and yet low enough to include all the ensemble members. Again, this demonstrates the utility of percentile thresholds.

# 3. Results for LBC and IC perturbations

### *a. dFSSmean and eFSSmean*

First we consider the realism of the spatial ensemble spread by comparing dFSSmean and eFSSmean for both cases. Both dFSSmean and eFSSmean were calculated over the section of the 2.2 km UK domain with radar coverage (highlighted by the dotted region in Figure 2). Figure 5 shows dFSSmean (left) and eFSSmean (right) for the organized Spring case (top) and flooding case (bottom) calculated for the $99^{th}$ percentile threshold over the whole UK domain. These results were computed for the 12 members of the flooding cases and 8 MOGREPS members with standard physics for the organized Spring case. To check the validity of comparing these differently sized ensembles, results were also produced for the flooding case when only considering the first 8 ensemble members (not shown). These 8 member results differed only in small details from those calculated from 12 members, and lead to the same conclusions, so it was decided to show the results from the full 12 member comparisons.

Comparison of the dispersion measures (dFSSmean) for the two cases (Figures 5a and 5c) shows that, although these cases are synoptically different, with different convective forcing, the overall behavior is broadly similar. At small scales ensemble members are very different resulting in low values of FSS. FSS values increase as the members become more similar when considered at larger scales. The temporal variability present in the ensemble spread, as measured by dFSSmean, is also clear at this threshold with the scale at which $FSS = 0.5$ varying between 50-500 km for the organized Spring case and 100-250 km for the flooding case. These scales are large because in both cases there is considerable uncertainty in the locations of the showers and showery areas. The temporal variability can be related to the evolution of physical processes. For example, in Figure 5a the area of larger ensemble spread (decrease in dFSSmean) at lead times 13-20 hrs can be linked to greater convective activity and the highest rainfall instances (compare with Figure 4b) and the increase in dFSSmean (lower spread) from 20-25 hrs can be related to a area of spatially predictable frontal precipitation moving into the domain.

Overall there is less temporal variability in the FSS for the flooding case. This can again be related to the meteorology of the cases: precipitation in the flooding case was the result of one mechanism, instability associated with a decaying low pressure system, whereas precipitation in the Spring case was associated with both convective showers and frontal passage. Coincidentally, for both cases, the spatial ensemble spread increases with forecast lead time after 20 hours. This up-scaling of forecast spatial differences should be expected from a statistical evaluation of a large number of cases, but not necessarily from individual case studies where the physical processes of the day dominate. Using dFSSmean for individual case studies allows these processes, and their effect on the spatial ensemble spread and upscale growth of forecast differences, to be examined.

The error measures (eFSSmean, Figures 5b and 5d) show a similar structure to the dispersion measures with a similar magnitude for ensemble spread and skill. There are times, such as for the Spring case at a lead time of 20 hrs (Figure 5b), or the flooding case at lead times 0-5 hrs (Figure 5d) when the ensemble is clearly under-spread. For the Spring case a timing error results from a front passing into the domain in all members earlier than seen in the radar; for the flooding case convective showers present in the radar have yet to spin up in the model. In both cases there is little evidence that the ensemble is over-spread.

For the flooding case dFSSmean and eFSSmean have also been calculated over the 100 km limited area domain containing the flooding event. Selecting a subdomain in this manner allows us to focus on the spatial predictability of a specific event which can be very different from the UK domain averaged results. Differences between the domains can also be seen in the values corresponding to each percentile threshold as discussed in Section 2d. dFSSmean and eFSSmean, calculated over the 100km domain are shown in Figures 6a and 6b respectively at forecast lead times 17 hrs - 26 hrs when convection was seen over Edinburgh. Comparison of Figure 6a and Figure 6b suggests that the ensemble spread and skill are similar and that, over this area, the ensemble is capturing the spatial variability of the precipitation well. This gives confidence in the ensemble for a spatially unpredictable flooding event. There are some differences between dFSSmean and eFSSmean, in particular that eFSSmean is more variable with time. This may be partly due to both the smaller number of comparisons in the error calculation, and also reflects differences between the model and observations in the temporal evolution of the storm. Note that, as the $99^{th}$ percentile threshold corresponds to different precipitation values over the UK and Edinburgh domains, we cannot do a direct comparison between Figures 5 and 6. This also suggests that we are indeed looking at different processes or phenomena with the different domains and confirms the need to use a suitable domain size to examine the spatial variability of particular features. The domain must be large enough to give representative results, but small enough to focus on the phenomena of interest. Of course, the same remarks will be true of any spatial measure.

### *b. dFSSstdev in addition to dFSSmean*

In this section we discuss the benefits of considering dFSSstdev in addition to dFSSmean. Figure 7 shows dFSSmean and dFSSstdev calculated for the organized Springcase (top) and flooding case (bottom) when considering the $99^{th}$ percentile threshold for hourly precipitation accumulations. The FSS was calculated over the whole UK domain. dFSSstdev is shown in Figure 7c and Figure 7d and presents results consistent with those from dFSSmean. For example, the largest values of dFSSstdev occur in areas where low dFSSmean values extend to large scales. The greater spatial spread associated with low values of dFSSmean results in a wider range of possible values for $FSS_i$ and larger dFSSstdev.

However, there is also some further information given by the standard deviation. In particular, for the flooding case (Figure 7d) there is an area of higher standard deviation seen in the first two hours of the forecast at neighborhood sizes up to 500km which is associated with the spin-up of the model. This effect is even more apparent in results for the $99.9^{th}$ percentile threshold (not shown) and is the result of the convection permitting model having to spin up showers during the first few hours of the forecast. Because the ensemble members spin-up showers at different locations, lower values of dFSSmean and a large range of values of $FSS_i$ (resulting in a large dFSSstdev) are obtained. For the spring case (Figures 7a and 7b) convective showers are not present at the forecast start time and do not need to be spun-up from initial conditions. Hence spin-up effects are not seen in the precipitation diagnostics. It is useful to examine how the standard deviation behaves at different scales. The smallest values are found at both the grid scale, where differences are so large that similarly low values of the FSS are expected for all member pairs, and also at the largest scales, where all members are effectively the same.

### *c. Other fields and thresholds*

The use of different percentile thresholds allows more information to be gained about the ensemble spread for different ranges of forecast values, for example a higher threshold will select more extreme values compared to a lower threshold which will select values that are more widespread. An example is given in Figure 8 for the organized Spring case where results for the top $99^{th}$ (LHS) and $85^{th}$ (RHS) percentiles are compared. This time we show a different diagnostic field, the 10 m horizontal wind speed. Like the hourly precipitation accumulations this field was selected as a suitable candidate for calculation of the FSS because of its high spatial variability. 10 m wind speeds are also used by the Met Office for routine forecast verification.

The $99^{th}$ percentile threshold selects only the highest wind speeds in the domain. At lead times 0-10 hrs these are found in to areas to the north of the UK near the low pressure center. The exact placement of the highest winds varied considerably between the ensemble members, with some placing them to the northwest and others to the northeast of the UK. Hence there were large spatial differences between the members resulting in low dFSSmean values extending to large neighborhoods at a lead time of 10 hrs as shown in Figure 8a. At lead times greater than 10 hours there is high spatial agreement amongst the ensemble members resulting in high values of dFSSmean. All members place the highest winds to the northwest of the UK associated with the frontal feature that enters the domain at this time.

Comparing Figure 8a and Figure 8b we see the unusual result that for a lead time of 12 hours, and after 28 hours, there is more agreement (larger FSS values) for the $99^{th}$ than for the $85^{th}$ percentile for a given neighborhood size. This behavior suggests that care must be taken in the interpretation the $99^{th}$ percentile threshold for the wind speed field. For the wind speed, local variability is superimposed upon a background gradient from the large scale situation. The $99^{th}$ percentile is likely to include both local variability from points where the background field is moderate and also larger scale variability where the background field is high. Consequently, unlike for precipitation, we cannot cleanly examine local features in the wind speed field simply by selecting a high threshold value. It is necessary to also consider a lower threshold that includes features of the larger scale flow such as, for this case, the $85^{th}$ percentile threshold. Figure 8b shows that, at lead times 12-20 hrs, the FSS values for the $85^{th}$ percentile are particularly high. These areas of small spatial spread can be related to the synoptic situation: at a lead time of 12 hrs a highly predictable frontal feature entered the domain from the NW and the top 15 % of wind speeds in the domain were closely associated with the flow in the vicinity of this front. Hence, there was very high spatial agreement between the members at these times. Before the front entered the domain the highest winds were associated with a less predictable decaying cold front. Moreover, after the front had progressed further into the domain greater differences between the members emerged at larger scales for the winds to the south of the occluded front.

The effect of different thresholds on the FSS for hourly precipitation accumulations can be seen by comparing Figures 5a and 5c with Figures 9a and 9b respectively. The latter show dFSSmean calculated for the $90^{th}$ percentile threshold. In particular, it can be seen that the large temporal variability seen in Figures 5a and 5c for the $99^{th}$ threshold has been replaced in the $90^{th}$ percentile results by a trend for ensemble spread to increase systematically with time. This trend is expected climatologically as forecast differences grow from small to larger scales with increasing forecast lead time. The rate of increase is different for the two cases. For the flooding case (Figure 9b) scales at which dFSSmean=0.5 increase gradually from 5 km to 100 km over 36 hours as forecast differences grow from small to larger scales. For the Spring case, dFSSmean values greater than 0.5 are seen even at the grid scale for lead times up to 25 hrs. After this time the scale at which dFSSmean=0.5 increases rapidly to 225km. This pattern is in agreement with the behavior seen for the $99^{th}$ threshold and has the same interpretation: after 25 hrs an area of precipitation moves out of the domain but with timing differences between the members. Overall, there is better spatial agreement between the ensemble members at the $90^{th}$ percentile threshold than at the $99^{th}$: the broader-scale features selected by the lower threshold are more predictable. When considering a range of different thresholds from the $99^{th}$ to $80^{th}$ percentile (not shown) the transition from large temporal variability to a trend of upscale growth of forecast differences with increasing lead time was found to be smooth: there is no sudden transition. It is likely that the range of thresholds over which such a transition occurs will be highly case dependent as the relative importance of local and large scale features changes. The FSS allows such behavior to be investigated.

# 4. Results assessing different physics configurations

In this section we present an application of dFSSmean to the comparison of the multiphysics and MOGREPS ensembles for the organized Spring case. Thus we compare the spatial ensemble spread associated with LBC and IC perturbations to that generated through different physics configurations as described in Section 2c. The examples presented are for the $99^{th}$ percentile threshold of precipitation accumulation: lower thresholds showed smaller spatial differences (larger dFSSmean values) but lead to the same general conclusions. Note that the purpose is not to evaluate the merits of particular physics configurations but to show a method that can be used to examine the behavior of stochastic processes or physics changes in ensembles.

Figure 10b shows dFSSmean comparing the configuration with restricted convection scheme and increased time step (conv+time) to that with the modified treatment of graupel (grp) using the Physics2 comparison strategy (comparison strategy ii in Section 2c). This comparison strategy is shown because it gives larger spatial differences than those found when comparing any other physics configuration pairs, or considering all physics configurations (the Physics5 comparison strategy). In Figure 10b FSS values of 0.5 are reached by a neighborhood size of 5 km, and no spatial differences are seen for neighborhoods greater than 100 km (where FSS=1). The lowest values of dFSSmean occur between lead times of 12 hrs and 16 hrs when the heaviest convective showers were present: it is during these events that modifications to the treatment of graupel are most noticeable.

Results from comparing only the MOGREPS members from conv+time and grp (comparison strategy MOGREPS2, i in Section 2c) are shown in Figure 10a. These differ only in minor details from those shown in Figure 7a (dFSSmean calculated for the MOGREPS ensemble with the standard physics configuration). The MOGREPS2 results show that FSS values of 0.5 are reached on scales greater than 60 km, scales at which the Physics2 members are almost identical. In other words, the spatial variation introduced through different physics configurations is only seen close to the grid scale. If we consider FSS values lower than $FSS = 0.5$ to represent fields so different that the forecast is no longer useful, then the different physics configurations applied here, for this particular case, are simply moving around features that are known to be unpredictable from the MOGREPS ensemble. Of course, this is not to say that physics changes in general are unimportant for improving model performance, or that using different physics configurations is not sometimes a valuable component of an ensemble system, or that adding small scale perturbations is undesirable or that, for another case or for other physics perturbations the effects might be very different. Our purpose is simply to demonstrate a methodology that allows the spatial effects of different ensemble configurations to be thoroughly investigated and set into the context of other aspects of forecast uncertainty.

It is possible that, although the evaluation of Physics2 only showed forecast differences at small spatial scales, combining the different physics configurations with those from the MOGREPS2 ensemble may lead to large changes in the growth of spatial differences. To assess this, the comparison strategy MOGREPS2+Physics2 (comparison strategy iii in Section 2c) is employed. Again, examples are shown for the physics configurations conv+time and grp which show the largest spatial differences. The results of MOGREPS2+Physics2 are shown in Figure 10c. Differences between Figure 10c and Figure 10a are very small and hence, to aid interpretation, Figure 10d shows the difference between the MOGREPS2 and the MOGREPS2+Physics2 results. The differences are over an order of magnitude smaller than the dFSSmean values in Figures 10a and 10c. It is interesting that both positive and negative differences are seen: modifying the different physics configuration both adds and removes spatial spread. From Figure 10d it can also be seen that differences between MOGREPS2 and the MOGREPS2+Physics2 extend, with similar magnitude, across all spatial scales. However, in terms of the fractional difference relative to dFSSmean the differences at small neighborhoods have more importance. At a lead time of 15 hrs the fractional difference in dFSSmean varies from 7% at 50 km to 3% at 250 km. It should be noted that these differences are still very small, especially at the larger more predictable scales (as indicated by the point where FSS$\geq$0.5 in the MOGREPS ensemble).

Analysis of the combined MOGREPS+Physics comparisons supports the conclusions drawn previously that the introduction of these differences in the physics only influences scales much smaller than the predictable scales of the system (in this particular experiment). In practical terms, the variability of those scales could be addressed with spatial post processing and without the need for additional ensemble members. On the other hand, if the scales of the physics changes were to upscale to scales greater then the system's predictable scales then the performance of the ensemble might benefit from more perturbed-physics members. Systematic application of the methods shown here would provide a sound basis for making these decisions.

# 5. Discussion and conclusions

In this paper we have presented, with examples, a new methodology for the detailed analysis of ensemble spread for high resolution forecasts focusing on spatial variability. In particular we focused on two different measures of ensemble spread: dFSSmean and dFSSstdev, the mean and standard deviation of the FSS calculated over all ensemble member-member pairs. dFSSmean gives a measure of the FSS value for the whole ensemble indicating the average spatial agreement within the ensemble over a particular size of neighborhood i.e at a given spatial scale. dFSSstdev provides some further useful information about the range of FSS values used in the calculation of dFSSmean. A large range of FSS values, corresponding to a large value of dFSSstdev, indicates that the ensemble members are unevenly distributed.

To demonstrate the utility of these measures results were presented from two case studies. It was shown that dFSSmean and dFSSstdev allowed investigation of, for example, the temporal evolution of ensemble spread, model spin up, and saturation of forecast differences. Considering different percentile thresholds allowed information to be gained about the spatial spread of the ensemble for different physical regimes. In particular it was found that, for hourly precipitation accumulations, the dFSSmean for the $99^{th}$ percentile threshold had high temporal variability. This contrasted with the dFSSmean for the $90^{th}$ percentile threshold for which spatial differences between the ensemble members increased with time.

The realism of the ensemble spatial distribution was also tested by comparison with another metric, the mean FSS calculated over all member-radar pairs, denoted eFSSmean. This error measure can be compared with dFSSmean to investigate the spread-skill relationship of the ensemble at different times and spatial scales. For the two cases considered here these measures suggested that ensemble spread was reasonable. On occasion the ensemble was under-spread and this was linked to timing errors between the simulations and the observations and to the need for spin up of showers in a convection permitting model.

For one case study, results were presented for a comparison of spread between differently generated ensembles, including multiple physics configurations. This application illustrates a methodology for identifying the spatial scales that are influenced by modifications to physical processes. Examining the FSS for different spatial scales and over a range of times allowed a quantification of the effects of using different physics configurations compared to LBC and IC perturbations. For the case described here it was concluded that modifying the physics for this case did not influence the ensemble evolution at scales where the forecast has skill. These results are not to be interpreted as general: well chosen physics modifications can and do improve forecasts as demonstrated by, for example by Stensrud et al. (2000); Keil et al.

(2013). The key point is that evaluation techniques presented here allow clear statements about the impacts of physics modifications to be made since different ensemble configurations can be thoroughly investigated and the spatial impact of the changes quantified.

The work presented here provides a framework through which spatial ensemble spread can be analyzed. There are some limitations to this study: in particular the consideration of two cases only and the limited consideration of physics perturbations. It is left to future work to apply these methods to a larger sample of cases, and different, more realistic, multiphysics ensembles or other model error inclusion schemes. Another limiting factor is the methodology of calculating a single value of the FSS that is representative of a comparison across a whole domain. As discussed above this can mean that different meteorological phenomena, such as convective and frontal precipitation, are compared together, when each individually may have an inherently different predictability and ensemble spread. It is possible to select a smaller domain to consider events of interest, as highlighted with respect to Figure 6, although this is only useful in hindsight once the event has occurred. Hence, future work is intended to develop a spatially varying and scale dependent measure of ensemble spread that does not suffer from this drawback.

Despite these limitations there are some important conclusions from this work. In particular, we have stressed how the ensemble spread is highly dependent on the scales considered for evaluation. Consequently, to investigate the ensemble behavior thoroughly it is necessary to consider multiple scales, and be mindful of the different expectations for skill at these scales. Forecasts should be verified, and the benefits of forecast model changes assessed, at scales that are believable. This paper has provided a methodology for determining such believable scales and their temporal evolution. With future movement to higher and higher resolution models the distinction between the grid scale and the believable scales is becoming increasingly important.

*Acknowledgments.*

The authors would like to thank the three anonymous reviewers for their detailed comments that have helped improve the quality and clarity of this article. S.Dey acknowledges support from a NERC PhD studentship with CASE support from the UK Met Office. Initial work contributing to this paper was completed by the same author during a Summer Placement at MetOffice@Reading. S.Migliorini acknowledges support from the NERC National Center for Earth Observation.

# REFERENCES


Ancell, B. C., 2013: Nonlinear characteristics of ensemble perturbation evolution and their application to forecasting high-impact events. *Wea. Forecasting*, **28 (6)**.

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

Berner, J., S.-Y. Ha, J. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139 (6)**, 1972–1995.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc*, **135 (640)**, 767–776.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc*, **134 (632)**, 703–722.

Caron, J.-F., 2013: Mismatching perturbations at the lateral boundaries in limited-area ensemble forecasting: A case study. *Mon. Wea. Rev.*, **141 (1)**, 356–374.

Clark, A. 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.

Craig, G. C., C. Keil, and D. Leuenberger, 2012: Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection. *Quart. J. Roy. Meteor. Soc*, **138 (663)**, 340–352.

Davies, T., M. J. P. Cullen, A. J. Malcolm, M. H. Mawson, A. Staniforth, A. A. White, and N. Wood, 2005: A new dynamical core for the met office's global and regional modelling of the atmosphere. *Quart. J. Roy. Meteor. Soc*, **131 (608)**, 1759–1782.

Derbyshire, S., A. Maidens, S. Milton, R. Stratton, and M. Willett, 2011: Adaptive detrainment in a convective parametrization. *Quart. J. Roy. Meteor. Soc*, **137 (660)**, 1856–1871.

Done, J. M., G. C. Craig, S. L. Gray, and P. A. Clark, 2012: Case-to-case variability of predictability of deep convection in a mesoscale model. *Quart. J. Roy. Meteor. Soc*, **138 (664)**, 638–648.

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

Duda, J. D. and W. A. Gallus, Jr, 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28 (4)**.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorol. Appl.*, **15 (1)**, 51–64.

Edwards, J. M. and A. Slingo, 1996: Studies with a flexible new radiation code. I: Choosing a configuration for a large-scale model. *Quart. J. Roy. Meteor. Soc*, **122 (531)**, 689–719.

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Z.*, **6**, 147–183.

Essery, R., M. Best, and P. Cox, 2001: MOSES 2.2 technical documentation. Tech. rep., Hadley Centre Technical Note.

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

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

Golding, B. W., 1998: Nimrod: a system for generating automated very short range forecasts. *Meteorol. Appl.*, **5 (1)**, 1–16.

Gregory, D. and P. R. Rowntree, 1990: A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Mon. Wea. Rev.*, **118 (7)**, 1483–1506.

Hacker, J., C. Snyder, S.-Y. HA, and M. Pocernich, 2011: Linear and non-linear response to parameter variations in a mesoscale model. *Tellus A*, **63 (3)**, 429–444.

Hanley, K., D. Kirshbaum, N. Roberts, and G. Leoncini, 2013: Sensitivities of a squall line over central europe in a convective-scale ensemble. *Mon. Wea. Rev.*, **141 (1)**, 112–133.

Hanley, K. E., D. J. Kirshbaum, S. E. Belcher, N. M. Roberts, and G. Leoncini, 2011: Ensemble predictability of an isolated mountain thunderstorm in a high-resolution model. *Quart. J. Roy. Meteor. Soc*, **137 (661)**, 2124–2137.

Harrison, D. L., S. J. Driscoll, and M. Kitchen, 2000: Improving precipitation estimates from weather radar using quality control and correction techniques. *Meteorol. Appl.*, **7 (2)**, 135–144.

Hohenegger, C., D. Lüthi, and C. Schär, 2006: Predictability mysteries in cloud-resolving models. *Mon. Wea. Rev.*, **134 (8)**, 2095–2107.

Hohenegger, C. and C. Schär, 2007a: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.*, **88 (7)**, 1783–1793.

Hohenegger, C. and C. Schär, 2007b: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64 (12)**, 4467–4478.

Johnson, A. and X. Wang, 2013: Object-based evaluation of a storm-scale ensemble during the 2009 NOAA hazardous weather testbed spring experiment. *Mon. Wea. Rev.*, **141 (3)**, 1079–1098.

Johnson, A., et al., 2014: Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Mon. Wea. Rev.*, **142 (3)**, 1053–1073.

Keil, C. and G. C. Craig, 2011: Regime-dependent forecast uncertainty of convective precipitation. *Meteorol. Z.*, **20 (2)**, 145–151.

Keil, C., F. Heinlein, and G. C. Craig, 2013: The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quart. J. Roy. Meteor. Soc.*, doi: 10.1002/qj.2143.

Kühnlein, C., C. Keil, G. C. Craig, and C. Gebhardt, 2013: The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quart. J. Roy. Meteor. Soc.*, doi:10.1002/qj.2238.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136 (9)**, 3408 – 3424.

Leoncini, G., R. S. Plant, S. L. Gray, and P. A. Clark, 2010: Perturbation growth at the convective scale for CSIP IOP18. *Quart. J. Roy. Meteor. Soc*, **136 (648)**, 653–670.

Leoncini, G., R. S. Plant, S. L. Gray, and P. A. Clark, 2012: Ensemble forecasts of a flood-producing storm: comparison of the influence of model-state perturbations and parameter modifications. *Quart. J. Roy. Meteor. Soc*, **139 (670)**, 198–211.

Leoncini, G., N. Roberts, and B. Golding, 2011: 8th July 2011 Floods in Scotland. report for Scottish Environment Protection Agency, Met Office, 18 pp.

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

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21 (3)**, 289–307.

Martin, W. J. and M. Xue, 2006: Sensitivity analysis of convection of the 24 May 2002 IHOP case using very large ensembles. *Mon. Wea. Rev.*, **134 (1)**, 192–207.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83 (3)**, 407–430.

Migliorini, S., M. Dixon, R. Bannister, and S. Ballard, 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. and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25 (1)**, 343–354.

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, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, **63 (2)**, 71–116.

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

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

Roberts, N. M., 2003: The impact of a change to the use of the convection scheme to high resolution simulations of convective events. Forecasting Research Technical Report 407, Met Office, 30 pp. URL `http://www.metoffice.gov.uk/archive/forecasting-research-technical-report-407`, available from the National Meteorological Libray and Archive.

Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136 (1)**, 78–97.

Schwartz, C. S., et al., 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25 (1)**, 263–280.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128 (7)**, 2077–2107.

Surcel, M., I. Zawadzki, and M. K. Yau, 2014: On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Mon. Wea. Rev.*, **142 (3)**, 1093–1105.

Vié, B., G. Molinié, O. Nuissier, B. Vincendon, V. Ducrocq, F. Bouttier, and E. Richard, 2012: Hydro-meteorological evaluation of a convection-permitting ensemble prediction system for mediterranean heavy precipitating events. *Nat. Hazards Earth Sys.*, **12 (8)**, 2631–2645.

Walser, A., D. Lüthi, and C. Schär, 2004: Predictability of precipitation in a cloud-resolving model. *Mon. Wea. Rev.*, **132 (2)**, 560–577.

Walser, A. and C. Schär, 2004: Convection-resolving precipitation forecasting and its predictability in alpine river catchments. *J. Hydrol. (Amsterdam, Neth.)*, **288 (12)**, 57 – 73.

Wilkinson, J., 2011: The large-scale precipitation parametrization scheme. Unified Model Documentation Paper 26, Unified Model Version: 8.0, Met Office, 53 pp.

Wilson, D. R. and S. P. Ballard, 1999: A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Quart. J. Roy. Meteor. Soc*, **125 (557)**, 1607–1636.

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

Zhang, F., 2005: Dynamics and structure of mesoscale error covariance of a winter cyclone estimated through short-range ensemble forecasts. *Mon. Wea. Rev.*, **133 (10)**, 2876–2893.

Zimmer, M., G. C. Craig, C. Keil, and H. Wernli, 2011: Classification of precipitation events with a convective response timescale and their forecasting characteristics. *Geophys. Res. Lett.*, **38**, doi:10.1029/2010GL046199.

# List of Figures

(a) (b)

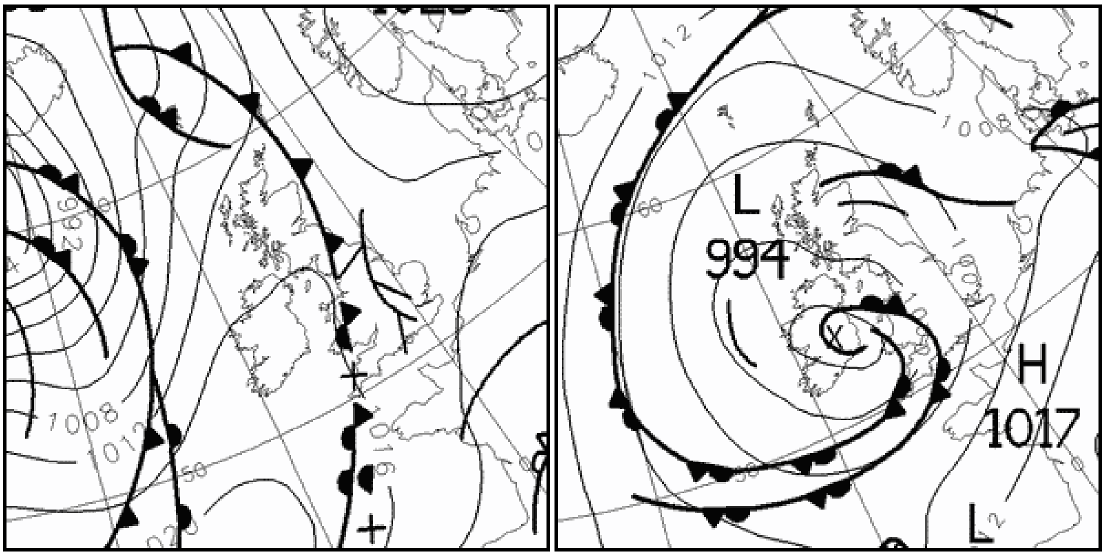


Fig. 1. UK Met Office surface analysis valid at (a) 18 UTC on 23 April 2011 and (b) 06 UTC on 8 July 2011. Courtesy of the Met Office. Crown copyright.

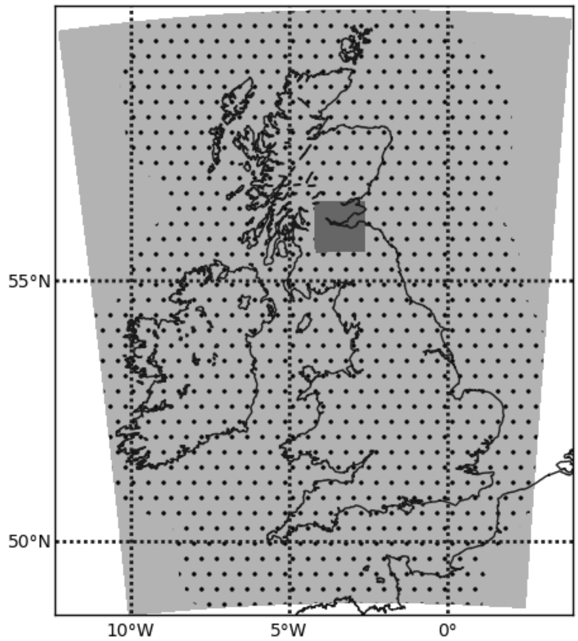


FIG. 2. Domains of the UK 2.2 km model (light gray), 100 km subdomain for the Summer flooding case (dark gray) and areas of radar coverage (dotted).

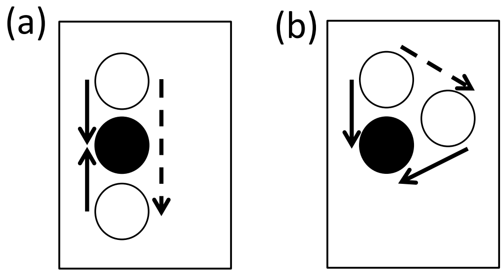


Fig. 3. Two different idealized spatial distributions of precipitation. Individual ensemble members (shown in white) position the precipitation in different spatial locations. The control simulation (shown in filled black) may produce precipitation in the center of that produced by individual ensemble members as shown in (a) or at the edge of the ensemble as shown in (b). Considering only the spatial separation of member-member pairs (solid arrows) indicates that (a) and (b) have the same spatial ensemble spread. Including both member-control and member-member pairs allows the differences in spread between (a) and (b) to be detected.

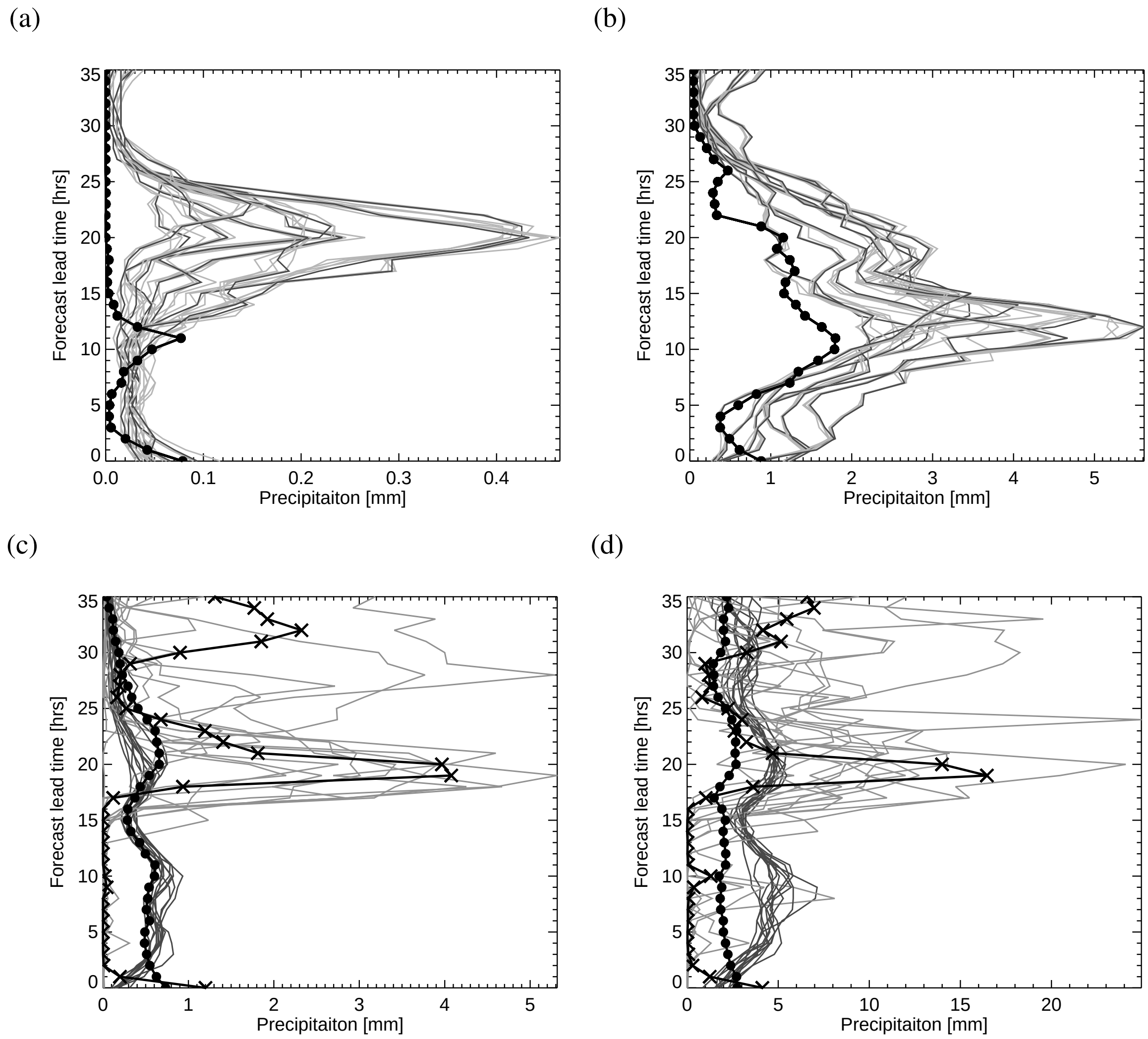


FIG. 4. Hourly precipitation accumulation values corresponding to the $90^{th}$ (a,c) and $99^{th}$ (b,d) percentile thresholds. For the organized Spring case (top) results from all the simulations are shown. To highlight the grouping of members those with the standard physics configuration are shown in dark gray and those from other physics configurations in light gray. For the flooding case (bottom) percentile thresholds calculated using data for the full the UK domain are shown in dark gray, and those for the limited area domain are shown in light gray. Radar data is shown from the area of the UK domain with radar coverage (black with circles) and, in (c,d) over the limited area domain (black with crosses).

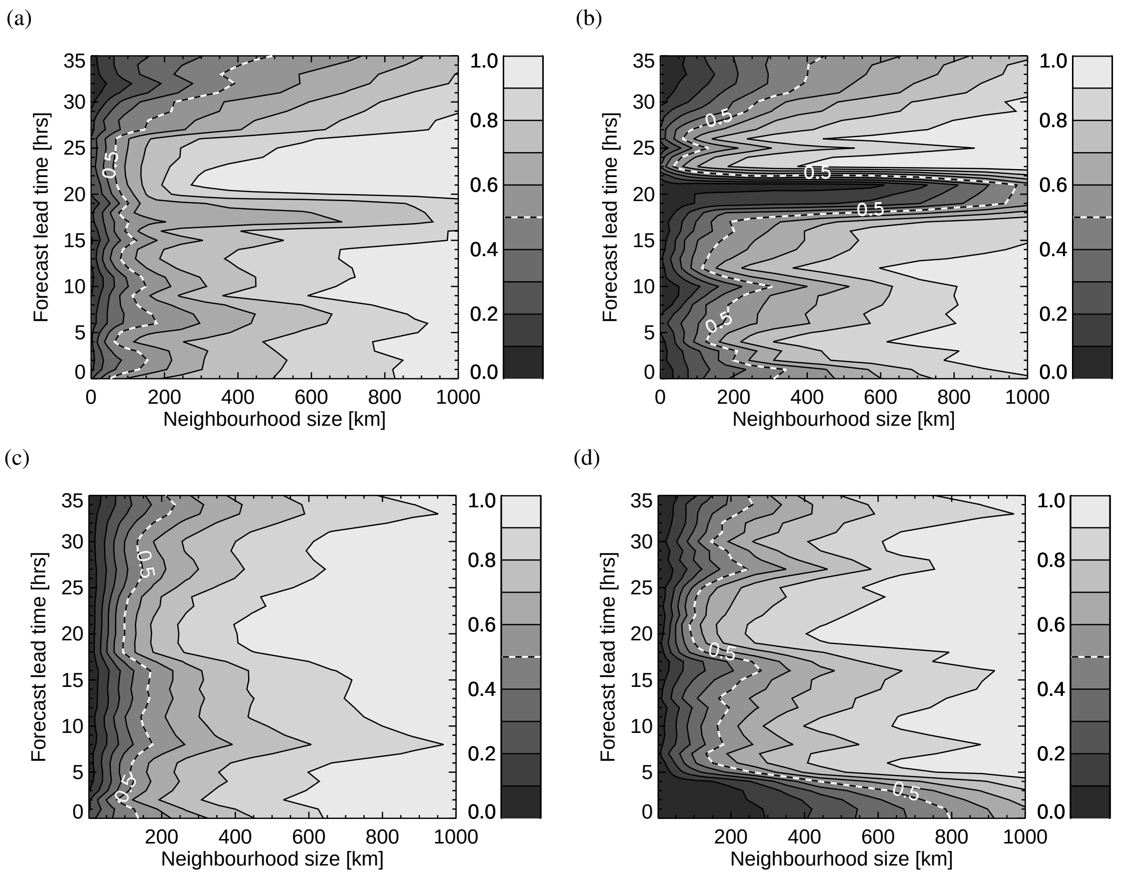


Fig. 5. dFSSmean (a,c) and eFSSmean (b,d) for the organized Spring case (top) and the Summer flooding case (bottom). The standard physics configuration and the $99^{th}$ percentile threshold are considered. The white dashed line at 0.5 represents the believable scale. Results were calculated over the area of the UK domain with radar coverage.

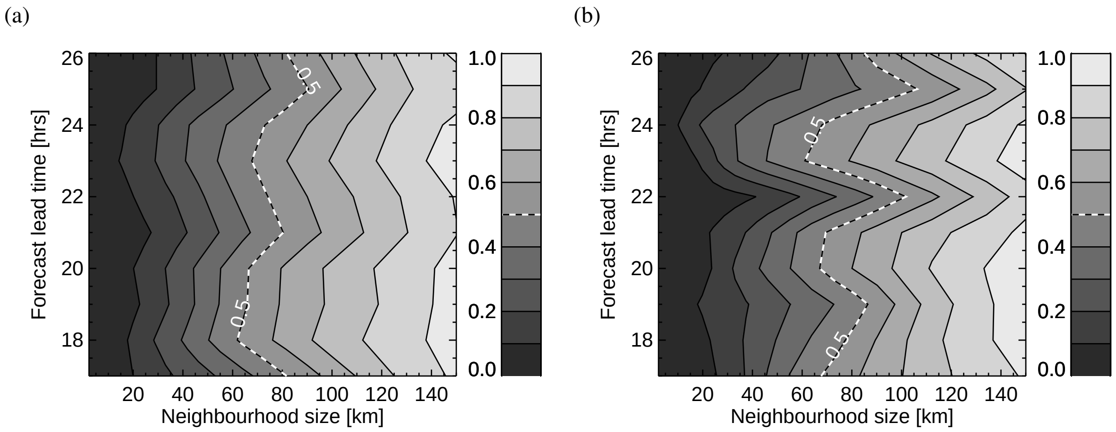


FIG. 6. FSS calculations over the Edinburgh subdomain: (a) dFSSmean and (b) eFSSmean. The $99^{th}$ percentile threshold is considered. The white dashed line at 0.5 represents the believable scale.

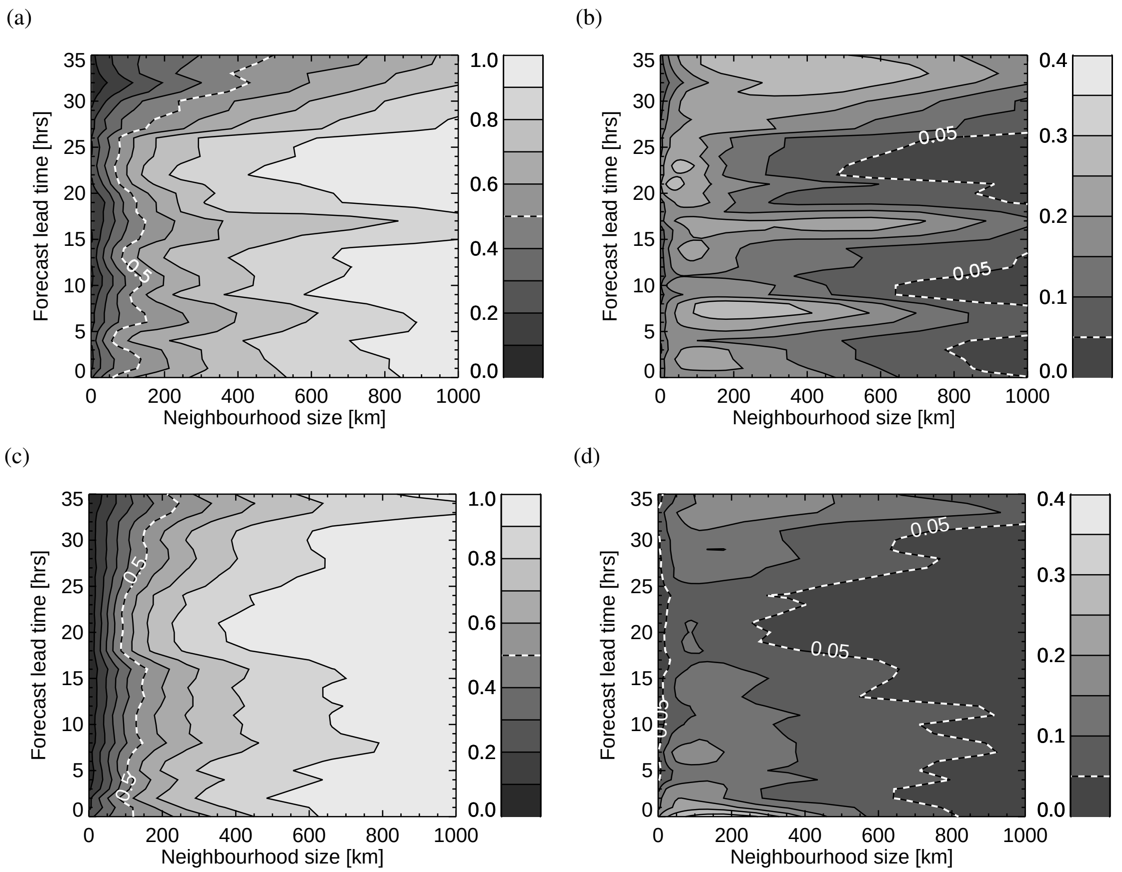


FIG. 7. dFSSmean (a,c) and dFSSstdev (b,d) for the organized Spring case (top) and the flooding case (bottom). The white dashed line in (a,c) at 0.5 represents the believable scale. To guide the eye, in (b,d) the white dashed line at 0.05 represents the neighborhood at which dFSSstdev is an order of magnitude smaller than the believable scale. The $99^{th}$ percentile threshold is considered and results are calculated over the whole UK domain.

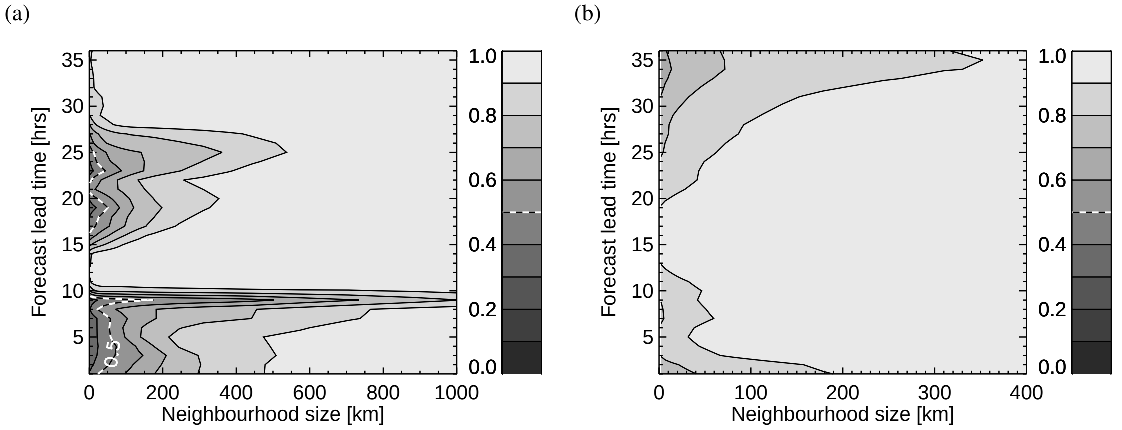


FIG. 8. Comparison of dFSSmean calculated for the (a) $99^{th}$ and (b) $85^{th}$ percentile thresholds for the 10 m horizontal wind speed field and the organized Spring case. Results are calculated over the whole of the UK domain and only the standard physics configuration is considered. The white dashed line at 0.5 represents the believable scale.

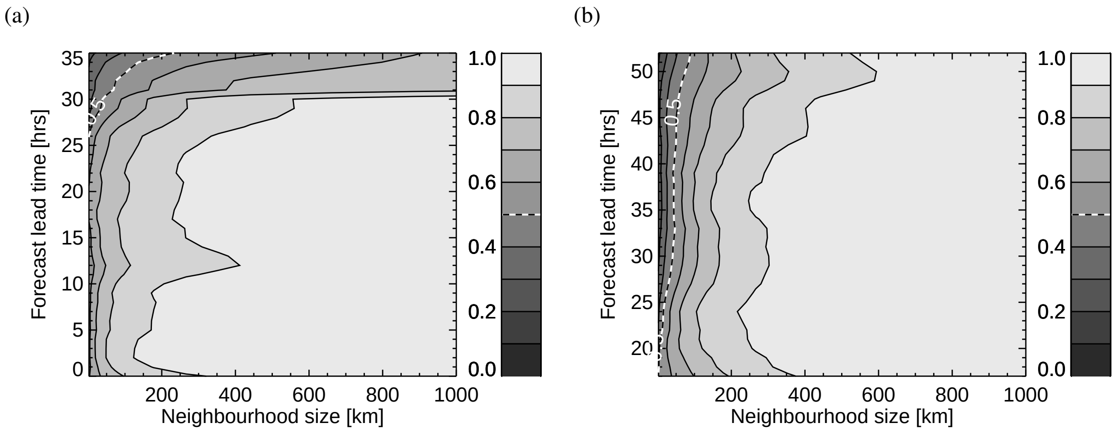


FIG. 9. dFSSmean calculated using the $90^{th}$ percentile threshold of hourly precipitation accumulations for (a) the organized Spring and (b) the Summer flooding case. Results are calculated over the whole of the UK domain and only the standard physics configuration is considered. The white dashed line at 0.5 represents the believable scale.

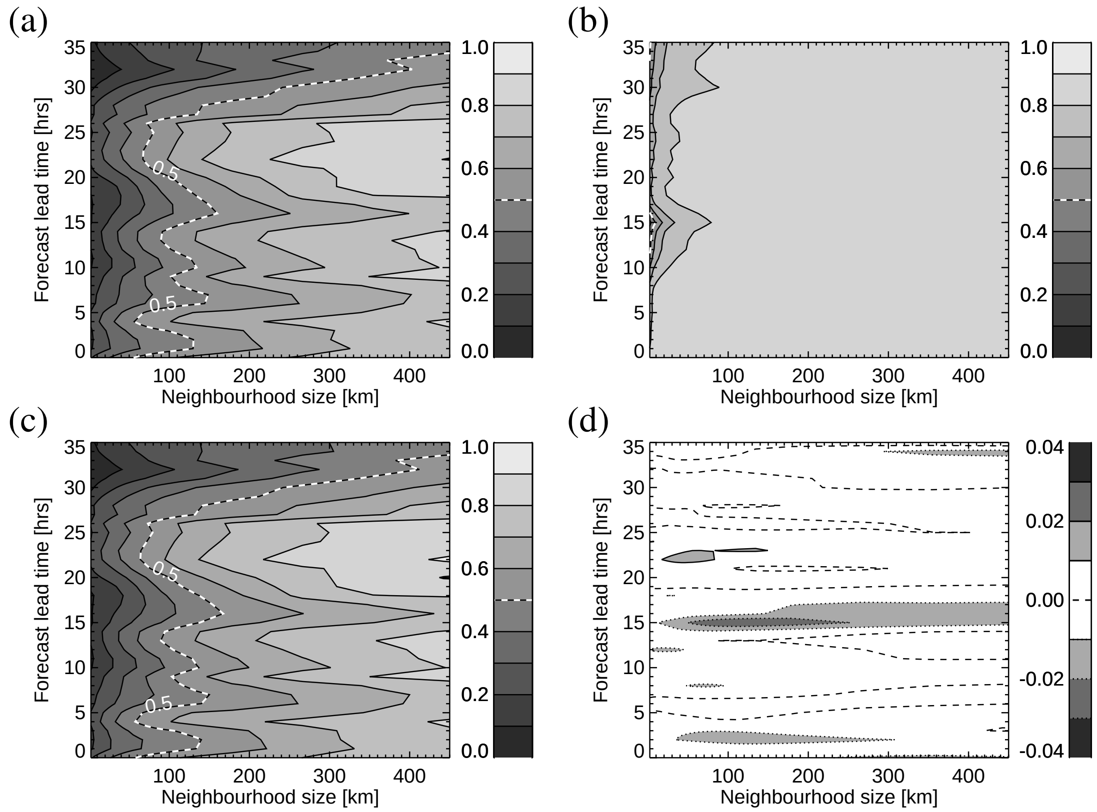


FIG. 10. dFSSmean comparisons of the restricted convection with increased time step and graupel physics configurations for the $99^{th}$ percentile threshold of hourly precipitation accumulations. Results from different comparison strategies are shown: (a) MOGREPS2, (b)Physics2 and (c) MOGREPS2+Physics2. (d) shows the difference between sub-figures (c) and (a). Results are calculated over the whole of the UK domain. The white dashed line at 0.5 represents the believable scale.