Ensemble prediction for nowcasting with a convection-permitting model—I: description of the system and the impact of radar-derived surface precipitation rates

By S T E F A N O   M I G L I O R I N I [1]*, M A R K   D I X O N [2], R O S S   B A N N I S T E R [1] and S U E   B A L L A R D [2],
[1]*Department of Meteorology, University of Reading, Reading RG6 6BB, UK;* [2]*Joint Centre for Mesoscale Meteorology, Met Office, Reading, UK*

ABSTRACT

A key strategy to improve the skill of quantitative predictions of precipitation, as well as hazardous weather such as severe thunderstorms and flash floods is to exploit the use of observations of convective activity (e.g. from radar). In this paper, a convection-permitting ensemble prediction system (EPS) aimed at addressing the problems of forecasting localized weather events with relatively short predictability time scale and based on a 1.5 km grid-length version of the Met Office Unified Model is presented. Particular attention is given to the impact of using predicted observations of radar-derived precipitation intensity in the ensemble transform Kalman filter (ETKF) used within the EPS. Our initial results based on the use of a 24-member ensemble of forecasts for two summer case studies show that the convective-scale EPS produces fairly reliable forecasts of temperature, horizontal winds and relative humidity at 1 h lead time, as evident from the inspection of rank histograms. On the other hand, the rank histograms seem also to show that the EPS generates too much spread for forecasts of (i) surface pressure and (ii) surface precipitation intensity. These may indicate that for (i) the value of surface pressure observation error standard deviation used to generate surface pressure rank histograms is too large and for (ii) may be the result of non-Gaussian precipitation observation errors. However, further investigations are needed to better understand these findings. Finally, the inclusion of predicted observations of precipitation from radar in the 24-member EPS considered in this paper does not seem to improve the 1-h lead time forecast skill.

1. Introduction

The rate of increase of computer power experienced to date has allowed meteorological agencies to run short-range numerical weather prediction (NWP) forecasts at increasing resolution. This has led to the development of operational 'deep-convection-permitting' models (e.g. Clark, 2009) and characterized by a grid length of about 1 km. These can represent deep convection explicitly rather than through a parameterization scheme. Because of computational costs, which for a constant domain size increase as the cube of the grid length size reduction ratio, these models are implemented over a limited area only and lateral boundary conditions are provided by nested model runs at lower resolution.

Another benefit of increasing the model horizontal resolution is the possibility of using higher resolution topography and a better description of surface fluxes, as well as of using high-resolution measurements capable of describing ongoing convection for assimilation (e.g. Stensrud et al., 2009). The main source of observations that are able to provide a detailed representation of convective activity is from radar, either in the form of radial velocity and reflectivity or in the form of precipitation measurements (e.g. Sun, 2005).

A major aim of considering models at resolutions capable of resolving deep convective systems (if not clouds) is to improve quantitative precipitation forecasting (e.g. Mass et al., 2002; Fritsch and Carbone, 2004) as well as hazardous weather such as severe thunderstorms (e.g. Stensrud et al., 2009). Over the last few years, a number of studies have demonstrated the utility of using convection-permitting models for convective-scale NWP. For example, Done et al. (2006) compared results with the Weather Research and Forecast (WRF) model at a

resolution of 10 and 4 km, with and without parameterization of convection, respectively. They showed that forecasts at 4 km resolution were characterized by a better correspondence between observed and simulated mesoscale convective systems (MCSs), as well as by a better determination of the occurrence of MCSs. Furthermore, Lean et al. (2008) studied the performance of nested high-resolution versions (12, 4 and 1 km grid lengths) of the Unified Model (UM), when no convection parameterization is used within the 1 km resolution model, for a number of summer-time case studies. By allowing the model to represent convection explicitly at 1 km resolution, they found that it was possible to achieve a better representation of convection initiation and a more realistic description of mesoscale features. However, a remaining difficulty was the presence of too many convective cells with low rain rate.

Despite the practical advantages and the increasing amount of evidence of improved performance, forecast skill of convection-permitting models is still limited by the large sensitivities to initial conditions (e.g. Hohenegger and Schär, 2007) and to model error (e.g. Park and Droegemeier, 2000; Gilmore et al., 2004; Tong and Xue, 2008). These large sensitivities reduce predictability, which is more pronounced at smaller horizontal scales: a 20 km wavelength feature (e.g. a thunderstorm) has a deterministic predictability horizon of 1–2 h (e.g. Lilly, 1990). The short predictability horizon means that forecasts are quickly affected by significant errors, which due to nonlinearities may not be normally distributed. In this case, a probabilistic approach based on a forecast ensemble can be effective in providing informative predictions.

This paper describes a convective-scale version of the Met Office Global and Regional Ensemble Prediction System (MOGREPS; Bowler et al., 2008) over the southern United Kingdom, set up to study forecast error and predictability. This will help design the future Met Office NWP-based nowcasting system. An initial characterization of its behaviour and skill at 1 h lead time via a series of experiments performed on two case studies is provided. The other main aim of this paper is to investigate the impact of predicted radar-derived surface precipitation rates on the forecast skill obtained with the convective-scale version of the MOGREPS. The structure of the paper is as follows. Section 2 is a description of the convective-scale ensemble prediction system used in this study is provided. Sections 3 and 4 describe the ensemble forecasting experiments performed in this work and their results, respectively. Finally, Section 5 is a summary of the results and the conclusions.

2. Convective-scale ensemble prediction system

In this work, the evolution of forecast error is estimated by means of an ensemble of forecasts from a convection-permitting version of the UM at 1.5 km horizontal grid spacing over the southern United Kingdom (Fig. 1), with 360×288 points over 70 levels (with variable vertical grid spacing and top model level height of
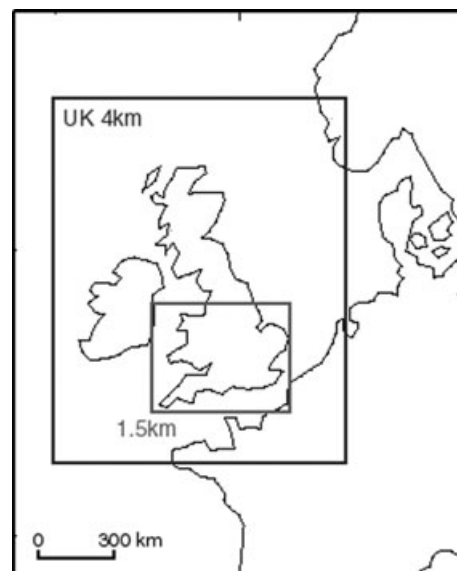

Fig. 1. The 1.5 and 4 km models domains.

about 38 km), initialized from a set of initial conditions as discussed below. The dynamic formulation of the UM is based on non-hydrostatic equations that are solved on a rotated equatorial latitude–longitude horizontal grid with Arakawa C staggering. The vertical levels are described by means of a hybrid-height, terrain-following coordinate with Charney-Philips staggering (Davies et al., 2005). The large-scale precipitation scheme used in this configuration is based on microphysical calculations involving water vapour, cloud liquid and ice water content, treated as prognostic variables (Wilson and Ballard, 1999). In addition, a prognostic rain variable is advected with the three-dimensional wind field. so as to limit the occurrence of grid-point convection (Lean et al., 2008, their section 2.d).

The high-resolution control analysis is produced using a version of the data assimilation system described in Dixon et al. (2009), adapted to an hourly cycle frequency. Lateral boundary conditions (LBCs) are provided by a 4 km grid spacing model over the United Kingdom (Fig. 1). In each cycle, the model is run from $T - 0.5$ with respect to the cycle validity time T (UTC), using the $T + 0.5$ background model state, $\mathbf{x}_b$, from the previous cycle as initial conditions (Fig. 2). A 3DVAR scheme (Lorenc et al., 2000) is used for the analysis of surface observations of pressure, 1.5-m temperature, 1.5-m relative humidity and 10-m u- and v-component of wind and aircraft observations of temperature and u and v-component of wind. This 3DVAR scheme uses a **B**-matrix whose formulation is suited to synoptic-scale rather than convective-scale dynamics as the Met Office does not yet have a truly convective-scale 3DVAR system. We assume for this paper that this preliminary system is still useful to study convective-scale forecasts as fine-scale features are expected to emerge as the ensemble of high-resolution forecasts progresses. Assimilation increments derived from 3DVAR, $\mathbf{x}'_{VAR}$, are added
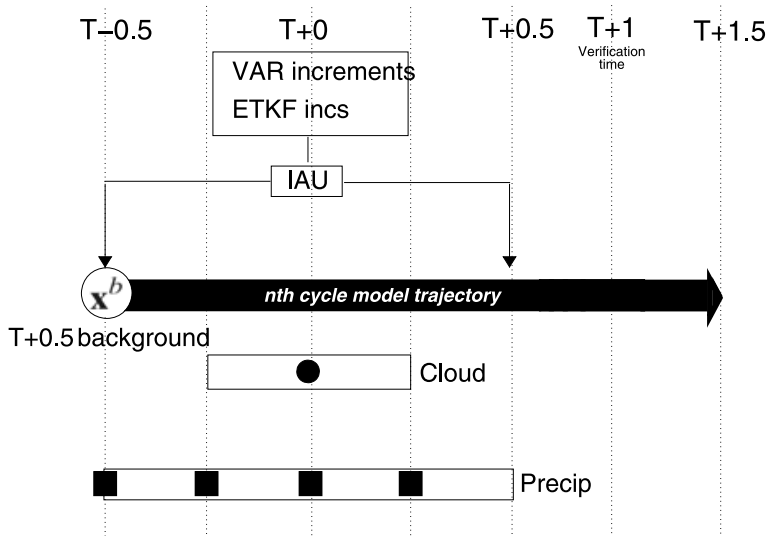
Fig. 2. Schema of the data assimilation and ensemble cycles. In the case of the data assimilation cycle, only 3DVAR increments are added via the IAU. In the case of each ensemble forecast, both 3DVAR and the relevant ensemble analysis perturbations are added via the IAU (refer to text for more details).

to the model state via the incremental analysis update (IAU) procedure (Bloom et al., 1996) (which reduces spurious imbalance) from $T - 0.5$ to $T + 0.5$. Convective-scale assimilation of cloud profiles and surface precipitation from the UK weather radar network (Golding, 1998) was performed via two nudging procedures. Cloud nudging was used to assimilate hourly humidity information produced using the Moisture Observation Pre-processing System (MOPS) (Macpherson et al., 1996). Radar-derived surface precipitation rates, available on a 5 km resolution grid every 15 min, are corrected, quality controlled (Harrison et al., 2000), re-sampled onto a 15 km resolution grid and assimilated with a latent-heat nudging scheme (Jones and Macpherson, 1997).

The initial conditions for the first cycle of the high-resolution ensemble forecasting experiments (described in S 3) are provided by a set of K initial fields (presently equal to 24) determined by adding the available $K - 1$ MOGREPS initial perturbations—with a horizontal resolution of 24 km, over 38 levels in the vertical—to the analysis at the same resolution. These 'large-scale' initial conditions are then interpolated on to the 'high-resolution' convection-permitting model grid. The analysis, also interpolated on to the high-resolution grid, is subtracted from the initial conditions, to derive a set of $K - 1$ perturbations. These perturbations are added to the high-resolution analysis, which includes high-resolution observational information from radar, via the IAU (Fig. 2).

New high-resolution perturbations are calculated at every 1 h forecast lead time by means of an ensemble transform Kalman filter (ETKF). According to this algorithm, the departures of the initial conditions, $\mathbf{x}_i^a$, from the analysis $\bar{\mathbf{x}}^a$ at a given time, are related to the departures of the model forecasts, $\mathbf{x}_i^f$, from the mean $\bar{\mathbf{x}}^f$ as

$$\mathbf{X}^a = \mathbf{X}^f \mathbf{T} \Pi, \tag{1}$$

where $\mathbf{X}^f = (\mathbf{x}_1^f - \bar{\mathbf{x}}^f, \ldots, \mathbf{x}_i^f - \bar{\mathbf{x}}^f, \ldots, \mathbf{x}_K^f - \bar{\mathbf{x}}^f)$ and $\mathbf{X}^a = (\mathbf{x}_1^a - \bar{\mathbf{x}}^a, \ldots, \mathbf{x}_i^a - \bar{\mathbf{x}}^a, \ldots, \mathbf{x}_K^a - \bar{\mathbf{x}}^a)$, $\mathbf{x}_i^f$ and $\mathbf{x}_i^a$ being the ith ensemble forecast and analysis, respectively. $\mathbf{T}$ is the ensemble transform matrix, which can be expressed as (Wang et al., 2004)

$$\mathbf{T} = \mathbf{C}(\mathbf{\Gamma} + \mathbf{I})^{-1/2} \mathbf{C}^T, \tag{2}$$

where $\mathbf{C}$ is a matrix whose columns are the right singular vectors of $\mathbf{E}$ and $\mathbf{\Gamma}$ is a diagonal matrix whose diagonal elements are the nonzero singular values of $\mathbf{E}$. Π is an 'inflation factor' needed to increase the analysis error variance and compensate the sampling error due to the relatively small ensemble size used in the ETKF (Section 2.1). The matrix $\mathbf{E}$ is given by

$$\mathbf{E} = \mathbf{R}^{-1/2} \mathbf{H} \left(\mathbf{P}_e^f \right)^{1/2}, \tag{3}$$

where $(\mathbf{P}_e^f)^{1/2} \equiv \mathbf{X}^f / \sqrt{K-1}$ is a square root of the forecast error covariance matrix estimated by means of an ensemble of K forecasts, $\mathbf{R}$ is the observation error covariance matrix and $\mathbf{H}$ is the observation operator, here assumed to linearly relate the observations with the true atmospheric state vector. Note that from (2) and (3) it is clear that only the components of the observation vector that are accurate enough to produce singular values of $\mathbf{E}$ that are significantly greater than 1 (or, in other words, with signal to noise ratio significantly greater than 1) will be able to reduce the analysis perturbation amplitude.

2.1. Inflation factor

The scheme used to calculate the inflation factor in this work is detailed in Bowler et al. (2008). For large-scale versions of MOGREPS, the inflation factor is based on sonde and satellite observation types. However, the temporal frequency at which these observations are available (every 6 h in the case of sondes) as well as their spatial coverage is inadequate for the 1-h

cycle nowcasting system over the southern UK presented here. For this reason, the inflation factor used in this study was based on selected observations from surface stations (which include pressure, temperature, horizontal wind components, relative humidity and visibility) and aircraft reports (which include temperature and horizontal winds). Note that precipitation observations from radar were not considered for the purpose of calculating the inflation factor, given their statistical distribution was more likely to be non-Gaussian (see Section 4.2.1 for more details).

The inflation factor is used to scale the analysis perturbations at time T with the aim of ensuring that at $T + 1$, the variance of the innovations σ_d^2—defined as the difference between observations and observations predicted by means of the ensemble mean forecast—is given by

$$\sigma_d^2 = \sigma_f^2 + \sigma_{tobs}^2, \tag{4}$$

where σ_f^2 is the variance of the ensemble forecast error (in observation space), and σ_{tobs}^2 is the variance of the 'true' observation error, when it is assumed that forecast errors are uncorrelated with observation errors. To avoid inflation factors that vary too much from cycle to cycle, each factor includes a correction given by the ratio of the mean square spread at the previous cycle to the mean square spread at the current cycle (see Bowler et al., 2008). Note that the term 'true' observation error is adopted to distinguish our estimates of actual observation error from those used in this work for variational assimilation, σ_{VAR}^2, which are usually chosen to be larger than σ_{tobs}^2, as a means of dealing with problems of correlation and representativeness. Estimates for σ_{tobs}^2 were obtained by first running the ensemble system for a relatively short number of cycles (four), using σ_{VAR}^2 as a proxy for σ_{tobs}^2. A linear regression of σ_d^2 against σ_f^2 was then performed using the output from these runs, to determine σ_{tobs}^2 (as the value of σ_d^2 extrapolated for $\sigma_f^2 = 0$). Reliable estimates (expressed in terms of $\sigma_{tobs}^2/\sigma_{VAR}^2$ in Table 1) were found for a subset only of the observations. In the case of the other observations, the regression method was not found to be useful as σ_f^2 was found to be approximately constant or to decrease with increasing σ_d^2 (contradicting to the assumptions of eq. 4). The inflation factor calculations performed in all subsequent experiments used (at most) only those observations and associated errors listed in Table 1.

2.2. Use of radar-derived surface precipitation data within the ETKF

The original ETKF scheme used in MOGREPS was not designed to be used with predicted observations (denoted as $\mathbf{H}\mathbf{x}_i^f$, in the case of the ith forecast ensemble member) of precipitation, so that a number of modifications to the Met Office operational observation processing system (OPS) had to be undertaken. The purpose of these modifications was to create a set of predicted observations, by using the model surface precipitation rate field at a given time, which was spatially averaged and interpolated so as to match the resolution and location of the radar observations that are assimilated in the UM. In this way, by using a set of K ensemble members, it is now possible to create a set of K predicted observations of surface precipitation rate at a given location and time and calculate the term $\mathbf{H}(\mathbf{P}_e^f)^{1/2}$ within the transform matrix, by taking the mean of all sets of K predicted observations.

To calculate the transform matrix with (3), an expression for the surface precipitation rate observation error covariance matrix $\mathbf{R}$ is also required. The latent heat nudging procedure does not require this information as it only makes use of a quality indicator that depends on the radar range and elevation of the measurement from the model freezing level (see Dixon et al., 2009 and references therein), which is not suitable for determining $\mathbf{R}$.

Over the last few years, the Japan Meteorological Agency (JMA) has been assimilating 2.5 km resolution hourly precipitation amounts from radar using 4DVAR with a mesoscale model centred over Japan (Koizumi et al., 2005). The JMA model horizontal resolution used in their paper is 10 km over 40 vertical levels up to 10 hPa, whereas the resolution of the tangent linear and adjoint model is reduced to 20 km, for computational efficiency. The departures of hourly precipitation amounts from radar and from a model short-range forecast at observation locations and times showed evidence of non-Gaussian observation errors, with a distribution closely approximated by a function of observation and model forecast values. This error distribution gives rise to a quadratic assimilation cost function with a smaller curvature (i.e. with a larger observation error variance) for forecast values that are larger than observation values. This evidence led the JMA to model precipitation observation error standard deviation from radar σ_r, to be used in variational assimilation, as

Table 1. Observations with corresponding error standard deviation used for 3DVAR assimilation and error ratio used for the inflation factor calculations

Observation type	VAR error	Error ratio
Surface temperature	1.1 K	1.0
Surface u- and v-wind	1.2 ms^{-1}	0.6
Aircraft temperature	0.84–1.9 K	1.0
Aircraft u-wind	1.45–2.8 ms^{-1}	1.0
Aircraft v-wind	1.45–2.8 ms^{-1}	1.0

Note: Error ratio is defined as the ratio between 'true' observation error standard deviation and that used for 3DVAR (see text for more details).

$$\sigma_r = \begin{cases} cr_c & x_r^f \le y_r^o \text{ and } y_r^o < r_c \\ cy_r^o & x_r^f \le y_r^o \text{ and } y_r^o \ge r_c \\ 3cr_c & x_r^f > y_r^o \text{ and } y_r^o < r_c \\ 3cy_r^o & x_r^f > y_r^o \text{ and } y_r^o \ge r_c \end{cases} \tag{5}$$

where y_r^o and x_r^f are the observation and the model forecast value, respectively, and $r_c = 1$ mm. Note that, according to Koizumi et al. (2005), the factor c in (5) was determined empirically to be 1.

To describe the error statistics of radar-derived surface precipitation rates used in this work we decided to adopt the same expression as in (5), but with $r_c = 1$ mmh^{-1} (note our units) and a variable c, so as to control the forecast spread reduction due to radar observations. If not otherwise specified, c should be supposed equal to 1.

3. Description of the experiments

In this section, we describe two case studies that were chosen to assess the performance of the ensemble prediction system, one focused on forecast skill of temperature, pressure and wind fields (on 27 July 2008, denoted as C1) and the other focused on precipitation (on 26 July 2007, denoted as C2).

3.1. The 27 July 2008 case study (C1)

For this case study, characterized by the presence of showers mainly concentrated over the eastern part of the domain (East Anglia), a series of four experiments were performed. Each experiment was run for a total of 11 cycles, starting from 0700 UTC and ending at 1700 UTC. Figure 3 shows the synoptic situation at 0600 UTC and 1800 UTC, corresponding to the beginning and the end of the experiments, respectively.

Four different configurations of the ensemble system have been tested for C1 (see Table 2 for a summary). In experiment C1E01, LBCs for the forecast ensemble were provided by the UK 4 km model, and the inflation factor calculation included both aircraft and surface observations (as summarized in Table 1). Experiment C1E02 differed from C1E01 in using surface observations only in the inflation factor. Experiments C1E03 and C1E04 made use of perturbed LBCs from the lower resolution (24 km grid) North-Atlantic and European (NAE) regional version of the MOGREPS, with both surface and aircraft

Table 2. Summary of experiments for the 27 July 2008 case study (C1)

Experiment	Obs source for inflation factor calculation	LBC
C1E01	Surface, aircraft	Deterministic (4 km model)
C1E02	Surface	Deterministic (4 km model)
C1E03	Surface, aircraft	Perturbed (NAE)
C1E04	Surface	Perturbed (NAE)

observations used in the inflation factor for C1E03 and a surface-observation inflation factor used in C1E04.

3.2. The 26 July 2007 case study (C2)

As discussed in Section 2, predicted observations at given locations and with given observation errors, determine the analysis perturbations according to their signal-to-noise ratio. For this case study, a number of experiments were carried out to investigate the impact of precipitation on the structure of the ETKF perturbations as determined by the ensemble transform matrix. Note that within MOGREPS, the perturbed state variables that represent the forecast uncertainty are velocity components (u' and v'), potential temperature (θ'), specific humidity (q') and Exner pressure (Π'). The effects of predicted observations—including those of precipitation—on forecast uncertainty of precipitation depends, therefore, on cross-correlations between ETKF-derived and precipitation perturbations as imposed by the model.

Our tests were based on the analysis of a rainfall event over the southern UK, which occurred on 26 July 2007. Figure 4 is the analogue of Fig. 3, but for 26 July 2007 and Fig. 5 shows the radar rainfall rate observations over the United Kingdom at 1200 UTC. The area is affected by a cyclone moving towards the north of Scotland, with a cold front and a warm front meeting at a triple point over Western England. In the 1800 UTC mean sea-level pressure chart (not shown), the two fronts are merged into an
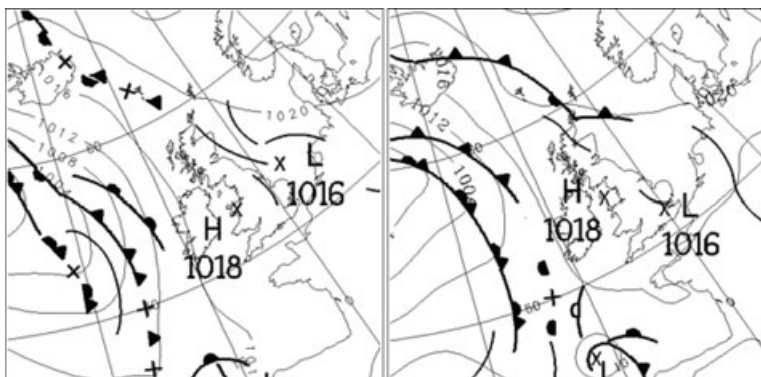

Fig. 3. Mean sea-level pressure (in hPa) analysis chart at 0600 UTC (left panel) and 1800 UTC (right panel) on the 27th July 2008. Courtesy of the Met Office.

Fig. 4. Mean sea-level pressure (in hPa) analysis chart at 0600 UTC (left panel) and 1200 UTC (right panel) on 26 July 2007. Courtesy of the Met Office.
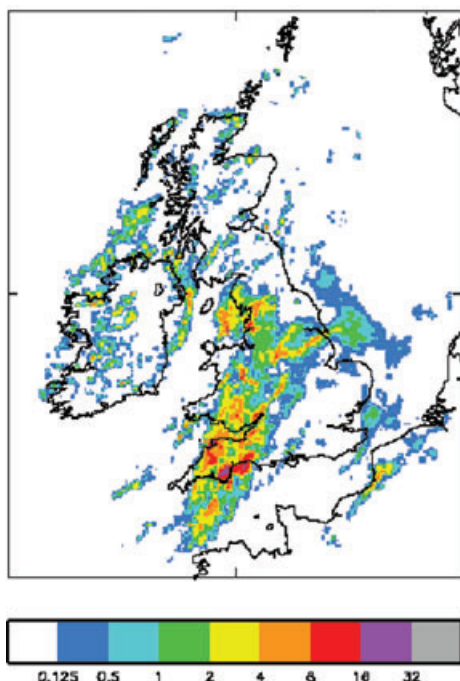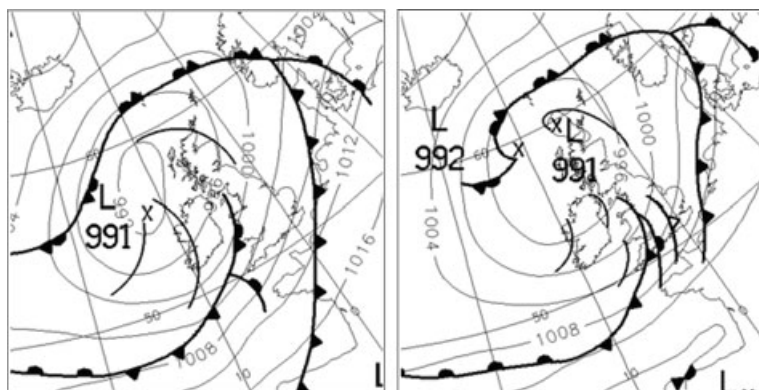

Fig. 5. Surface rainfall rate (in mmh^{-1}) from the radar UK network at 1200 UTC on 26 July 2007. Courtesy of the Met Office.

occluded front over the North Sea. An upper-level split front is present over South East England. The event is also characterized by embedded instability and line convection.

As in C1, each ensemble experiment consisted of 11 cycles, from 0700 UTC to 1700 UTC. Fifteen different configurations of the ETKF system have been tested for C2 (see Table 3 for a summary). The first experiment, denoted as experiment C2E01, did not include predicted observations of precipitation, and made use of both surface and aircraft observations to determine the inflation factor Π, as explained in Section 2.1. Also, the boundary conditions were set to be different for each ensemble forecast member, and were determined by downscaling an ensemble of MOGREPS NAE forecasts. As for the experiments in the previous case study, the ETKF perturbations and the 3DVAR analysis

increments were added to each forecast trajectory at each time step over a period of an hour—between $T - 0.5$ and $T + 0.5$—by means of the IAU procedure, whereas latent heat and cloud increments were nudged at each time step between $T - 0.5$ and $T + 0.5$ and between $T - 0.25$ and $T + 0.25$, respectively.

Other experiments (C2E08–C2E11) made use of the same boundary conditions for all ensemble members, which were determined from the downscaling of the same 4 km horizontal resolution forecast that was used to prescribe the boundary conditions for the data assimilation experiment. The sensitivity of the results to the choice of the observations informing the inflation factor was also tested by removing aircraft observations from the inflation factor calculations (C2E06–C2E07 and C2E10–C2E11). Furthermore, the effect of cloud and latent heat nudging on the ensemble forecast spread of the surface precipitation field at $T + 1$ was determined (C2E12–C2E13). Experiments were performed with latent heat nudging only acting between $T - 0.5$ and $T + 0.25$, and cloud nudging only between $T - 0.25$ and T, while keeping $T - 0.5$ and $T + 0.5$ as the time interval between successive ETKF and 3DVAR increments. Finally, the effect of adding the ETKF increments instantaneously at time T (i.e. without the use of the IAU procedure) was tested in the case when the nudging time interval ended at time T for cloud and at $T + 0.25$ for latent heat profiles (C2E14–C2E15). The IAU time interval for 3DVAR increments was kept between $T - 0.5$ and $T + 0.5$, as these increments include large-scale components of the flow, which need to be filtered for the effect of spurious gravity waves.

The effect on the forecast ensemble of including predicted precipitation observations in the ETKF was tested on the cases described earlier. Note that the resolution of the predicted precipitation observations was usually set to a 15 km × 15 km grid size, equal to the resolution of the precipitation observations used for assimilation, except for one experiment (C2E04) where the resolution of the predicted precipitation observations was set to 4.5 km × 4.5 km, comparable to the resolution of the original radar measurements (i.e. before averaging). The effects of including predicted precipitation observations corresponding to larger (i.e. above a given threshold) precipitation observation

Table 3.   Summary of experiments for the 26 July 2007 case study (C2)

Experiment	Resolution of predicted precip obs	Precip error factor, c	Obs source for inflation factor calculation	LBC	Nudging timescale
C2E01	n/a	n/a	Surface, aircraft	Perturbed (NAE)	Standard
C2E02	15 km	1	Surface, aircraft	Perturbed (NAE)	Standard
C2E03	15 km (predicted obs > 1 mmh^{-1} only)	1	Surface, aircraft	Perturbed (NAE)	Standard
C2E04	4.5 km	1	Surface, aircraft	Perturbed (NAE)	Standard
C2E05	15 km	2	Surface, aircraft	Perturbed (NAE)	Standard
C2E06	n/a	n/a	Surface	Perturbed (NAE)	Standard
C2E07	15 km	1	Surface	Perturbed (NAE)	Standard
C2E08	n/a	n/a	Surface, aircraft	Deterministic (4 km model)	Standard
C2E09	15 km	1	Surface, aircraft	Deterministic (4 km model)	Standard
C2E10	n/a	n/a	Surface	Deterministic (4 km model)	Standard
C2E11	15 km	1	Surface	Deterministic (4 km model)	Standard
C2E12	n/a	n/a	Surface, aircraft	Perturbed (NAE)	Reduced
C2E13	15 km	1	Surface, aircraft	Perturbed (NAE)	Reduced
C2E14	n/a	n/a	Surface, aircraft	Perturbed (NAE)	Instantaneous
C2E15	15 km	1	Surface, aircraft	Perturbed (NAE)	Instantaneous

values (C2E08), as well as the use of a larger precipitation observation error variance, were also tested (C2E05).

4. Results of the experiments

Here the results of the experiments performed in this work for the two case studies introduced in Section 3 are presented.

4.1. The 27 July 2008 case study (C1)

4.1.1. The baseline experiment C1E01. Rank histograms (Talagrand et al., 1997; Hamill, 2001) provide a useful measure of ensemble spread. Shown in Fig. 6 is a temporal series of histograms for an ensemble of $T + 1$ forecasts of surface temperature, u-wind, v-wind and relative humidity corresponding to experiment C1E01. Note that the effect of observation error was taken into account by adding to each forecast ensemble member a random number drawn from a Gaussian distribution with zero mean and standard deviation equal to 1.1 K (for temperature observations), 1.2 ms^{-1} (wind observations) and 6.2% (relative humidity). These values are all equal to the corresponding standard deviation value used for variational assimilation, except for wind data, for which a lower value was used, in accordance with the error ratio values reported in Table 1. The histograms for surface temperature seem to show that, at the beginning of the experiment, a cold bias overpopulates the right extreme ranks. During the course of the experiment the temperature ensemble spread becomes more consistent with the spread of the temperature observations, probably owing to an increased value of the inflation factor (see below for a discussion of the inflation factor)

or to a dependence of the bias on the diurnal cycle. A similar picture can be observed for relative humidity, except that the initial moist bias is on the left extreme ranks. The horizontal wind components both have a fairly flat rank histogram, which seems to indicate that the ensemble prediction system achieves a good estimate of the horizontal wind ensemble spread. In the case of surface pressure, the rank histograms for experiment C1E01 (Fig. 7, left panel) when the value of the surface pressure observation error is drawn from an unbiased Gaussian distribution with standard deviation equal to 100 Pa (consistent with the corresponding value used for assimilation), seems to show that the ensemble forecasts are overspread.

4.1.2. Experiment C1E03. A similar result is produced when the boundary conditions are determined from the nested NAE MOGREPS ensemble (instead of the fixed boundary conditions from the 4 km model; Fig. 7, right panel). This insensitivity of surface pressure spread to boundary conditions may indicate that the overspreading problem is from elsewhere. It may, for example, be due to a σ^2_{tobs} value that is too large, indicating the need to reduce the error ratio for surface pressure from the value given in Table 1.

4.1.3. The remaining experiments in C1.. Experiments C1E01 and C1E02 exhibit similar spread for all observations (of both surface and aircraft type). The rank histogram results for the C1E03 and C1E04 ensembles are also fairly similar to those for C1E01 and C1E02, with minor differences that are likely to be related to the different temporal behaviour of the inflation factors (Fig. 8).

4.1.4. Brier skill scores for C1E01. The skill of the forecasts produced by these experiments is assessed in Fig. 9 with
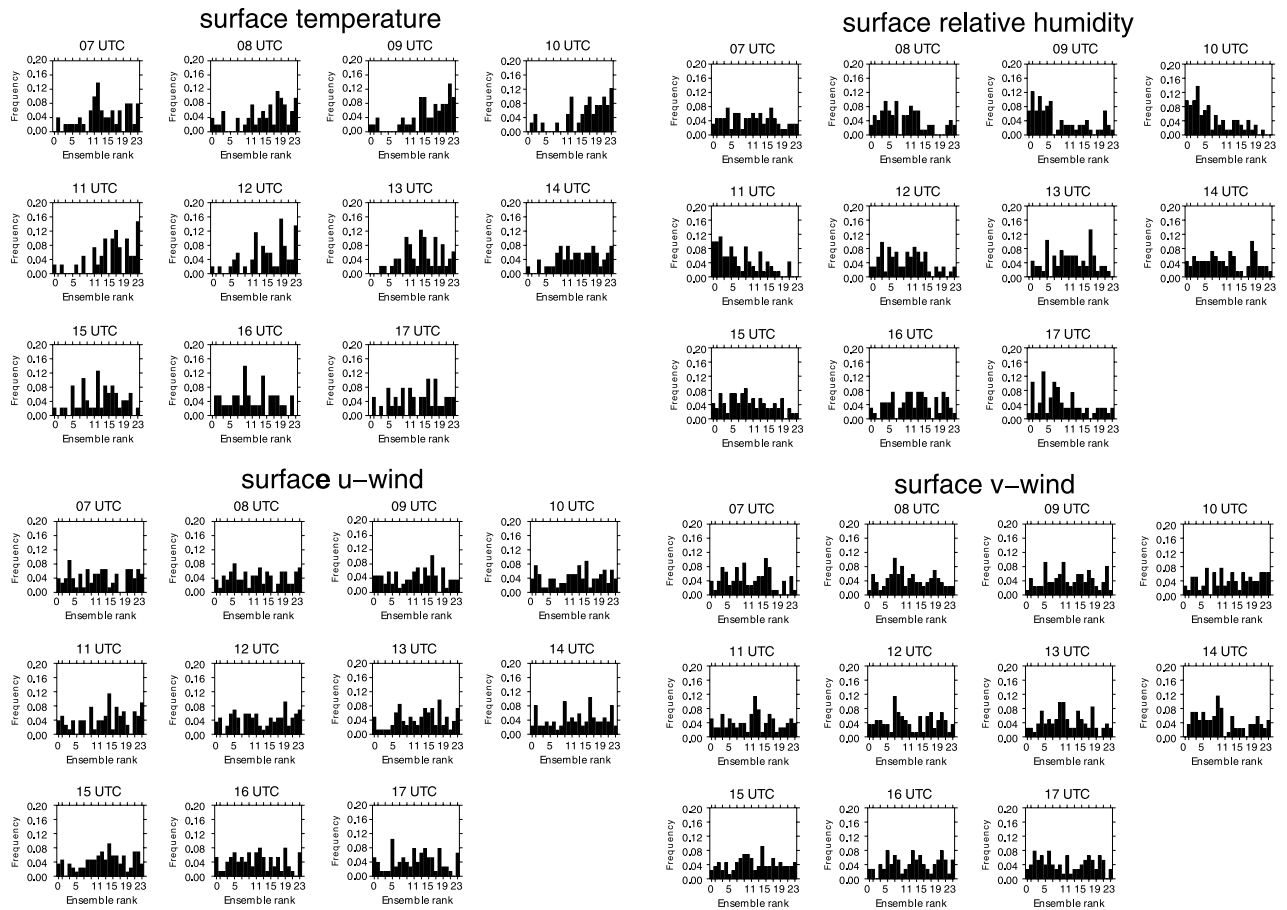
Fig. 6. Rank histograms for surface temperature (first panel), surface *u*-wind (second panel), surface *v*-wind (third panel) and surface relative humidity (fourth panel) produced by C1E01. Cycle hour increases from left to right on each panel, starting at 7Z and ending at 17Z.

respect to observations via the Brier skill score (BSS) (e.g. Wilks, 2006), using the in-sample climatology as a reference. We restrict attention to the surface observations, because the spatial coverage afforded by the aircraft data is poor. Even in the case of surface data, however, it is difficult to interpret the BSS, as the surface stations are widely separated. By way of illustration, the left panel of Fig. 9 shows the BSS produced by C1E01 for wind speed exceeding its median value at a given time. The central panel shows the time-dependent median values (solid line) used as thresholds for the BSS. The right panel shows the BSS (dashed line) for wind speeds exceeding the climatological median value of 2.1 ms^{-1} shown in the central panel (dashed line) found from 910 observations. Both the left and right panels show peaks of skill at 1200 UTC and 1500 UTC, but it is only in the left panel that it is possible to note a trend of increasing skill, likely to be due to the effect of the inflation factor (Fig. 8, top left panel) on the ensemble forecast spread. In the right panel, this effect is masked by the presence of a number of values above the threshold (i.e. above the median climatological wind speed) which increases with time.

We note that these results are useful mainly to characterize the evolution of the skill score during the experiment (thereby justifying the use of the in-sample climatology) and are not meant to provide an absolute measure of skill. As a matter of fact, even if a convective-scale forecast is relatively accurate in predicting the existence of an observed feature, a score such as the BSS will be poor if there is a small (subjectively tolerable) error in the positioning of that feature. In Section 4.2, the validity of the verification procedure is shown to be enhanced by the relatively high spatial density of surface precipitation rate observations.

4.2. The 26 July 2007 case study

Given the considerable number of different experiment configurations, results for experiment C2E01—referred to as the baseline experiment—are discussed extensively, whereas the differences between the other experiments and the baseline, particularly regarding the precipitation field, are highlighted briefly. An in-depth discussion of the statistical characteristics of forecast errors as well as their balance relationships, for a forecast

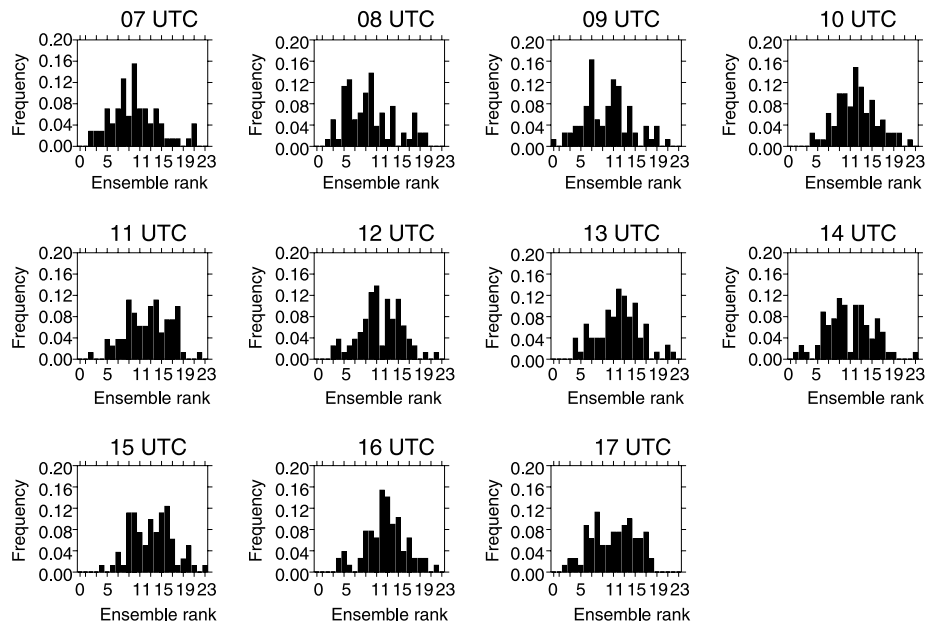# C1E01 surface pressure
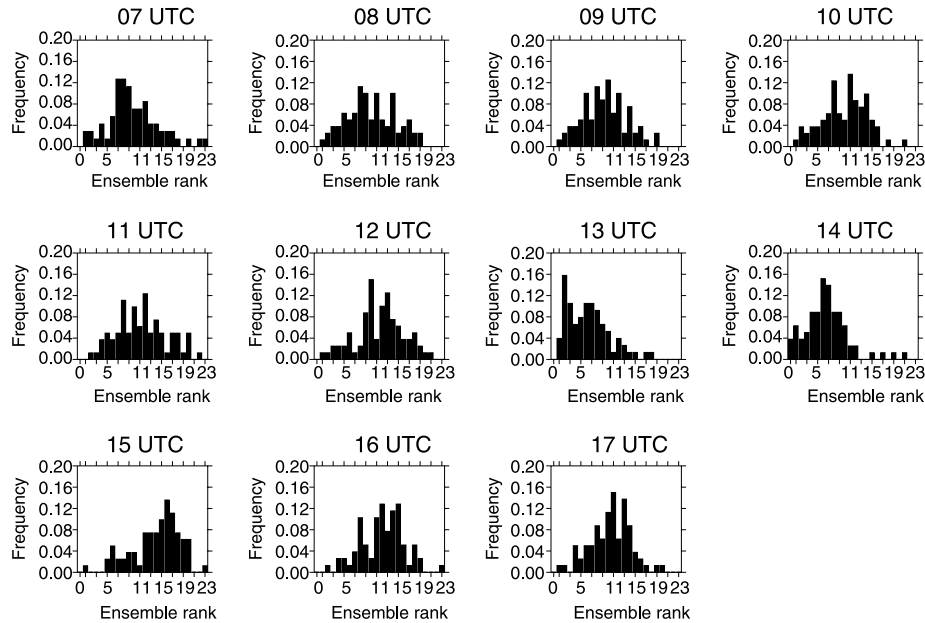
C1E03 surface pressure

Fig. 7. Rank histograms for surface pressure produced by C1E01 (top panel) and by C1E03 (bottom panel) at 17Z.

ensemble valid at 1800 UTC from experiment C2E01, is presented in Bannister et al. (2011).

4.2.1. The baseline experiment C2E01. As discussed in Section 2.1, the ensemble forecast standard deviation at time T is inflated by a factor Π in the hope that, at a later time here chosen

to be $T + 1$, the variance of the innovations becomes comparable to the variance of the ensemble forecast error in observation space plus the variance of the observation error. In Fig. 10, the estimated inflation factor values for experiment C2E01 as a function of time are shown. Note that, in this case, the initial
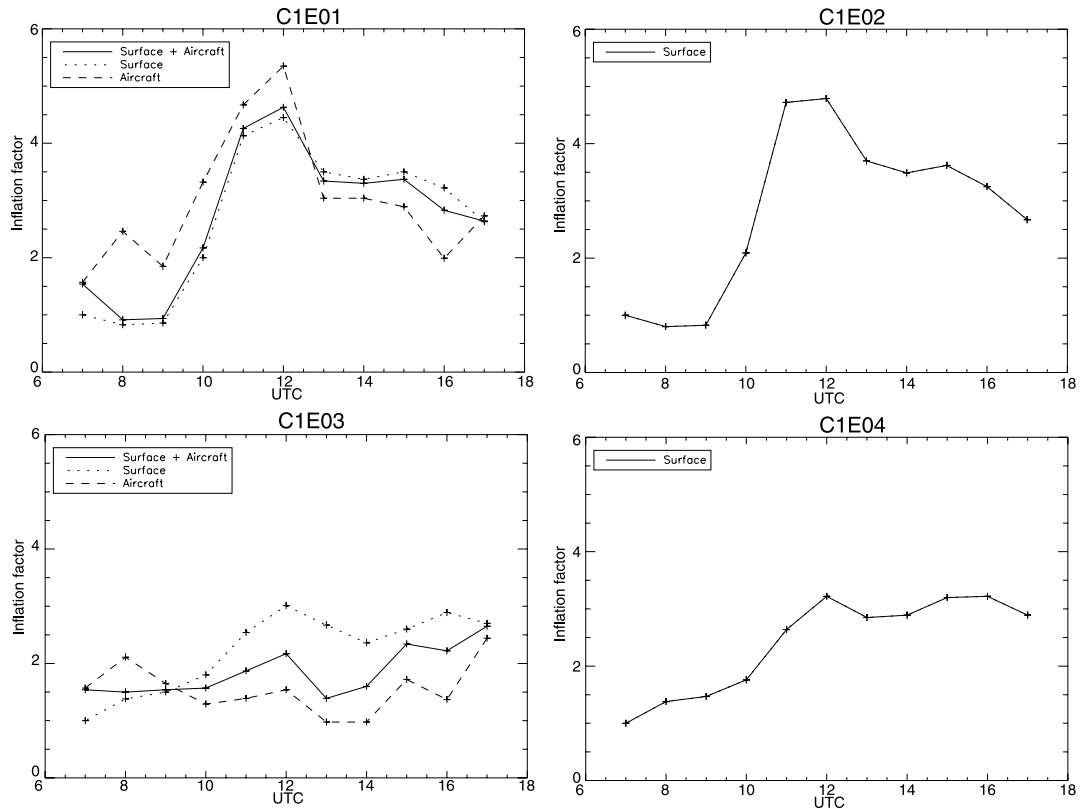
Fig. 8. Inflation factor of experiments C1E01 (first panel), C1E02 (second panel), C1E03 (third panel) and C1E04 (fourth panel): total (solid line), surface (dotted) and aircraft (dashed).

value of Π was set to 1 but, thereafter, both surface and aircraft observations were used to determine it. In the initial phase of the experiment, the inflation factor increases with time until about 1100 UTC, when the inflation factor peaks at a value of about 6. From Fig. 10 it also follows that the innovation variance due to surface observations shows the need for a larger forecast ensemble spread than in the case of aircraft observations. Results from experiments when only surface observations are used to inform the inflation factor are discussed in Section 4.2.3.

To illustrate the geographical distribution of the observational information available at 1200 UTC, the observation locations and values of the zonal wind component from surface stations and aircraft are shown in Fig. 11, together with the location and values of surface precipitation rate observations from radar.

As for the previous case study, discussed in Section 4.1, rank histograms were calculated to assess the reliability of the ensemble spread. Figure 12 shows the rank histogram results for surface temperature, surface horizontal wind speed components, surface relative humidity and surface pressure for experiment C2E01 valid at 1200 UTC. Note that the same values of observation error standard deviation as in C1E01 were used for these rank histograms. Figure 12 seems to confirm the results obtained for C1E01 shown in Figs 6 and 7, as the spread of all shown variable except surface pressure is fairly flat. The surface

pressure histogram seem to indicate that the surface pressure forecast ensemble is overspread, possibly due to the use of a too large surface pressure observation error (see discussion in Section 4.2).

The $T + 1$ precipitation rate forecasts for the 1100 UTC cycle, valid at 1200 UTC, from the 23 ensemble members and the control forecast (denoted as member 0) initialized at 1030 UTC are shown in Fig. 13. Note that, as explained in Section 2, between 1030 UTC and 1130 UTC the latent heat nudging procedure forces the precipitation forecasts to converge towards the radar rainfall observations. Despite this, the precipitation field for the different ensemble members presents a substantial degree of variability, which compensates for some of the shortcomings that are evident in the control forecast, identical to the 'deterministic' forecast from the previous analysis except for the use of boundary conditions from the NAE version of MOGREPS. For example, the control forecast does not capture the heavy rainfall over the north coast of South West England, while members 2, 3, 4, 6 and 18 show, over this area, levels of precipitation that are consistent with the spatial distribution and intensity measured by the radar network (Figs 5 and 11c). This is also the case for light precipitation events, such as those over the eastern part of the domain, which are not present in the member-0 forecast while are captured by a number of other forecasts members,
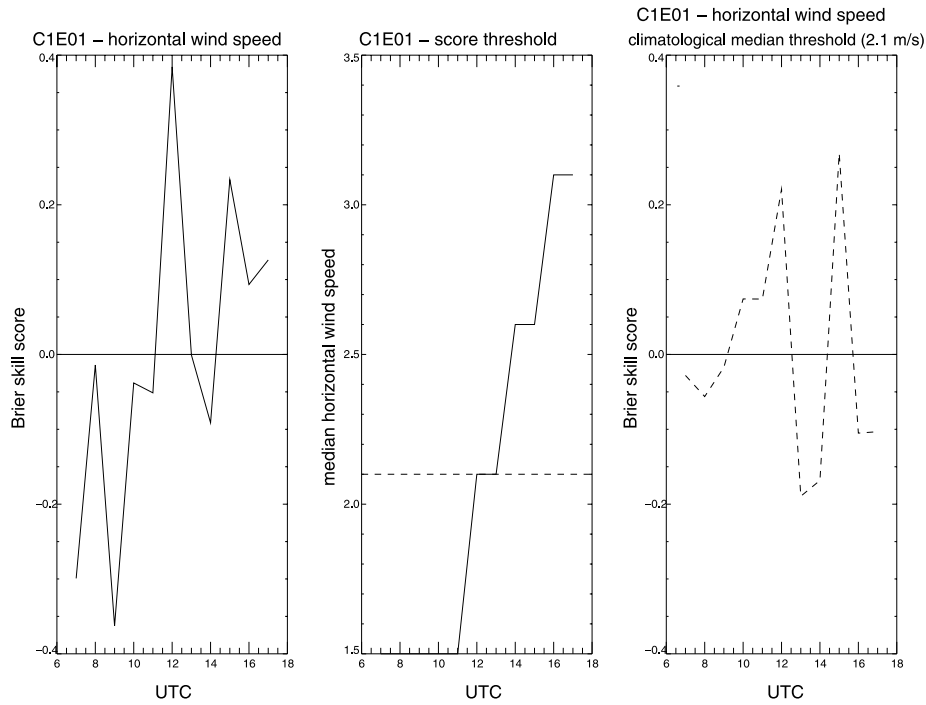
Fig. 9. Brier skill score (BSS) with respect to the in-sample climatology (left panel), for horizontal wind speed exceeding the threshold (equal to the median of the observational distribution at a given time) shown in the middle panel, produced by experiment C1E01. The right panel shows the BSS for a climatological threshold of 2.1 ms^{-1}, equal to the median of the observational distribution over the whole case study.
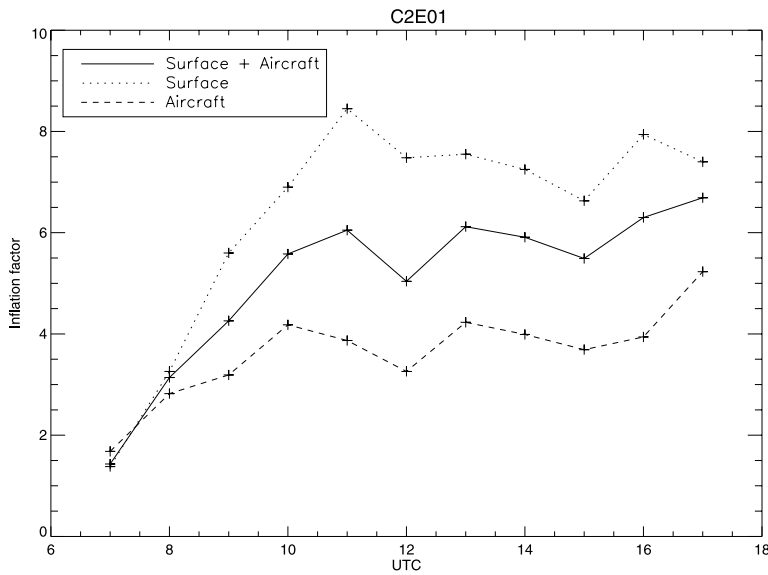

Fig. 10. Inflation factor values used within experiment C2E01, determined by using surface and aircraft observations (solid line), surface observation only (dotted line) and aircraft observation only (dashed line).

particularly by member 11. Finally, various members show evidence of spurious precipitation over the boundaries of the domain (e.g. over the eastern and southern boundary). This is likely to be due to inconsistencies arising from the high nesting ratio between the horizontal resolution of the outer and the inner model grid (being equal to 16).

The previous findings can be effectively summarized by means of (uncalibrated) probability plots, generated by calcu-

lating the number of members that exceed a given precipitation threshold at a given location and time. Note that in our experiments the probability resolution is equal to about 4%. In Fig. 14, the probability of occurrence of precipitation greater than 0.125, 1 and 5 mmh^{-1} are shown, along with the minimum, maximum and mean precipitation rate value for each location, valid at 1200 UTC. The great majority of members predicts for the intense event over the South West of England a precipitation
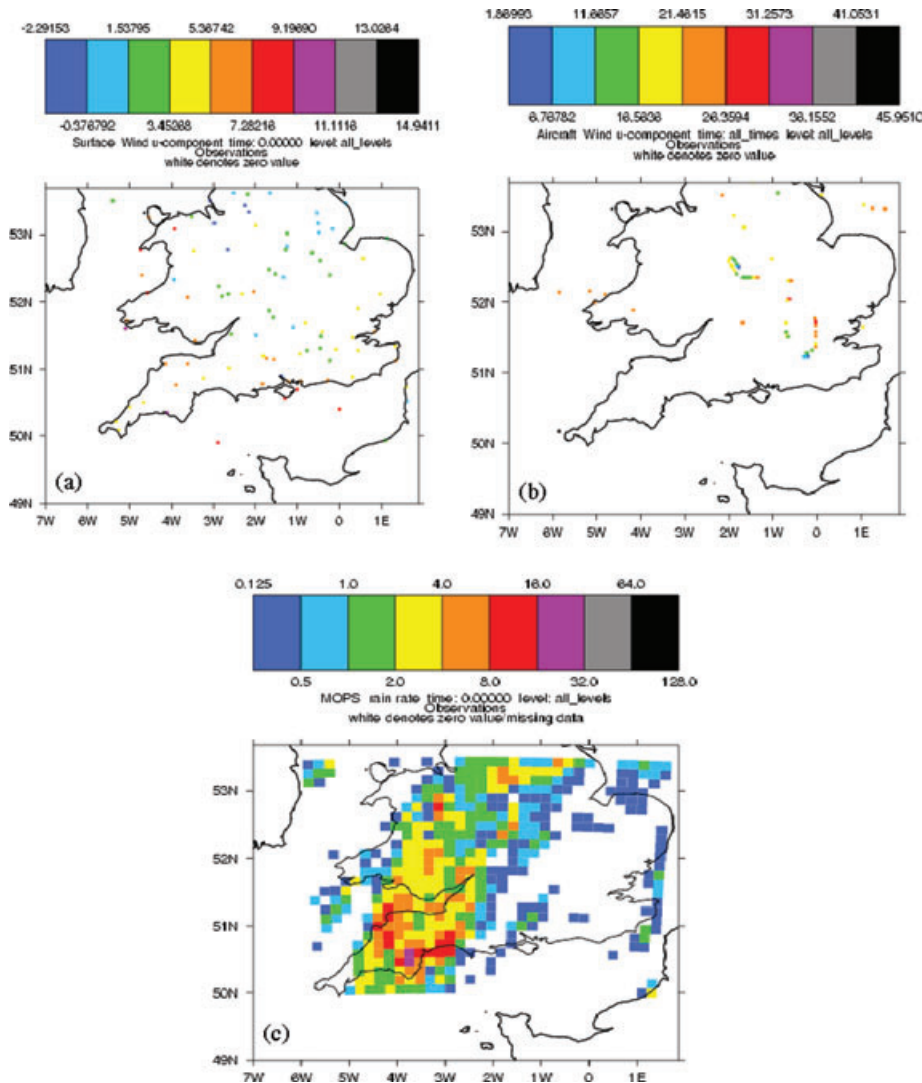
Fig. 11. Observations of the zonal component (denoted as *u*) of the wind vector (in ms^{-1}) from surface stations (panel *a*) and from aircraft (panel *b*). Panel (c) shows rainfall rate observations from the radar UK network (in mmh^{-1}) as in Fig. 5 but after being averaged over a 15 km grid length. Surface (panel a) and radar (panel c) observations were taken at 1200 UTC whereas aircraft observations (panel b) were taken between 1130 UTC and 1230 UTC, on 26 July 2007.

greater than 0.125 mmh^{-1}. However, the area where most members predict precipitation of higher intensity is relatively small (as compared to observations).

The ensemble mean forecasts an area over South West England with intense precipitation that is similar to that from radar observations, where the deterministic forecast (member 0 in Fig. 13) shows more localized events. This is the result of the fact that different ensemble members predict the same event over slightly different locations. The precipitation over the sea, on the eastern part of the domain, is also captured by the ensemble mean, although with less intensity than observed. Finally, the minimum precipitation field—that is, the field with minimum values of precipitation among all ensemble members—shows that all members predict precipitation over the western

and southern boundary of the domain, where the flow is advected into the domain (Fig. 4). This supports the hypothesis that the predicted precipitation over these areas is an artefact due to the lower resolution boundary conditions.

A more quantitative analysis of the forecasting performance of the ensemble system can be achieved, for example, by comparing the innovation variance against the ensemble forecast variance, for the period when the inflation factor has reached a quasi-stationary value, that is between 1100 UTC and 1700 UTC (Fig. 10). To do this the forecast variances are arranged in ascending order and divided into a number of bins, each containing the same number of variance and innovation values. The variances in each bin are averaged, and the root-mean-square innovation is calculated for each bin. Figure 15 shows the values
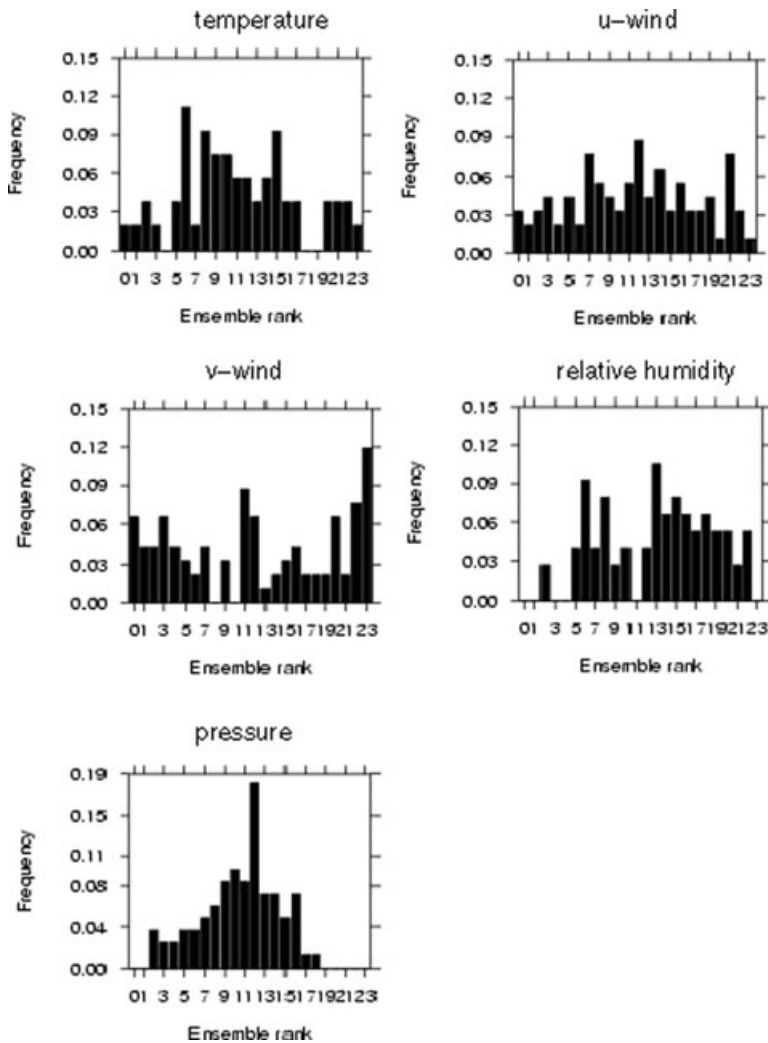
Fig. 12. Rank histograms for surface temperature (top-left panel), surface *u*-wind (top-right panel), surface *v*-wind (middle-left panel), surface relative humidity (middle-right panel) and surface pressure (bottom-left panel) produced by C2E01 at 1200 UTC.

of $T + 1$ forecast variance and of root-mean-square innovation of surface rainfall rate from radar on a 15 km resolution grid for the 6188 observations available between 1030 UTC and 1730 UTC (a quarter of the total number of 15 km resolution radar observations available between 1030 UTC and 1730 UTC; Table 4). The results for the case when only 5 bins are considered (asterisks), each bin representing 1237 forecast variances (except for the bin with largest variances, which represent 1240 forecast variances), are superimposed on the results for the case when 20 bins are considered (diamonds), each bin now representing 309 forecast variances (except for the bin with largest variances, which represent 317 forecast variances).

As discussed in Section 2.1, when forecast errors are uncorrelated with observation errors, innovation variance, σ_d^2, is linearly related to forecast error variance in observation space, σ_f^2, with unit gradient and $\sigma_f^2 = 0$ intercept of the observation error variance, σ_{tobs}^2 (see eq. 4). From Fig. 15 it appears that a linear relationship approximates quite well the dependence of forecast ensemble error variance on innovation variance, although with

a gradient that is less than unity. This could simply be the result of the use of ensemble variance as an imperfect representation of the true forecast error variance (see Wang and Bishop, 2003, their section 8b). As expected, the relationship becomes noisier when the sample size in each bin is smaller (see Wang and Bishop, 2003, their section 8c). The intercept of the linear fit between innovation and ensemble variance provides an estimate of the observation error variance, which resulted to be equal to 0.075 mm^1 h^{-2} for the 5-bin case and to 0.140 mm^1 h^{-2} for the 20-bin case. Based on these results, it was decided to use a value of 0.1 mm^1 h^{-2} for the observation error variance, to account for observation error in the evaluation of the skill of the ensemble prediction system performed in this paper. The distribution of innovation was investigated to determine whether it was sufficiently well approximated by a Gaussian distribution. To do so, the sample skewness and kurtosis (with a constant of 3 subtracted) were calculated, for the whole innovation sample (6188 data points), finding values of -1.431 and 9.091, respectively. Similarly to Pires et al. (2010), these values were compared
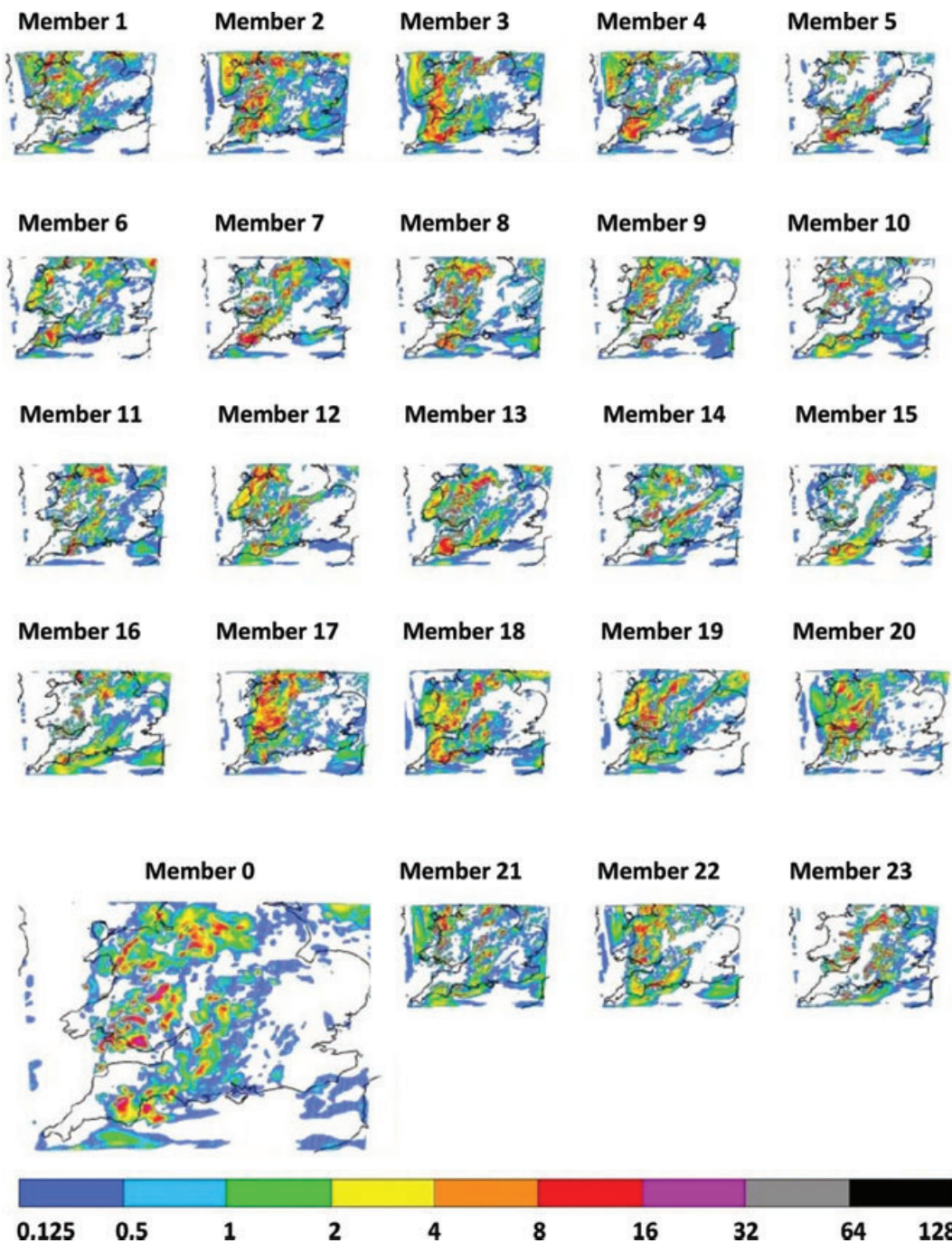
Fig. 13. Ensemble forecasts of surface precipitation rate (in mmh^{-1}) valid at 1200 UTC on 26 July 2007, for experiment C2E01. Member 0 is the forecast from the previous analysis.

with the mean values of sample skewness and sample kurtosis obtained for sample of 6188 pseudo-random numbers drawn from a Gaussian distribution with mean and variance equal to those of the innovation sample, for an ensemble of 1000 realizations. Values of -0.001 ± 0.001 for the mean sample skewness and of -0.005 ± 0.002 were found. This means that the values of skewness and kurtosis of the innovation sample differ signif-

icantly from the values expected for a sample of the same size that is normally distributed.

From the above results we conclude that the $T + 1$ forecast ensemble standard deviation (also denoted as 'spread') for surface precipitation represents a reasonable estimate of the standard deviation of the difference between the ensemble mean (in observation space) and the observations, which is usually taken as
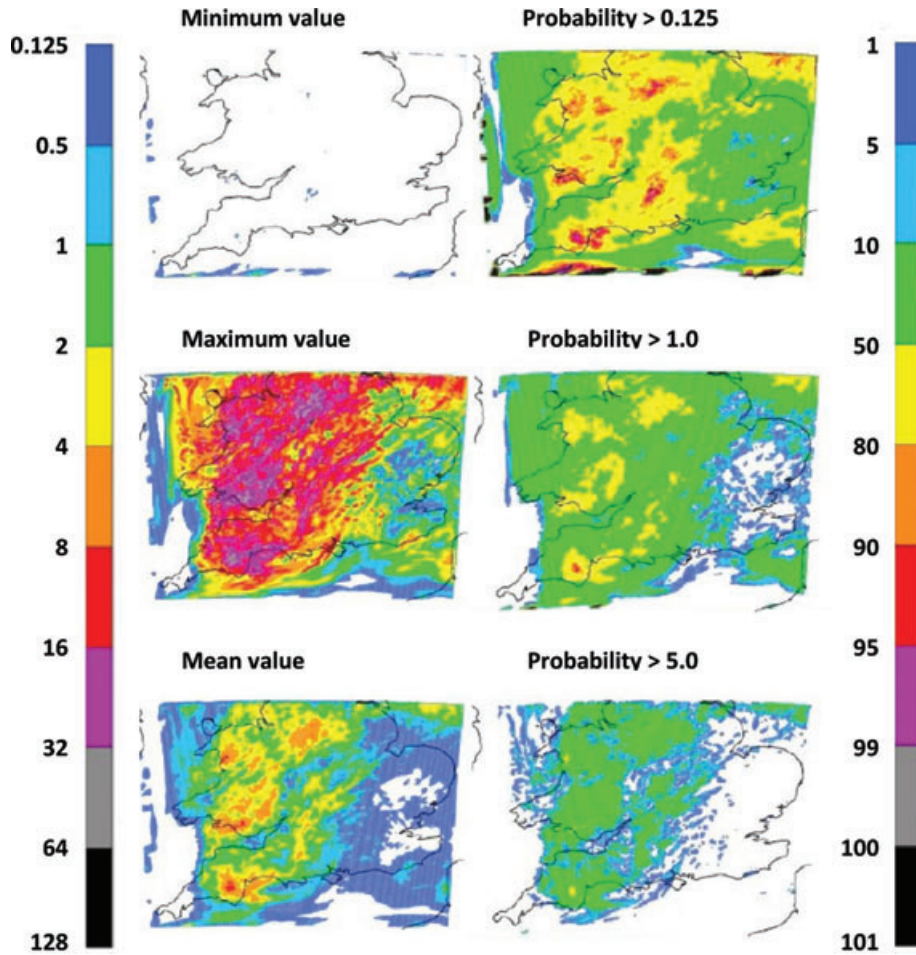
Fig. 14. Minimum, maximum and mean value over the forecast ensemble (right panels) and probability of occurrence for different threshold values (left panels) of surface rainfall rate at 1200 UTC on 26 July 2007, for experiment C2E01.
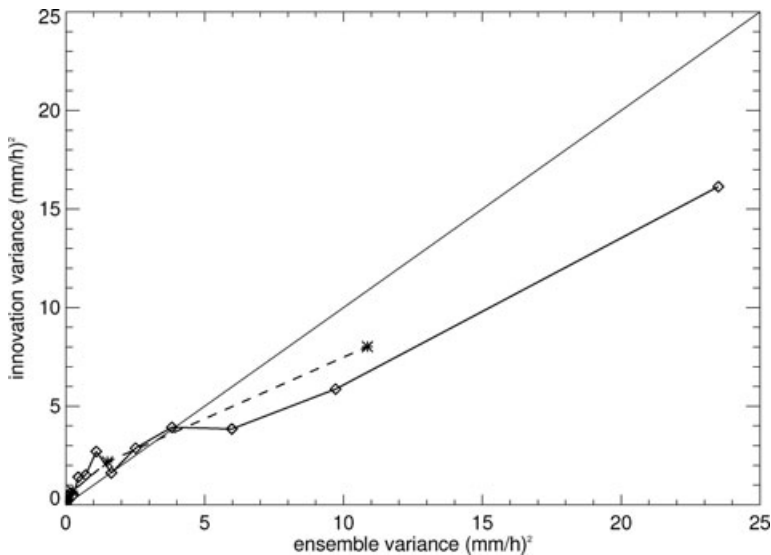

Fig. 15. Scatter plot of innovation variance against ensemble variance values for observations and forecasts of surface precipitation rate valid between 1030 UTC and 1730 UTC on 26 July 2007 for experiment C2E01, for a total of 6188 data point. Diamonds joined by a solid line represent the results when the data are binned into 20 bins, while asterisks joined by a dashed line represent the results when the data are binned into 5 bins. A reference line of unit gradient has also been plotted.

Table 4. Number of observations within every hour centred between 0700 UTC and 1700 UTC on 26 July 2007, that are simulated over the set of the experiments performed in this work

	07	08	09	10	11	12	13	14	15	16	17
Surface pressure	82	83	81	85	83	83	81	82	83	84	84
Surface temperature	42	42	54	54	53	54	40	40	54	53	54
Surface relative humidity	76	77	75	77	76	76	74	75	76	77	77
Surface wind u-component	80	80	90	93	91	92	78	78	93	92	93
Surface wind v-component	80	80	90	93	91	92	78	78	93	92	93
Aircraft temperature	55	54	109	57	24	56	45	37	37	29	63
Aircraft wind u-component	55	54	108	56	24	57	43	37	36	29	63
Aircraft wind v-component	55	54	108	56	24	57	43	37	36	29	63
15 km resolution radar rain rate	3536	3536	3536	3536	3536	3536	3536	3536	3536	3536	3536
15 km resolution radar rain rate >1 mmh^{-1}	440	590	820	930	1049	973	888	804	676	448	249
15 km resolution radar rain rate modified nudging interval	2652	2652	2652	2652	2652	2652	2652	2652	2652	2652	2652
15 km res radar rain rate > 1 mmh^{-1}; modified nudging interval	320	427	604	682	795	745	665	615	515	369	192
4.5 km resolution radar rain rate	44368	44368	44368	44368	44368	44368	44368	44368	44368	44368	44368
4.5 km resolution radar rain rate >1 mmh^{-1}	5513	6987	9507	11013	12949	12473	10854	9599	8581	6184	3649

an estimate of the skill of the forecast ensemble. However, the statistically significant deviation from Gaussianity of the innovations decreases the importance of the spread as a predictor of the skill of the surface precipitation forecast ensemble mean. This calls for the need to evaluate the performance of an ensemble of forecasts of precipitation by means of transformed variables that are normally distributed, although the choice of which transformation to use is a non-trivial task.

A substantial discrepancy between the magnitude of the observation error standard deviation inferred from the innovations using the above procedure and that used within the ETKF was found: the square root of the former value is only about one third of the minimum value of precipitation error standard deviation used in this work for the radar-derived surface precipitation observations within the ETKF (Section 2.2). However, as discussed in Section 2.2, the choice for determining the radar-derived surface precipitation observation error standard deviation used in the ETKF was driven by the need to make the precipitation innovation distribution more Gaussian, at the cost of introducing correlations between forecast and observation errors. This means that our estimate of observation error standard deviation for precipitation—which assumes observation errors are independent of forecast errors—which are used within the ETKF need not be consistent.

In Fig. 16, the distribution of the observations is compared with the distribution of the $T + 1$ forecast ensemble members for all precipitation observations, through a temporal series of rank histograms. Note that here the observation error added to each forecast ensemble member is drawn from a Gaussian distribution with zero mean and variance equal to 0.1 mm^2 h^{-2}.

At the beginning of the experiment, the histograms seem to show that the forecast error probability sampled by the forecast ensemble has a larger spread around the mean than the observation distribution—resulting in an overpopulation of the middle ranks—and, at the same time, a lower occurrence of extreme values than the observation distribution, resulting in a overpopulation of the extreme ranks. This may be a consequence of the forecast error distribution and the observation distribution being non-Gaussian and of the forecast error distribution having a smaller kurtosis than the observation distribution. The former point is consistent with our findings regarding the innovation distribution. At later times in the experiment, the increasing magnitude of the inflation factor, Fig. 10, seems to have the effect of decreasing the occurrence of observations with extreme values, making the forecast error distribution overspread. It is interesting to note that these findings seem to be not entirely consistent with those shown in Fig. 15, where only a minority of mean ensemble variances are greater than the corresponding innovation variances. This is likely to be a consequence of the innovations being non-normally distributed, making the mere comparison of the innovation variance with the ensemble variance presented in Fig. 15 inadequate to characterize fully the skill of the forecast ensemble.

4.2.2. Experiments C2E02, C2E03, C2E04 and C2E05. Here the results of four different experiments, with configuration identical to experiment C2E01 except for the inclusion of precipitation data are discussed. Experiment C2E02 includes the use of 15 km resolution simulated precipitation data for the calculation of the ensemble transform matrix **T**. From Table 4 it is evident that among all observation types considered in the ETKF
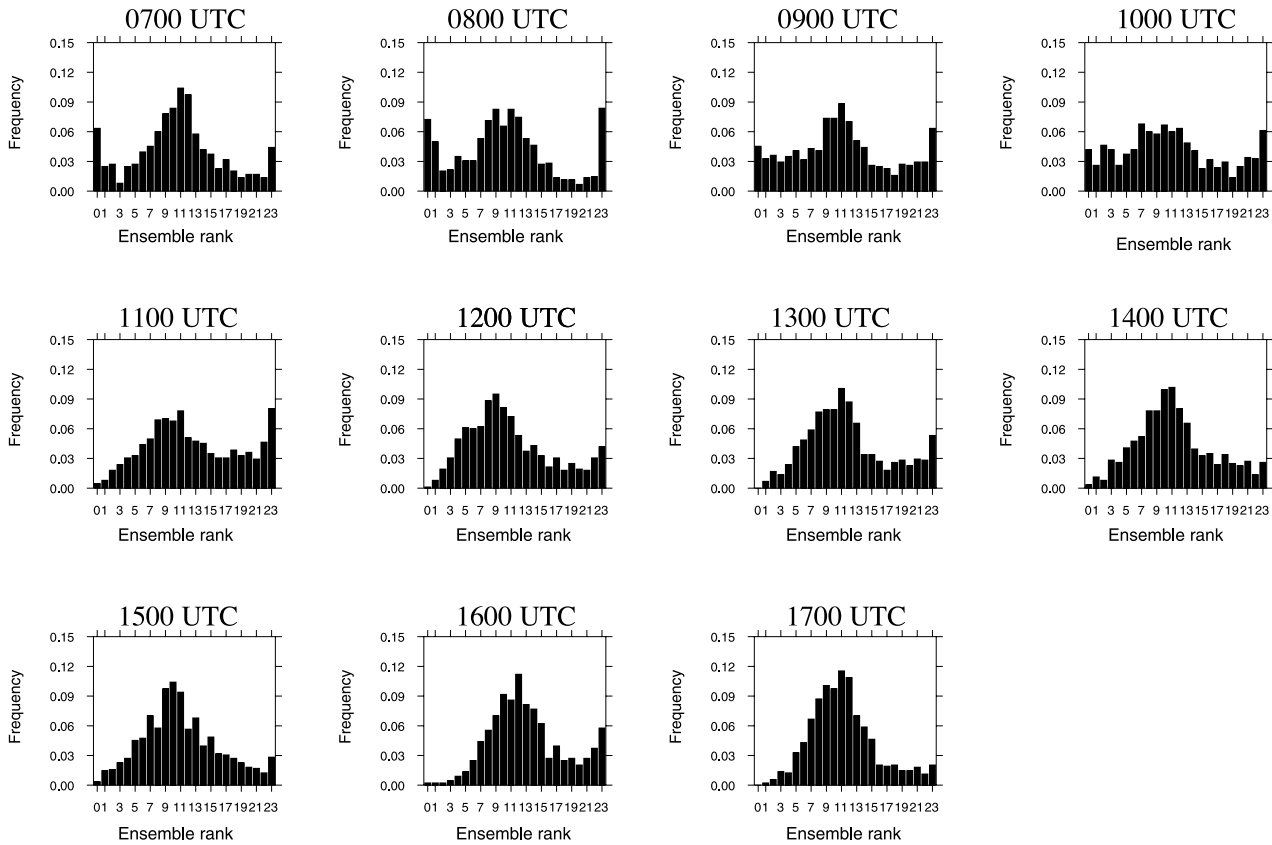
Fig. 16. Rank histogram for surface precipitation rate forecasts and observations at between 0700 UTC and 1700 UTC on 26 July 2007, for experiment C2E01.

calculations, rainfall observations from radar are by far the most numerous.

As shown in Fig. 17, the inflation factor found for experiment C2E02 is about twice as large as for experiment C2E01 (cf. Fig. 10), and it takes longer to reach a quasi-stationary value. Experiments C2E01 and C2E02 still indicate the need for larger inflation values to explain the variability due to surface observation data only.

First, the differences in the analysis perturbations between experiment C2E01 and C2E02 that result from the use of precipitation data are investigated. The perturbation of the total (i.e. kinetic plus available potential) energy density with respect to an isothermal atmosphere at rest is given by (e.g. Bowler et al., 2008, their eq. 7, modified to consider potential temperature rather than temperature perturbations)

$$E = \frac{1}{2}\rho_0(u'^2 + v'^2) + \frac{RT_0\rho_0}{2\kappa}\frac{\theta'^2}{\theta_0^2} + \frac{RT_0\rho_0}{2\gamma}\frac{p'^2}{p_0^2}, \qquad (6)$$

where $T_0 = 300$ K, p_0 is here the pressure corresponding to the ensemble mean Exner pressure $\overline{\Pi}$, $\theta_0 = T_0/\overline{\Pi}$ and $\rho_0 = p_0/(RT_0)$ with R being the specific gas constant. Note that the vertical velocity perturbations have

been neglected as here they are not determined by the ETKF.

The energy column density perturbation *EC*—resulting from the integration of E in the vertical between the bottom and the top model level by assuming hydrostatic balance—for the ensemble-mean total energy perturbation at 1200 UTC from experiments C2E01 and C2E02 are presented in Fig. 18, panels (a) and (b), respectively. As expected, the results show that the region that presents the most variability is the part of the domain that is better observed, mainly inland. To investigate the significance of the differences in energy between the two experiments, and to try to relate them to the impact of predicted observations of precipitation from radar used in the ETKF for C2E02, the energy difference $EC_{C2E02} - EC_{C2E01}$ was calculated (not shown). Also, the signal-to-noise ratio of predicted observations from radar—defined as the forecast error standard deviation in observation space (e.g. interpolated at observation location and time) divided by the observation error standard deviation—was computed (Fig. 18, panel c). These quantities are closely related to the singular values of **E** that appear in the transform matrix **T** (eqs. 2 and 3). The interpretation of the energy difference results is complicated by the effects of the use of different inflation factors for experiment C2E01 and C2E02, as well as the use of
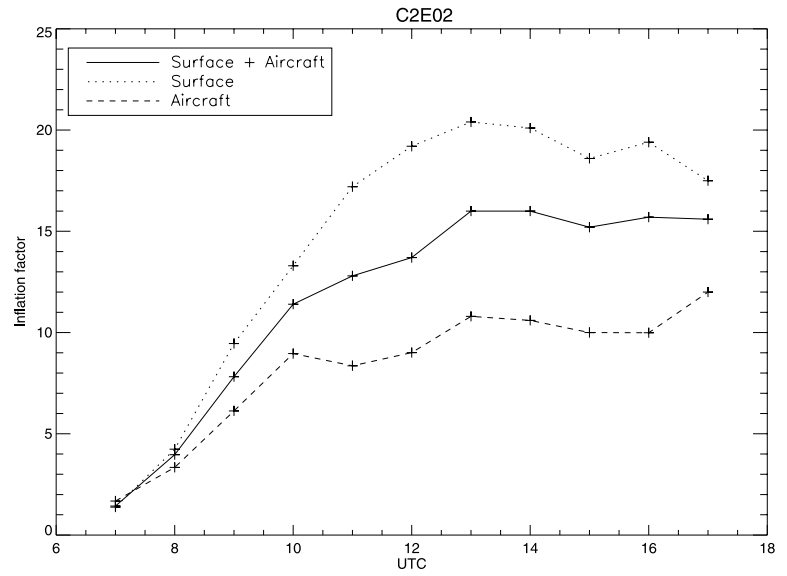
Fig. 17. Inflation factor values used within experiment C2E02, determined by using surface and aircraft observations (solid line), surface observation only (dotted line) and aircraft observation only (dashed line).
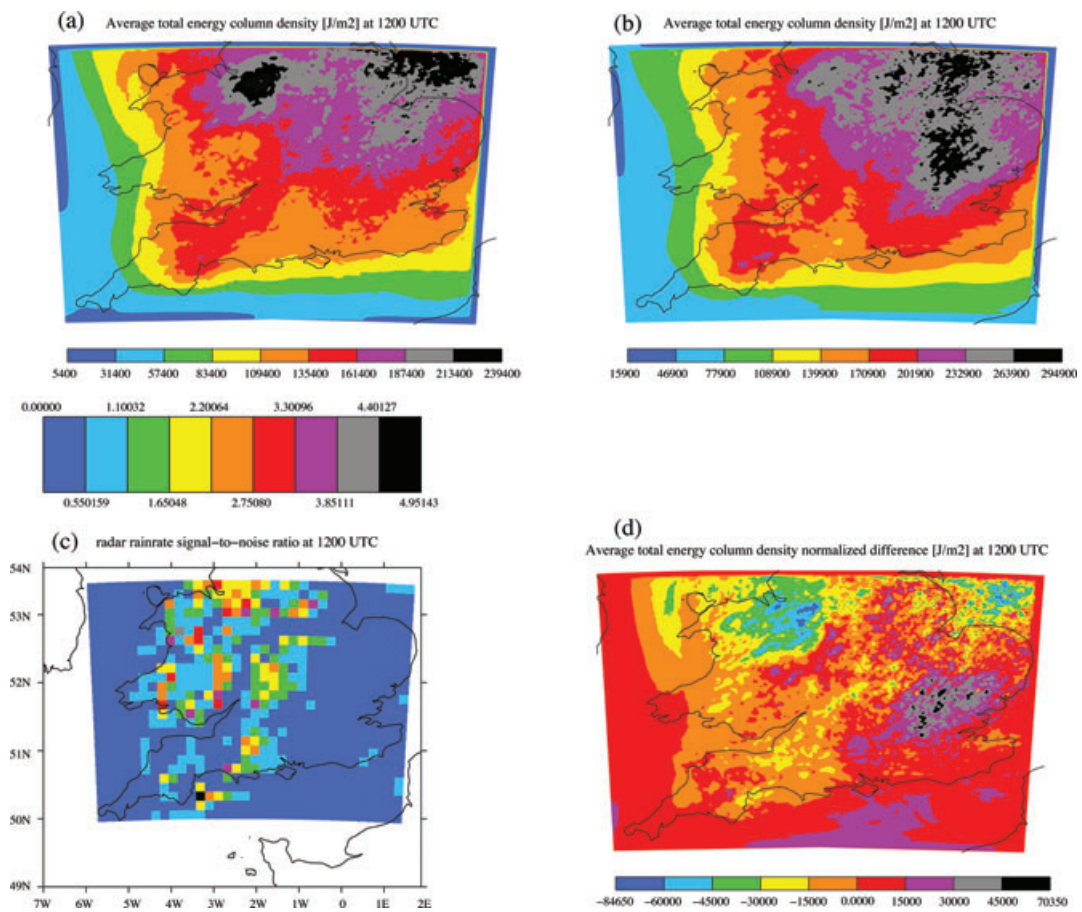

Fig. 18. Average total energy column-density perturbation (in J m^{-2}) at 1200 UTC from experiment C2E01 (panel a) and from experiment C2E02 (panel b); radar rain-rate signal-to-noise ratio at 1200 UTC (panel c) and average total energy column-density perturbation normalised difference (in J m^{-2}) between experiments C2E02 and C2E01 (panel d).

different initial conditions which may also introduce differences in the way features from the LBCs are advected into the domain. An attempt to compensate for these differences was made by multiplying EC_{C2E02} with the square of the ratio between the inflation factor used for C2E01 (equal to 5.04) and that used for C2E02 (equal to 13.7) at 1200 UTC. This attempt still did not produce an energy difference with patterns that match panel c as the new energy difference was dominated by the EC_{C2E01} values, whose magnitude are always larger than those of EC_{C2E01}. This is a result of the cumulative effect of the inflation factor in time and it is of little use to account only for the different inflation factor at 1200 UTC. A more meaningful approach is arguably to first calculate the spatial average of the EC_{C2E01} and EC_{C2E02} values, denoted as $\overline{EC_{C2E01}}$ and $\overline{EC_{C2E01}}$, respectively. Then multiply each EC_{C2E02} value by the ratio $\overline{EC_{C2E01}}/\overline{EC_{C2E02}}$ (equal to 0.755). This procedure effectively scales each C2E03 perturbation such that the energies $\overline{EC_{C2E01}}$ and $\overline{EC_{C2E01}}$ are the same. The resulting difference is shown in Fig. 18(d), which we refer to as the normalized energy difference. The areas where negative values of the normalized energy difference are found, roughly correspond to the regions where the signal-to-noise of radar observations is highest. This is consistent with the expected effect of the predicted observations from radar on the ensemble of analysis perturbations for experiment C2E02 (eq. 1), as the average total energy of the ETKF perturbations is proportional to the analysis error variance.

In Fig. 19, the 24-member ensemble of precipitation forecasts valid at 1200 UTC for experiment C2E02 are shown. Except for the control forecast (which is not affected by the ETKF perturbations), each ensemble member presents some differences from the corresponding member shown for C2E01 in Fig. 13. This is the result, valid at one-hour forecast lead time, of using eq. (1) at 1100 UTC with a different $\mathbf{T}$ matrix. However, this cannot be taken as an indication of differences in the information contained in the two forecast ensembles, given that $\mathbf{T}$ is not unique, as the analysis error covariance is invariant when $\mathbf{T}$ is multiplied on the right by an arbitrary orthonormal matrix (e.g. Sakov and Oke, 2008). A more meaningful way to investigate this aspect is to compare the characteristics of the probability distribution of precipitation forecast error at a given time for experiments C2E01 and C2E02. By comparing Fig. 14 for C2E01 with Fig. 20 for C2E02 it is possible to see that the areas having probability greater than 50% of being affected by precipitation with intensity greater than $0.125 \, \mathrm{mmh}^{-1}$ are quite similar between the two experiments. However, areas having probability greater than 90% to be affected by precipitation intensity greater than $0.125 \, \mathrm{mmh}^{-1}$ are larger in the case of experiment C2E02, and for these locations the radar shows presence of rain. Also, experiment C2E02 shows larger areas with zero probability of precipitation greater than $1 \, \mathrm{mmh}^{-1}$ over the South East of England, where the radar measurements show no precipitation or precipitation lower than $1 \, \mathrm{mmh}^{-1}$. Figure 21 shows the rank histograms for forecasts of precipitation rates for C2E02.

Comparing Figs 16 (for C2E01) and 21 though, shows that the inclusion of simulated precipitation observations does not produce any substantial improvement in the skill of precipitation forecasts.

Motivated by the results obtained in experiment C2E02, we decided to perform an experiment (denoted as C2E03) where the ensemble transform matrix included information from surface and aircraft data and from simulated precipitation data corresponding to precipitation observations with intensity greater than $1 \, \mathrm{mmh}^{-1}$ only. However, an inspection of the rank histograms for C2E03 (not shown) produced no evidence of significant changes due to the effect of discarding low or zero precipitation intensity observations with respect to the baseline experiment: the only differences are reasonably due to the use of different random numbers to account for observation error in the two rank histograms.

It is also interesting to check the sensitivity of our results on the spatial resolution of predicted precipitation data used in the ETKF. As explained in Section 2, the radar data that are used within the latent-heat nudging scheme are presented on a 15 km resolution grid, while their original resolution is nine times higher. A new experiment, denoted as C2E04, was devised to study the effects of using predicted precipitation data in the ETKF at a 4.5 km resolution, achieved by averaging the original model precipitation field over a 3×3 grid box for each ensemble member. For this experiment, at each ETKF cycle, 11 092 predicted precipitation observations are considered. The rank histogram results for C2E04 (Fig. 22) show a larger forecast bias for high intensity precipitation observations between 0900 UTC and 1300 UTC, with respect to C2E01. This larger bias may simply be an effect of an insufficient inflation factor before 1300 UTC, when the inflation factor reaches his peak value of about 60, from an initial value of the order of one at 0700 UTC. Note, incidentally, that the need for such a large inflation factor (about four times the peak value for experiment C2E02) may arise from the use of too few ensemble members to represent the distribution of forecast errors in observation space, compared to the rank of the observation error covariance matrix (eq. 3).

From eqs (1) and (2) it follows that when $\mathbf{\Gamma}$ is the zero matrix, the analysis ensemble coincides with the forecast ensemble (assuming $\Pi = 1$). From eq. (3) it is possible to see that this happens when forecast error variances in observation space are negligible with respect to observation error variances, when projected along the directions of the right singular vectors of $\mathbf{E}$. This means that, for a given forecast spread, the larger the observation error variance, the closer the analysis error is to forecast error spread.

From the above discussion it follows that the initial condition ensemble spread determined by the ETKF is sensitive to the magnitude of the observation error variance. As from eq. (5), the precipitation observation error standard deviation is determined by means of an empirical coefficient c, set initially equal to 1. A
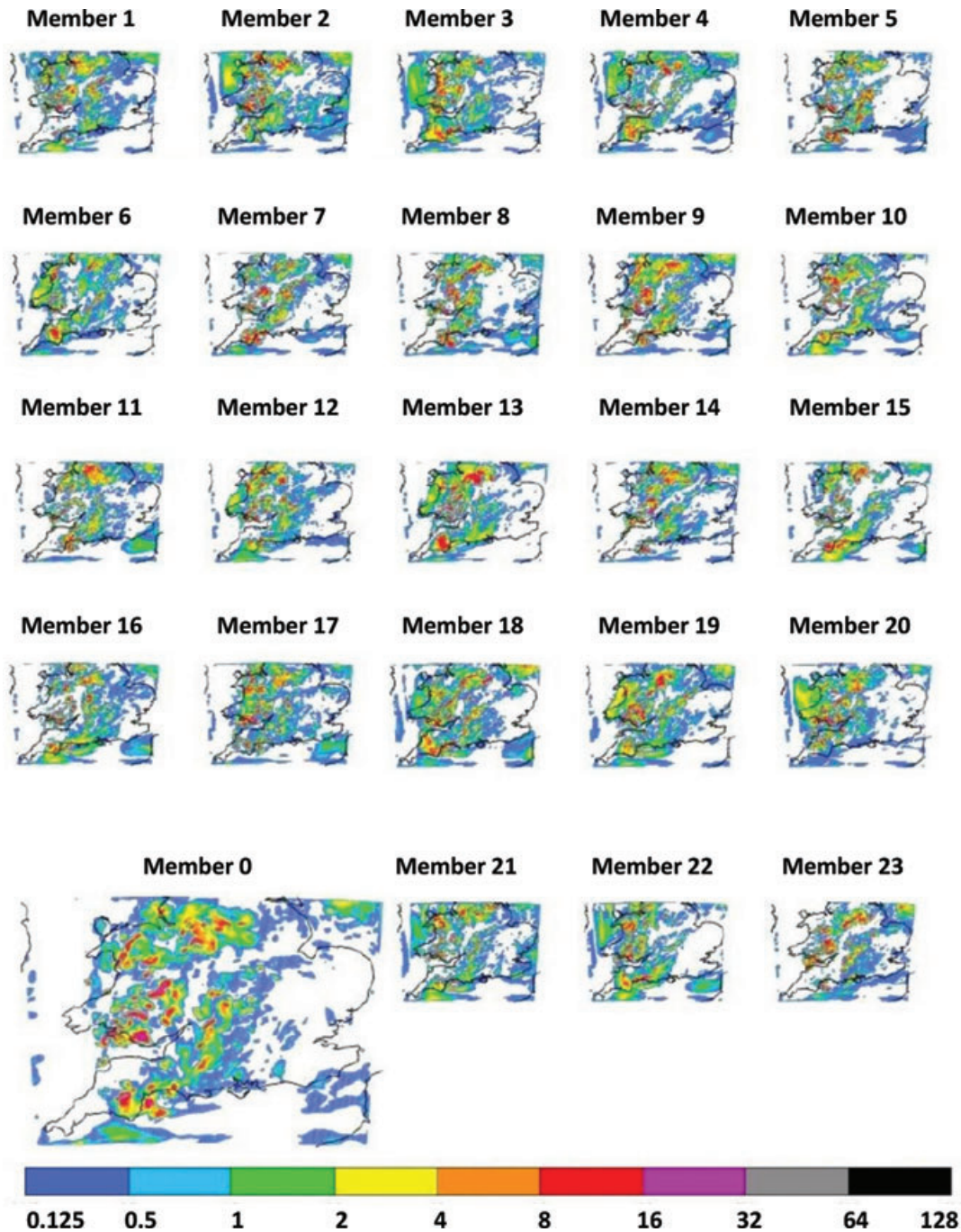
Fig. 19. Ensemble forecasts of surface precipitation rate (in mmh^{-1}) valid at 1200 UTC on 26 July 2007, for experiment C2E02. Member 0 is the forecast from the previous analysis.

new experiment, denoted as C2E05, was designed to investigate the sensitivity of the ensemble forecast spread to the characteristics of precipitation observation error. For this experiment, a value of $c = 2$ was chosen. However, over the whole duration of the experiment, a doubling of the precipitation observation error standard deviation seems not to change significantly the shape of the forecast error distribution with respect to the observation

distribution. This can be seen from the series of rank histograms for experiment C2E05 (not shown), where, for consistency with the previous case-2 experiments, the variance of the random observation error added to the forecast ensemble for the rank histograms was kept equal to 0.1 mm^2 h^{-2}.

The results for the experiments presented in this section can be summarized in a quantitative manner by means of the Brier
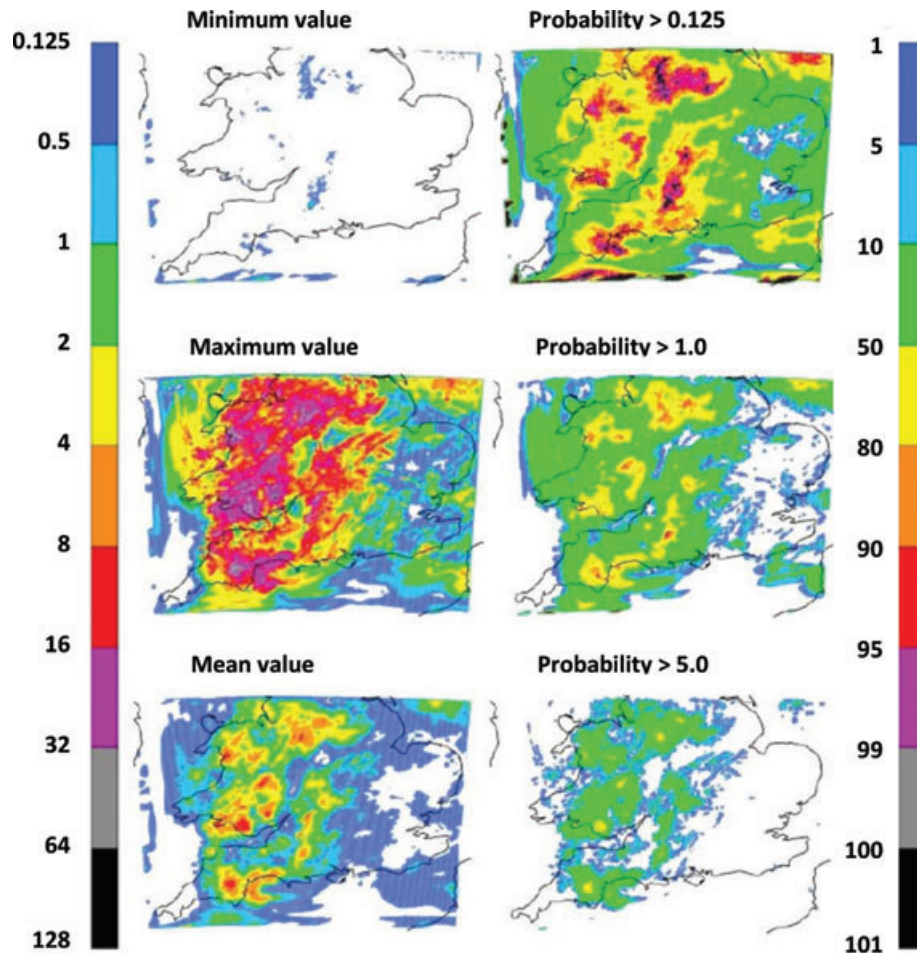
Fig. 20. Minimum, maximum and mean value over the forecast ensemble (left panels) and probability of occurrence for different threshold values (right panels) of surface rainfall rate at 1200 UTC on 26 July 2007, for experiment C2E02.

Score (BS) for precipitation. A lower BS indicates a better set of ensemble forecasts. The BS results for events defined by the occurrence of observations with precipitation rate greater than 0.0, 1.0 and 5.0 mmh^{-1} for experiment C2E01 are shown in Fig. 23. The score for the 0.0 mmh^{-1} precipitation rate observations threshold improves (i.e. decreases) during the experiment, consistently with a improvement seen also in other indicators, such as the rank histogram results. This improvement during the experiment is 'robust', given that it happens when more observations measure precipitation, that is when the rain band starts crossing the domain. The results found for the BS at higher thresholds may seem surprising: the larger the threshold, the better the score. However, this result is likely to be a consequence of the relative scarcity of precipitation observations of moderate or high intensity, so that the BS for higher thresholds is dominated by the results for non-occurring events.

Given that we are here interested in determining the effect on forecast skill of including precipitation in the ETKF calculations, it is useful to define a precipitation skill score (denoted as PSS)

as

$$\text{PSS}_{ij} = 1 - \frac{\text{BS}_{\text{C2E0}i}}{\text{BS}_{\text{C2E0}j}} \qquad (7)$$

for experiment C2E0i, calculated with respect to the BS for experiment C2E0j. Note that perfect forecasts imply $\text{BS}_{\text{C2E0}i} = 0$ and $\text{PSS}_{ij} = 1$, when observation error is not considered.

The PSS results for experiments C2E02, C2E03, C2E04 and C2E05 with respect to C2E01, shown in Fig. 24 for a 0.0 mmh^{-1} precipitation threshold, indicate a limited impact on forecast skill of the precipitation data used within the ETKF calculations. The temporal variations of skill are very similar in the case of experiments C2E02, C2E03 and C2E05, with variations confined between ±10%. Worse results are obtained for experiment C2E04—where the number of simulated precipitation observations is more than 12 times larger than for the other three experiments—especially between 0900 UTC and 1300 UTC, consistently with the rank histogram results comparison between experiments C2E04 and C2E01. Finally, the PSS results for a 1.0 mmh^{-1} threshold (not shown) show a consistently lower skill
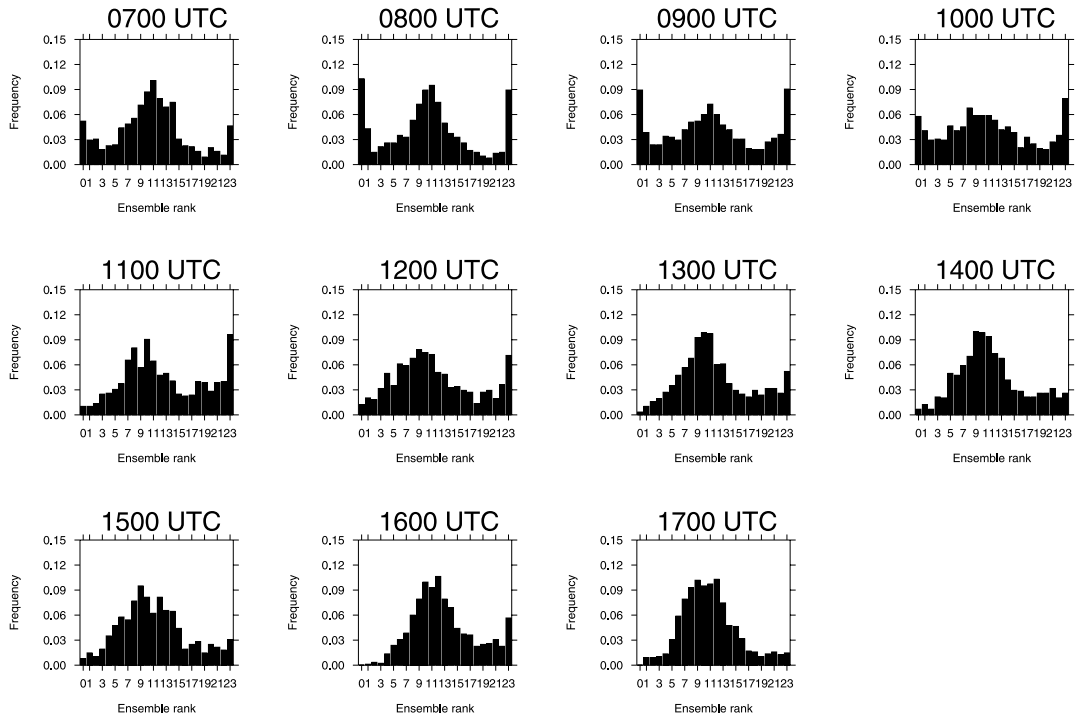
Fig. 21. Rank histograms for surface precipitation rate forecasts and observations between 0700 UTC and 1700 UTC on 26 July 2007, for experiment C2E02.
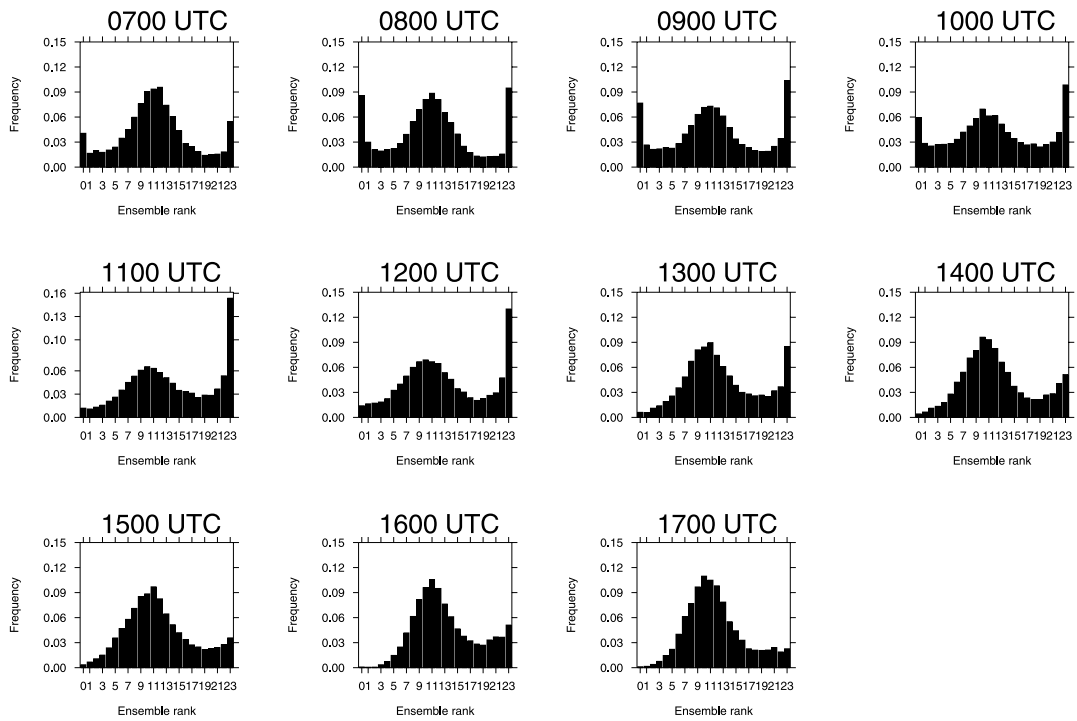

Fig. 22. Rank histograms for surface precipitation rate forecasts and observations between 0700 UTC and 1700 UTC on 26 July 2007, for experiment C2E04.
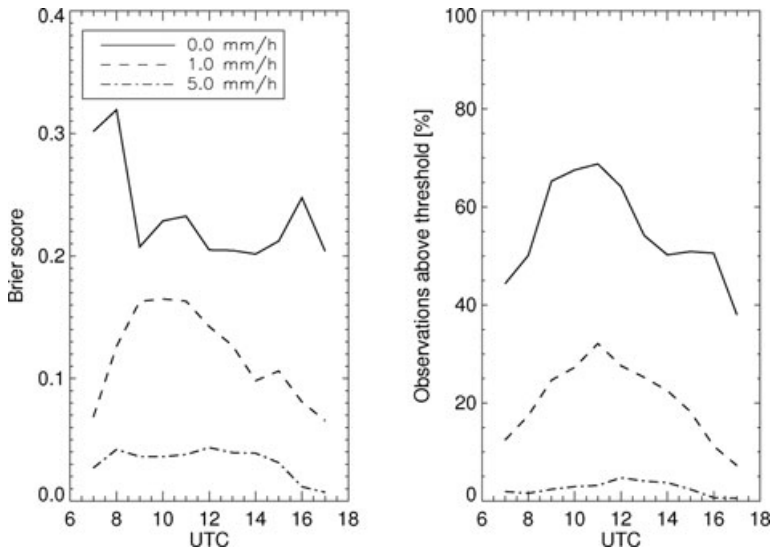
Fig. 23. Brier score results for experiment C2E01, for precipitation events with intensity greater than 0.0 (solid line), 1.0 (dashed line) and 5.0 (dash-dot line) mmh^{-1} (left-hand panel); number of precipitation events (%) in experiment C2E01 with intensity greater than 0.0 (solid line), 1.0 (dashed line) and 5.0 (dash-dot line) mmh^{-1} (right-hand panel).
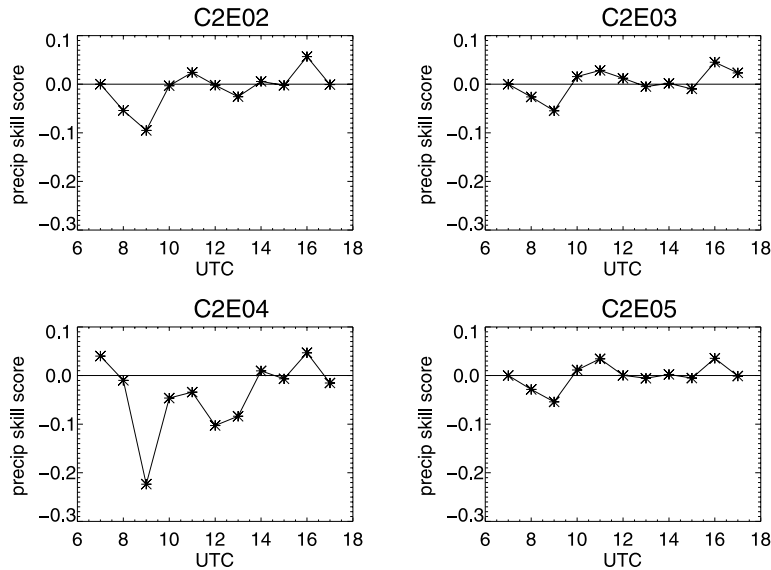

Fig. 24. Precipitation skill score (PSS, defined in eq. 7) with respect to experiment C2E01, for zero precipitation intensity threshold, for experiment C2E02 (top-left panel), C2E03 (top-right panel), C2E04 (bottom-left panel) and C2E05 (bottom-right panel).

for each experiment, with respect to that for zero precipitation threshold. Similar results are obtained for a 5.0 mmh^{-1} threshold (not shown), with the exception of experiment C2E04, which proves to be more skilful at this threshold than at the two lower ones, before 1200 UTC. However, this surprising higher skill of C2E04 for the higher precipitation threshold (at odds with the rank histogram results shown in Fig. 22) may merely be the effect of better (i.e. lower) Brier score results for non-occurring events: the quadratic nature of the Brier score makes the score 12 times smaller for 12 times more events with the same forecast probability of occurrence

4.2.3. Experiments C2E06 and C2E07. As discussed in Sections 4.2.1 and 4.2.2, from Figs 10 and 17 it is possible to see a clear dependence of the inflation factor magnitude on the type of observation that is used to determine the inflation. It is inter-

esting to check the forecast performance when only surface data are used to calculate the inflation factor, both in the case when no precipitation is used in the ETKF (experiment C2E06) and when predicted precipitation observations are included in the ensemble transform matrix calculations (experiment C2E07).

In Fig. 25, the inflation factor values for experiments C2E06 and C2E07 are shown. As expected, the inflation factors obtained when only surface data are taken into account are very similar to those shown in Figs 10 and 17 in the case when only surface data are considered at each given time.

To determine the impact of using the surface-observation-based inflation factor, the PSS results for experiment C2E07 were calculated with respect to experiment C2E01 and C2E06, which are shown in Fig. 26. The results are very similar in the two cases, and also similar to the PSS results for C2E02 with
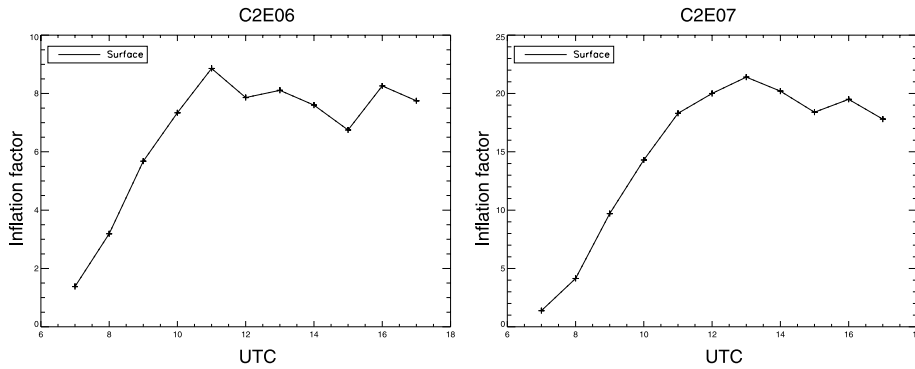
Fig. 25. Inflation factor values used within experiments C2E06 (left panel) and C2E07 (right panel), determined by using surface observations.
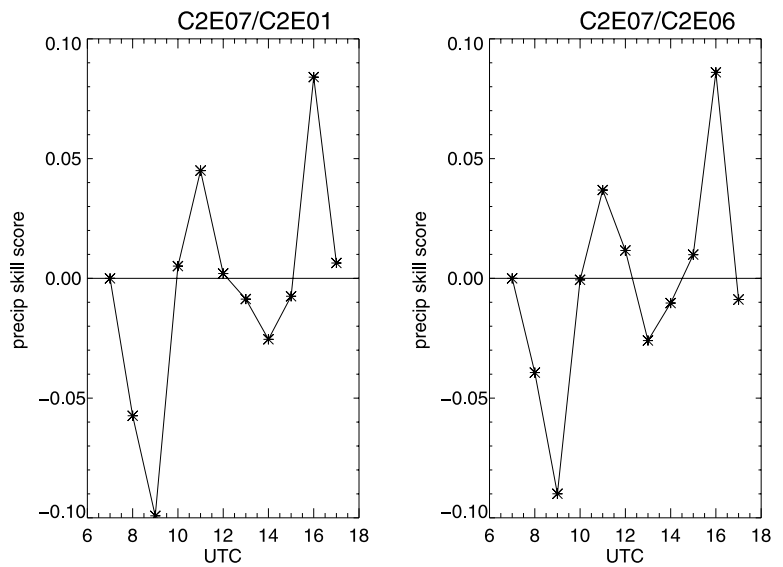

Fig. 26. Precipitation skill score of experiment C2E07, with respect to experiments C2E01 (left-hand panel) and experiment C2E06 (right-hand panel).

respect to C2E01, so that no evident benefit is experienced owing to the use of a larger inflation factor.

4.2.4. Experiments C2E08, C2E09, C2E10 and C2E11. As summarized in Table 3, these experiments are all characterized by the use of the same boundary conditions at a given time, for all ensemble members. The boundary conditions are derived from forecasts on a 4 km resolution grid over the whole UK with a 15-min temporal resolution, updated every 6 h. In Fig. 27, the BS for precipitation events with intensity greater than 0.0 mmh^{-1} for these experiments is shown, along with the same BS for experiment C2E01. Except for the 0800 UTC case, when the BS for C2E08 and C2E10—which do not include precipitation data in the ETKF calculations—is slightly lower, the precipitation forecast skill when higher resolution and deterministic boundary conditions are used in the simulations is lower than in the case of the baseline experiment. Consistently, between 1100 and 1500 UTC the PSS values for C2E09 with respect to C2E01 (Fig. 28, left panel) are substantially lower than those for experiments that
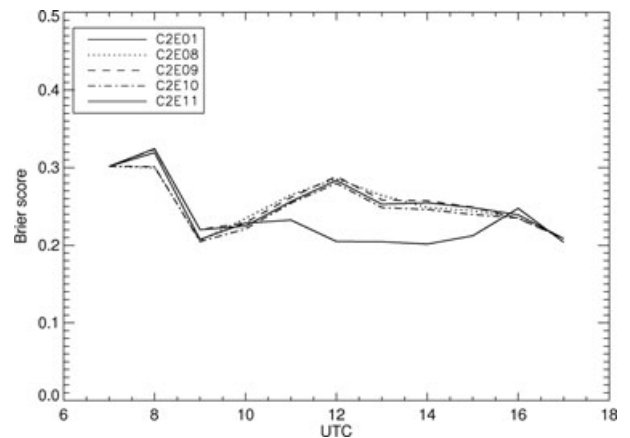

Fig. 27. Brier score for precipitation events with intensity greater than 0.0 mmh^{-1}, for experiments C2E01 (solid line), C2E08 (dotted line), C2E09 (dashed line), C2E10 (dash-dotted line) and C2E11 (dash-triple-dotted line).
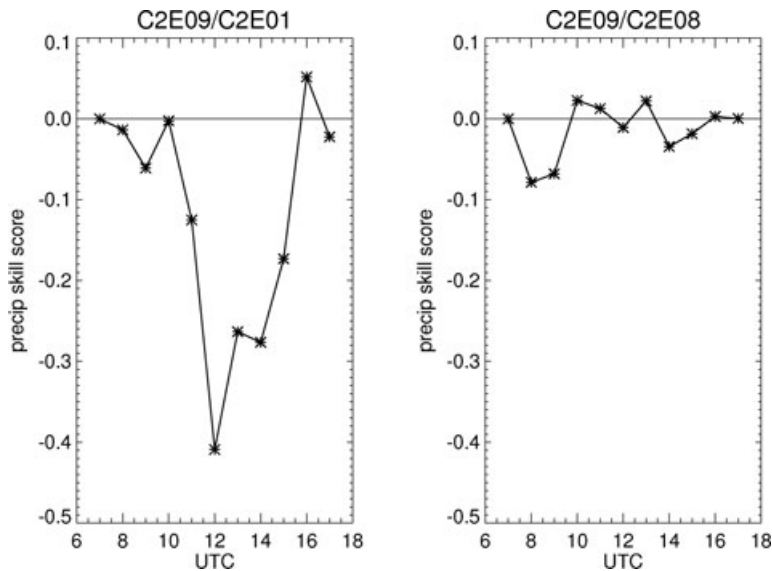
make use of perturbed boundary conditions. Not surprisingly, PSS values for C2E09 with respect to C2E08 (Fig. 28, right panel) are instead similar to those for C2E02 with respect to C2E01.

A possible explanation for the decrease of skill experienced when deterministic boundary condition are used for the case study under investigation, can be inferred from the inspection of Figs 14 and 29. Compared to experiment C2E01, experiment C2E08 shows over South West England larger areas with zero probability of precipitation with intensity greater than 1.0 and 5.0 mmh^{-1}, where instead the radar indicates the presence of intense rainfall (Fig. 5). At the same time, experiment C2E08 shows over the north eastern corner of the domain larger areas with non-zero precipitation probability with intensity greater than 5 mmh^{-1}, where the radar measures only light precipitation.

It is interesting to check whether the regions affected by rain, located in close proximity of the boundaries of the simulation domain, that can be seen in Figs 13, 14, 19 and 20, are still present in the case of one of the 4 km boundary condition experiment, such as C2E08. By comparing the 0.125 mmh^{-1} threshold precipitation probability values shown in Figs 14 and 29, it is possible to see that experiment C2E08 assigns zero precipitation probability on most of the region close to the west boundary of the domain. Over the same location, experiment C2E01 shows localized region with 100% probability of rainfall with intensity greater than the threshold. This confirms the conjecture that these features are artefacts due to the large resolution difference of the boundary conditions used in the two experiments and suggests that the resolution ratio of the nested model grids should relatively small. On the other hand, experiment C2E08 is characterized by having a larger proportion of features, mainly in the southwest part of the domain, that have a 100% probability of precipitation above a given threshold. This may indicate that

the features are being advected into the interior of the domain through the use of the same boundary conditions.[1]

4.2.5. Experiments C2E12, C2E13, C2E14 and C2E15. It might be argued that the effect of latent heat and cloud nudging on each member of the forecast ensemble may reduce the impact of including predicted precipitation observations in the ETKF. As a result of the latent heat nudging procedure, the vertical latent heat profile ensemble at a given time is scaled according to the ratio between the surface precipitation rate from observations and from the model. The result of the cloud nudging procedure is instead the forcing of the model relative humidity—through the addition of a forcing term to the specific humidity prognostic equation—towards relative humidity fields from a set of observations (Dixon et al., 2009). As discussed in Section 3, in the experiments described so far, a given ensemble forecast valid at time $T + 1$ is constrained by the cloud nudging procedure up to $T + 0.25$ and by the latent heat procedure up to time $T + 0.5$. This means that the specific humidity and the potential temperature—which affect the vertical velocity and consequently precipitation—evolve in an unconstrained fashion (and, potentially, increase spread) only during the last 45 and 30 min in the forecast, respectively.

In experiment C2E12 and C2E13, the cloud and latent heat nudging was stopped at time T and time $T + 0.25$, respectively.[2] In Fig. 30, the PSS values for experiment C2E13 with respect

---

[1] Note that the use of the same LBCs is also responsible for the very high (if not 100%) precipitation probability at the northern and eastern boundary of the domain (especially for the 0.125 mmh^{-1} threshold).

[2] It was decided to keep using the radar observations at time T, which nudge the model latent heat for the following 15 min, so as to be able to compare precipitation forecasts and observations at time T, as done for the previous experiments.
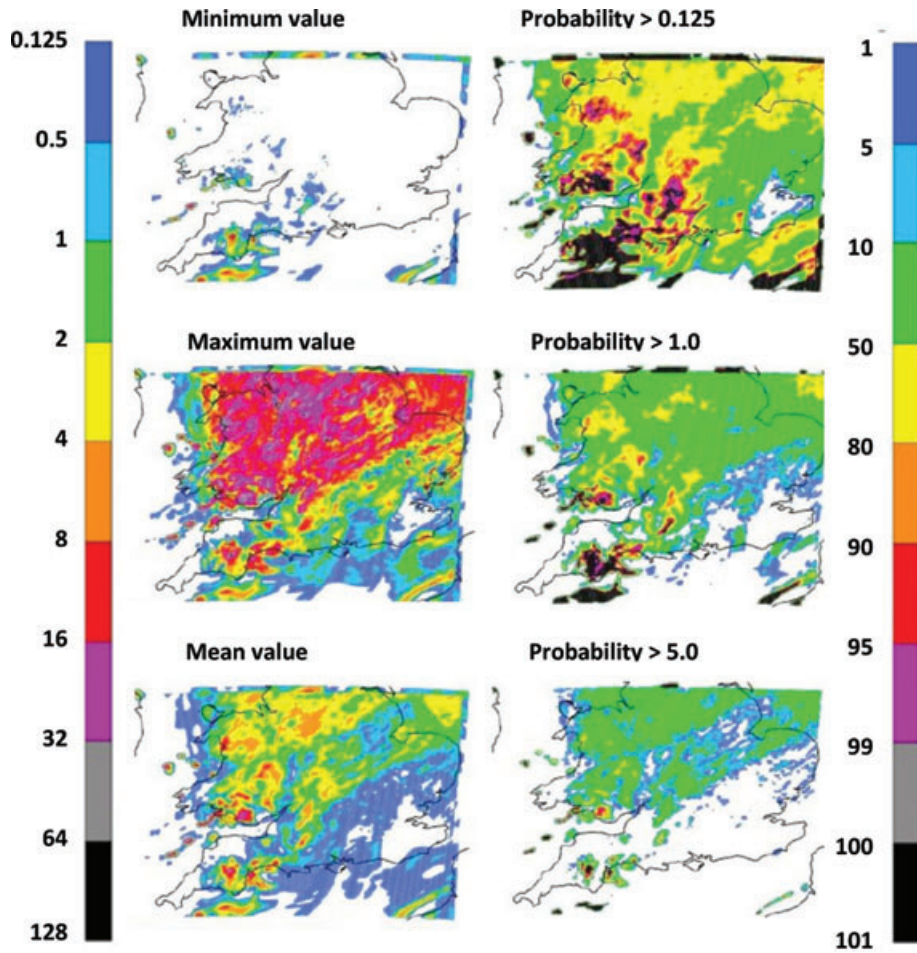
Fig. 29. Minimum, maximum and mean value over the forecast ensemble (left panels) and probability of occurrence for different threshold values (right panels) of surface rainfall rate at 1200 UTC on 26 July 2007, for experiment C2E08.
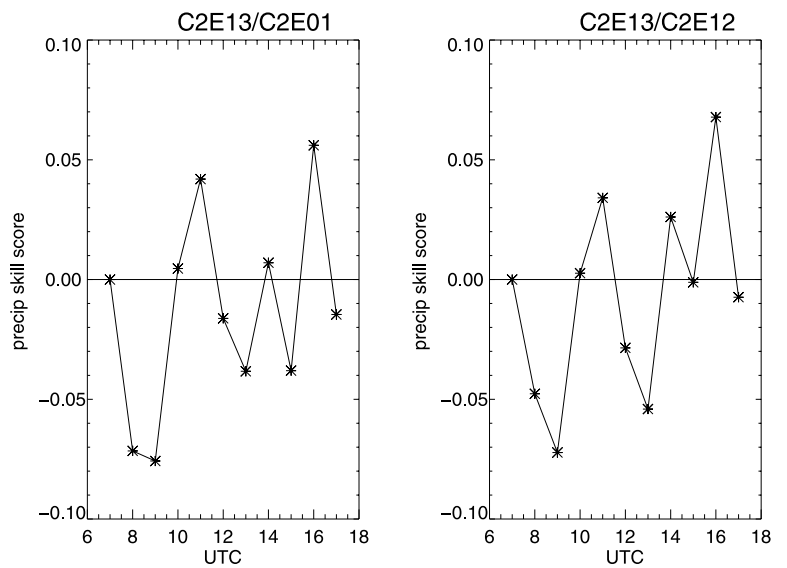

Fig. 30. Precipitation skill score of experiment C2E13, with respect to experiments C2E01 (left-hand panel) and experiment C2E12 (right-hand panel).
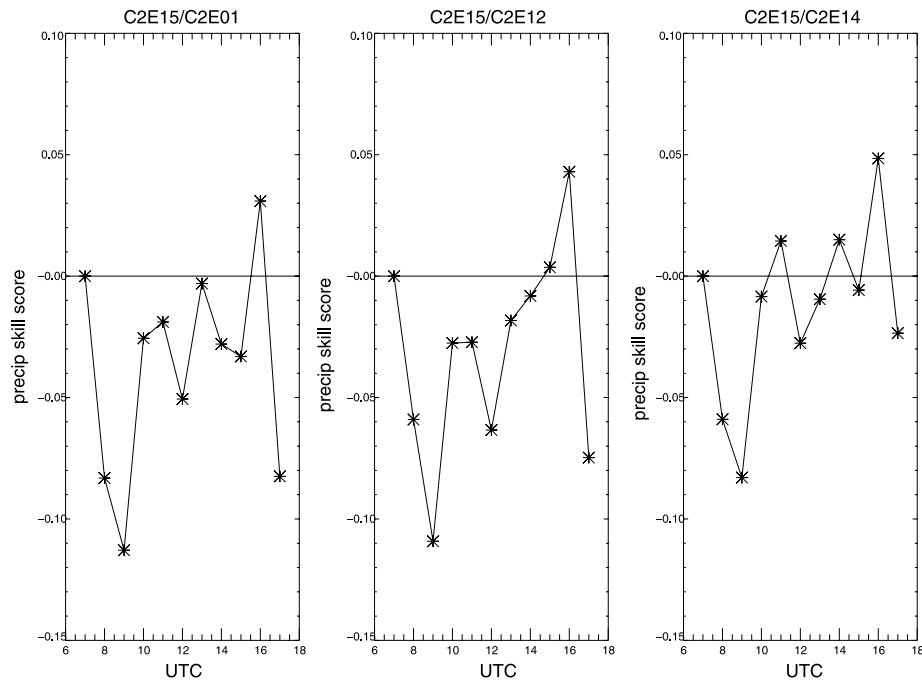
Fig. 31. Precipitation skill score of experiment C2E15, with respect to experiments C2E01 (left-hand panel), experiment C2E12 (middle panel) and experiment C2E14 (right-hand panel).

to experiments C2E01 and C2E12 are shown. Results are fairly similar to those obtained for previous experiments, the only notable difference being a slight increase of skill at 1400 UTC. Also, the modified nudging period does not produce evidence of improvements in precipitation forecast skill for precipitation threshold of 1.0 and 5.0 mmh^{-1} (not shown). The lack of sensitivity of the spread of precipitation forecasts at $T + 1$ on the duration of the latent heat nudging period is surprising. A shorter nudging period not only increases the period over which the evolution of specific humidity and potential temperature is unconstrained, but arguably also increases the degree of imbalance in the model fields, as the standard nudging implementation used at the Met Office constrains only the mass fields (namely, specific humidity and potential temperature) without adjusting the momentum fields. (When hydrostatic balance holds, a consistent coupling between corrections to the mass fields and to the wind fields is important for the prediction of precipitation even at high resolution, see e.g. Pagé et al., 2007) A possible explanation for this behaviour is that the length of the time interval needed to adjust the momentum field due to changes in the mass field is significantly shorter than the time interval during which the model state evolves in an unconstrained fashion.

Finally, in experiment C2E14 and C2E15, the modified nudging period was used in conjunction with the instantaneous addition at time T of the ETKF increments. As discussed in Section 3, for the experiments presented so far it was decided to avoid the introduction of relatively large discontinuities in the fields as a result of the addition of the ETKF increments to the model

forecasts, so as to minimize the possibility of introducing spurious inertia-gravity waves in the subsequent model forecasts. On the other hand, this meant that the ETKF increments were not fully added to the forecast ensemble until $T + 0.5$, thereby leading to a potential reduction of the spread that could be otherwise achieved at verification time $T + 1$, in the case when the increments were added more quickly, if not instantaneously. We decided to test this possibility by means of experiments C2E14 and C2E15, where the ETKF increments were added instantaneously. In Fig. 31, the precipitation skill score values for experiment C2E14 with respect to experiments C2E01 and C2E13 are shown. The PSS scores with respect to BS_{C2E01} and BS_{C2E12} are for experiment C2E15 consistently worse than the same scores obtained for experiment C2E13 (shown in Fig. 30), whereas the precipitation skill for experiment C2E15 with respect to the similarly configured experiment C2E14 is comparable to that obtained with other experiment configurations.

5. Summary and Conclusions

In this paper, an experimental ensemble prediction system for nowcasting purposes based on a convection-permitting version of the UM with a 1.5 km grid length over the southern UK is presented. An initial assessment of the performance of the system based on two summer time case studies is provided. The first case study is mainly devoted to assessing forecast spread as determined by the forecast ensemble, for a set of surface fields (namely, temperature, horizontal wind, relative humidity and

pressure). Our results show that, as the ETKF cycles progress, the forecast error distributions estimated by the forecast ensemble at lead times $T + 1$ become fairly consistent with the distribution of the innovations valid at the same time. This is a result of the forecast-error-variance-enlarging effect of the inflation factor. Problems still remain with the surface pressure forecast spread, which seems to be too large. However, this may simply be a consequence of assuming a surface pressure observation error variance which is too large. In light of this a more thorough validation of the ensemble system, presently ongoing, is needed to provide a more confident assessment of the ensemble forecast system.

The second case study is focused on investigating the surface precipitation forecast skill of the system and the effects of including predicted observations of precipitation in the calculation of the ensemble transform matrix **T** for an ETKF at convective scales. First, the 1-h precipitation forecast skill of the 'baseline' experiment C2E01 (defined in Table 3), which does not include predicted precipitation observations in the ETKF calculations, was evaluated. An analysis of the different precipitation forecast patterns that are characteristic of the different forecast ensemble members clearly shows the advantages of a probabilistic nowcasting perspective in terms of its ability to provide non-null probability of occurrence of precipitation at a given time T, over regions where it is possible to observe precipitation at the same time. However, there is also some evidence of spurious precipitation at the boundaries of the forecasting domain, which calls for the need to keep the nesting ratio between the lower resolution and the limited-area higher resolution models as small as possible. A more quantitative indicator of forecast skill can be determined by comparing the one-hour lead time forecast error variance with the innovation variance for precipitation intensity, after that the inflation factor has reached a quasi-stationary value. The comparison shows that the $T + 1$ ensemble standard deviation (denoted as 'spread') for surface precipitation provides a good estimate of the standard deviation of the difference between ensemble mean forecast (in observation space) and the observations (denoted as 'skill'). However, evidence of non-normally distributed precipitation forecast errors makes this result less robust. As a matter of fact, the inspection of the rank histograms seem to provide evidence of precipitation forecast error with spread that is too large.

The effects of the inclusion of predicted observations of precipitation intensity from radar within the predicted observation categories used to constrain the initial forecast error distribution were investigated for different experimental configurations. Experiment C2E02 shows that the inflation factor required to get realistic $T + 1$ precipitation forecast spread in the presence of predicted precipitation observations is about twice as large as that needed when only predicted surface and aircraft observations are considered in C2E01. In other words, the increase of the number of predicted observations in the ETKF due to the inclusion of predicted precipitation observations provides a sig-

nificantly worse estimate of $T + 1$ precipitation forecast spread, which requires an inflation factor about twice as large.

As expected, the ensemble-mean total energy column densities obtained with and without the inclusion of predicted precipitation observations at 1200 UTC show small scale variability in correspondence of the observation locations, as well as features advected into the domain from the LBCs or originating from the use of different initial conditions and inflation factor values. Also, the normalized difference (i.e. adjusted to account for the different spatial average of the energy column densities) between the C2E02 and C2E01 energy column densities is negative over locations where the observation signal-to-noise ratio is highest. This can be explained with the effect of predicted precipitation observations on reducing the analysis error variance, which is proportional to the average energy perturbations. A qualitative inspection of precipitation probability maps obtained at 1200 UTC shows that the use of predicted precipitation observations gives rise to more locations with very high probability of rain that are affected by precipitation with intensity greater than 0.125 mmh^{-1}. Also, we found some evidence of better performance of the forecast ensemble when perturbed boundary conditions are used, especially with a low nesting ratio. However, a detailed analysis of a number of skill indicators for the experiments described in Table 3 seems to indicate that no definite skill improvement for $T + 1$ probabilistic forecasts at convective scale is achieved when predicted precipitation observations are included in the ETKF formulation. From eq. (3) it follows that this finding is likely to be due to the detrimental effects of sampling noise associated with the use of too few ensemble members, resulting in a rank-deficient representation of the square root of the forecast error covariance matrix in observation space. To produce the hoped improvement in skill deriving from the use of radar data it is of paramount importance the rebalancing of the ratio between the number of predicted observations and number of ensemble forecast members used in the ensemble prediction system.

6. Acknowledgments

References

Bannister, R., Migliorini, S. and Dixon, M. 2011. Ensemble prediction with a convection-permitting model for nowcasting. Part II: Forecast error statistics. *Tellus* **63A**, this issue.

Bloom, S., Takacs, L., da Silva, A. and Ledvina, D. 1996. Data assimilation using incremental analysis updates. *Mon. Wea. Rev.* **124**, 1256–1271.

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

Clark, P. 2009. Limited area modelling in CRM mode. In: *Proceedings of the ECMWF Seminar on Parametrization of Subgrid Physical Processes*. ECMWF, 165–192. Available at: http://www.ecmwf.int/publications/.

Davies, T., Cullen, M. J. P., Malcolm, A. J., Mawson, M. H., Staniforth, A., and co-authors. 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Quart. J. Roy. Meteor. Soc.* **131**, 1759–1782.

Dixon, M., Li, Z., Lean, H., Roberts, N. and Ballard, S. 2009. Impact of data assimilation on forecasting convection over the United Kingdom using a high-resolution version of the Met Office Unified Model. *Mon. Wea. Rev.* **137**, 1562–1584.

Done, J. M., Craig, G. C., Gray, S. L., Clark, P. A. and Gray, M. E. B. 2006. Mesoscale simulations of organized convection: importance of convective equilibrium. *Quart. J. Roy. Meteor. Soc.* **132**, 737–756.

Fritsch, J. M. and Carbone, R. R. 2004. Improving quantitative precipitation forecasts in the warm season: a USWRP research and development strategy. *Bull. Amer. Meteor. Soc.* **85**, 955–965.

Gilmore, M. S., Straka, J. M. and Rasmussen, E. N. 2004. Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. *Mon. Wea. Rev.* **132**, 2610–2627.

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

Hamill, T. M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.* **129**, 550–560.

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

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

Jones, C. D. and Macpherson, B. 1997. A latent heat nudging scheme for the assimilation of precipitation data into an operational mesoscale model. *Meteor. Appl.* **4**, 269–277.

Koizumi, K., Ishikawa, Y. and Tsuyuki, T. 2005. Assimilation of precipitation data to the JMA mesoscale model with a four-dimensional variational method and its impact on precipitation forecasts. *SOLA* **1**, 45–48.

Lean, H. W., Clark, P. A., Dixon, M., Roberts, N. M., Fitch, A. and co-authors. 2008. The Met. Office global three-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteor. Soc.* **131**, 1759–1782.

Lilly, D. K. 1990. Numerical prediction of thunderstorms–has its time come? *Quart. J. Roy. Meteor. Soc.* **116**, 779–798.

Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingleby, N. B., Andrews, P. L. F. and co-authors. 2000. Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.* **136**, 3408–3424.

Macpherson, B., Wright, B. J., Hand, W. H. and Maycock, A. J. 1996. The impact of MOPS moisture data in the U.K. Meteorological Office Mesoscale Data Assimilation Scheme. *Mon. Wea. Rev.* **124**, 1746–1766.

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

Pagé, C., Fillion, L. and Zwack, P. 2007. Diagnosing summertime mesoscale vertical motion: implications for atmospheric data assimilation. *Mon. Wea. Rev.* **135**, 2076–2094.

Park, S. K. and Droegemeier, K. K. 2000. Sensitivity analysis of a 3D convective storm: implications for variational data assimilation and forecast error. *Mon. Wea. Rev.* **128**, 140–159.

Pires, C. A., Talagrand, O. and Bocquet, M 2010. Diagnosis and impacts of non-Gaussianity of innovations in data assimilation. *Phys. D: Nonlinear Phenom.* **239**, 1701–1717.

Sakov, P. and Oke, P. R. 2008. Implications of the form of the ensemble transformation in the ensemble square root filters. *Mon. Wea. Rev.* **136**, 1042–1053.

Stensrud, D. J., Xue, M., Wicker, L. J., Kelleher, K. E., Foster, M. P. and co-authors. 2009. Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc.* **90**, 1487–1499.

Sun, J. 2005. Convective-scale assimilation of radar data: progress and challenges. *Quart. J. Roy. Meteor. Soc.* **131**, 3439–3463.

Talagrand, O., Vautard, R. and Strauss, B. 1997. Evaluation of probabilistic prediction systems. In: *Proceedings, ECMWF Workshop on Predictability*, Shinfield Park, Reading, UK, 1–25.

Tong, M. and Xue, M. 2008. Simultaneous estimation of microphysical parameters and atmospheric state with radar data and ensemble Kalman filter. Part I: Sensitivity analysis and parameter identifiability. *Mon. Wea. Rev.* **136**, 1630–1648.

Wang, X. and Bishop, C. H. 2003. A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci* **60**, 1140–1158.

Wang, X., Bishop, C. H. and Julier, S. J. 2004. Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble?. *Mon. Wea. Rev.* **132**, 1590–1605.

Wilks, D. S. 2006. *Statistical Methods in the Atmospheric Sciences*, 2nd edition, Volume 59. International Geophysics Series. Academic Press.

Wilson, D. R. and Ballard, S. P. 1999. A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Quart. J. Roy. Meteor. Soc.* **131**, 1607–1636.