RMetS
Royal Meteorological Society

Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts

K. Kober,[a]* G. C. Craig,[a,b] C. Keil[b] and A. Dörnbrack[a]

[a]*Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany*
[b]*Meteorologisches Institut, Ludwig-Maximilians-Universität, München, Germany*

*Correspondence to: K. Kober, Meteorologisches Institut, Ludwig-Maximilians-Universität, München, Germany.
E-mail: kirstin.kober@lmu.de

A seamless prediction of convective precipitation for a continuous range of lead times from 0–8 h requires the application of different approaches. Here, a nowcasting method and a high-resolution numerical weather prediction ensemble are combined to provide probabilistic precipitation forecasts. For the nowcast, an existing deterministic extrapolation technique was modified by the local Lagrangian method to calculate the probability of exceeding a threshold value in radar reflectivity. Numerical forecasts were obtained from an experimental high-resolution ensemble that provides 20 different deterministic forecasts of synthetic radar reflectivity. Probabilistic information was calculated by different approaches from the ensemble output. The probabilistic forecasts based on the ensemble were calibrated with the reliability diagram statistics method. The skill of the probabilistic nowcasts and forecasts was evaluated using three quality measures. Finally, a seamless probabilistic forecast was generated as an additive combination of nowcast and forecast, using a weighting function based on their relative skills. The skill of the seamless forecast was greater than or equal to that of the nowcast or ensemble forecast in all quality measures and at all lead times. **Copyright © 2011 Royal Meteorological Society**

Key Words: short-range forecasting; blending; ensemble prediction; forecast calibration

Received 7 May 2010; Revised 25 August 2011; Accepted 2 September 2011; Published online in Wiley Online Library 4 October 2011

Citation: Kober K, Craig GC, Keil C, Dörnbrack A. 2012. Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Q. J. R. Meteorol. Soc.* **138**: 755–768. DOI:10.1002/qj.939

1. Introduction

An accurate forecast of the future atmospheric state at different forecast lead times is of great societal and economic significance. Convective precipitation forecasts affect daily life in various sectors including aviation, construction and leisure, but their utility may be limited by uncertainty. The quantification of forecast uncertainty in a probabilistic forecast enables more precise decision-making, taking into account each user's needs.

To provide reliable methods to perform more accurate short-term forecasts of convective precipitation is an ongoing challenge in atmospheric research (Fritsch and Carbone, 2004). The most commonly used forecast methods are nowcasting and numerical weather prediction (NWP). Both show different forecast skills depending on the forecast lead time.

Nowcasts are short-term forecasts initialized with observed patterns in remote-sensing data, for example areas of high radar reflectivity. These patterns represent convective elements with their own characteristic lifetimes. Usually, the forecasts are spatio-temporal extrapolations for a lead time of up to around 2 h. For very short lead times compared with the mean lifetime of an observed pattern, linear extrapolation shows very high forecast skill.

However, since only advective transport is considered in most nowcasting methods, the continuous temporal evolution of the precipitation field cannot be taken into account. Attempts to include life-cycle effects have shown ambiguous results, and forecast errors typically increase quite rapidly with forecast lead time (Pierce *et al.*, 2004; Wilson *et al.*, 2004).

On the other hand, forecasts based on NWP models simulate the temporal evolution of the precipitation field. However, even with advanced data assimilation techniques the initial humidity fields deviate from the true state. Furthermore, the parametrized model physics limits the predictive skill of the precipitation forecasts. Most importantly, convective elements may develop during the model integration from initially small-scale cells to larger patterns. Their evolution and the turbulent character of the flow limit predictability in the first few hours of the integration. Nevertheless, the ability of the numerical forecast to represent the evolution of the larger-scale environment of convection allows NWP model forecast skill to outperform nowcasting methods after some lead time (about 6 h in the study of Lin *et al.*, 2005).

The intrinsic uncertainty in both methods, as well as the stochastic nature of convection, requires a probabilistic approach to prediction. If the probabilities accurately represent the uncertainty of the nowcast and numerical forecast, they can be blended seamlessly to produce skilful predictions across a wide range of lead times.

Traditional nowcasting methods provide deterministic forecasts of objects defined by the respective observation method (Wilson *et al.*, 1998). The objects are identified either in radar, satellite or lightning data by applying one or a combination of several thresholds. Most nowcasting methods are radar-based and rely on the assumption that the evolution of the detected precipitation field is primarily governed by advection, e.g. Dixon and Wiener (1993), Li *et al.* (1995), Golding (1998), and Kober and Tafferner (2009).

In contrast to deterministic forecasts, probabilistic approaches predict the probability of exceeding a threshold in the observed field. The most straightforward method is to calculate a probability of precipitation based on the fraction of precipitation pixels in a region around a point of interest (Andersson and Ivarsson, 1991; Schmid *et al.*, 2000; Germann and Zawadzki, 2004). Germann and Zawadzki (2004) introduced and compared four methods of providing probabilistic forecasts based on continental radar observations. They concluded that the most skilful method was the local Lagrangian method, which has since been adapted by others, e.g. Megenhardt *et al.* (2004). In addition, uncertainty in nowcasts resulting from errors in the observations can be quantified by creating ensembles of precipitation fields (Germann *et al.*, 2009). In this method, stochastic ensemble members are the sum of the observed deterministic radar precipitation field and stochastic perturbations derived on basis of the radar error covariance matrix. Radar ensembles are of great interest for hydrological applications (Rossa *et al.*, 2010). These approaches do not incorporate precipitation forecasts from NWP models.

A significant change in numerical prediction of cumulus convection has occurred with the introduction of models with kilometre resolution, which operate without the use of cumulus parametrization. Initial experience suggests that such models offer improved forecast skill in comparison with coarser resolution models (Lean *et al.*, 2008; Dixon *et al.*, 2009; Weusthoff *et al.*, 2010). However, the skill of the deterministic forecasts depends on many aspects of the model configuration, including resolution and parametrizations (Done *et al.*, 2004; Gebhardt *et al.*, 2011). Furthermore, the representation of the initial fields and their discrepancies from observations is influential, and data assimilation methods have been found to have strong impacts on the behaviour of high-resolution numerical forecasts (Sokol and Rezacova, 2006; Stephan *et al.*, 2008; Dixon *et al.*, 2009).

In order to quantify the uncertainty in NWP model predictions, ensemble methods have been developed at weather prediction centres and matured to a well-established approach (see e.g. the review article of Lewis, 2005). Several approaches exist to design ensembles. Perturbations of the initial or boundary conditions or perturbations of the model physics (Stensrud *et al.*, 2000) in a linear or stochastic way (Bright and Mullen, 2002) can be applied to create different forecasts. Different forecast models (multimodel ensemble), runs of the same forecast model starting at different times (time-lagged ensemble, e.g. Mittermaier, 2007) and combinations thereof can also be utilized (Roebber *et al.*, 2004).

Although ensemble prediction systems (EPS) have matured to a standard technique for large and mesoscales, only a few high-resolution, i.e. convection-permitting, ensembles exist (e.g. Gebhardt *et al.*, 2011). The design of convection-permitting ensembles differs from that of mesoscale ensembles with parametrized convection, because of the different mechanisms of error growth at smaller scales (Hohenegger and Schär, 2007). The experimental EPS of the Deutscher Wetterdienst (DWD) is based on the Consortium of Small-scale Modeling (COSMO) deterministic forecast model with 2.8 km horizontal resolution. In the version of COSMO-DE-EPS used in this study, boundary conditions and physical parametrizations are varied with the aim of maximizing the spread in precipitation forecasts at short lead times (Gebhardt *et al.*, 2011).

The skilful combination of nowcasting methods and NWP models to forecast precipitation has the potential to maintain the overall predictive skill for a continuous range of lead times from 0 to more than 8 h. Most of the published methods for combining nowcasts and forecasts use a weighted sum of the two fields. The weighting functions are determined by the skill of the predictions derived from suitable quality measures. Several studies have identified the forecast skill of nowcasting and NWP models using deterministic (Golding, 1998; Kilambi and Zawadzki, 2005) or probabilistic (Bowler *et al.*, 2006) quality measures. The evaluated quantity was either radar reflectivity (Wilson and Xu, 2006), rainfall rate (Golding, 1998) or probability of precipitation (Pinto *et al.*, 2006). The combination has been performed by applying linear (Wong *et al.*, 2009) or exponential (Golding, 2000) weights. Additionally, a scale-dependent stochastic approach to calculate a probabilistic precipitation forecast was applied by Bowler *et al.* (2006). Note that most of these studies used coarse-resolution NWP models (larger 10 km) where convection is parametrized.

The aim of this article is to develop a method for combining a probabilistic nowcast with a probabilistic numerical forecast, in a way that preserves the skill of the individual

methods. A key element of the work presented here is the use of an ensemble of forecasts using a so-called cloud-resolving or convection-permitting model with a resolution of a few km. In addition to providing a better representation of the physics of convection, it is anticipated that high-resolution models may be more effectively combined with radar data, since the resolution is comparable. The use of an ensemble of forecasts allows various sources of forecast uncertainty to be taken into account. In combining the two data sources, care is taken to prepare the nowcast and numerical forecast output in a similar way. Each is presented as a forecast of the probability of reflectivity (observed or simulated) exceeding a specified threshold at each point on a high-resolution grid. The probabilities are then combined using a time-varying weighting function, based on the measured performance of the nowcast and numerical ensemble forecast. The result is a probabilistic forecast that transitions smoothly from one data source to the other and reflects the increasing uncertainty in the prediction with increasing lead time. By combining probabilities, we avoid inconsistencies associated with differences in how the probability distributions are represented in the two forecasting systems. If the nowcasting system used an ensemble to represent uncertainty, as is the case for the numerical modelling system, it would also be possible to construct a probabilistic forecast from the combined ensemble of nowcasts and forecasts, as was done by Bowler *et al.* (2006).

In this work, probabilistic nowcasts are created by extending the deterministic radar tracker Radar TRAcking and Monitoring (Rad-TRAM) and combining these with probabilistic forecasts based on the output of COSMO-DE-EPS. Section 2 describes the data and the methods used to derive probabilistic forecasts from Rad-TRAM and from COSMO-DE-EPS. The quality of the forecasts is evaluated and compared with different probabilistic quality measures in section 3. In section 4, the concept for the combination of the two forecasting methods is introduced and results of the blending procedure are presented. In section 5, the findings of this study are interpreted and discussed. Finally, in section 6 short conclusions are drawn.

2. Data and methods

This study will consider predictions of the probability of radar reflectivity exceeding a specified threshold. It is important that the same variable is used in both the nowcast and the numerical forecast output that will be blended. Reflectivity is a convenient variable, since no conversion of the observations is required. For the model forecasts, a forward operator is necessary to convert the variables to reflectivity (e.g., Seifert and Beheng, 2006). In other applications, the reflectivity threshold could be converted to a precipitation threshold using a Z–R relationship, or to a hazard threshold, for example in aviation.

2.1. *Observations: European radar composite*

The European radar composite is the data basis for the nowcasting method Rad-TRAM and for the quality evaluation of both forecasts. It is provided by the Deutscher Wetterdienst (DWD) and encompasses an area of 1800 km × 1800 km over Europe. In this study, a smaller subdomain (about 650 km × 650 km) covering a large part of Germany is selected for quality evaluation (Figure 1). The European

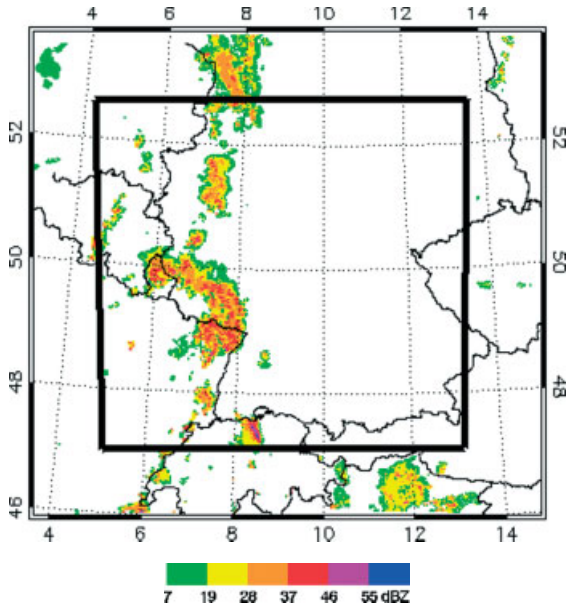

Figure 1. Observed radar reflectivity on 12 August 2007, 2315 UTC during the passage of a cold front over western Germany that caused pre-frontal convection. The domain for forecast quality evaluation is shown in black. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

radar composite provides radar reflectivities given in six dBZ classes [7, 19, 28, 37, 46, 55 dBZ] on a horizontal grid with a regular resolution of 2 km × 2 km. More details about the creation of the composite, the underlying radar measurements and the inherent errors are given by Kober and Tafferner (2009).

2.2. *Radar tracker Rad-TRAM and the new probabilistic module*

The deterministic tracking and nowcasting algorithm Rad-TRAM (Radar TRAcking and Monitoring) has been developed recently (Kober and Tafferner, 2009) and uses the European radar composite. Rad-TRAM consists of 4 parts: (1) the extraction of the motion field by solving the optical flow equation, (2) the detection of convective cells, (3) the tracking of cells and (4) the nowcasting of these cells for one hour. The motion field derived in part (1) is obtained by an optical flow technique (Keil and Craig, 2007; Zinner *et al.*, 2008) of the box- or region-based matching type (Barron *et al.*, 1994), and is based on the pyramidal method of Anandan (1989). In its original version, Rad-TRAM identifies severe convective cells through a threshold criterion of 37 dBZ. In this study, Rad-TRAM is upgraded to calculate probabilistic precipitation forecasts of reaching a lower threshold of 19 dBZ, indicating areas of rainfall. In general, a range of thresholds will be of interest for different applications, but since the purpose of this article is to demonstrate the method, a single threshold will be used. The lower value is chosen since events are more frequent than for higher thresholds and thus statistical verification is easier.

To create a probabilistic nowcast, a method similar to the *local Lagrangian* approach (Germann and Zawadzki, 2004) is developed and implemented. Germann and Zawadzki (2004) identify two main error sources in extrapolation forecasts: incorrect displacements and thermodynamic processes other than advection. The temporal evolution of the precipitation field (onset, growth or decay) cannot

be represented by extrapolation methods. Germann and Zawadzki (2004) relate the overall errors to the spatial variability of the precipitation field itself, without directly quantifying specific error sources. Following this reasoning, in our approach the probabilistic precipitation forecast is based on first estimating the fraction of precipitation pixels in a predefined area around each radar grid point with values larger than 0 dBZ. This *local* approach is followed by considering the movement of the precipitation field (*Lagrangian* approach).

In Rad-TRAM, the calculation of probabilistic forecasts is implemented as an optional module. To determine the displacement of the precipitation probability field, the module uses the scale-dependent displacement vector field derived in the first part of Rad-TRAM. The identified convective cells (part 2–4) are not considered in the probabilistic module.

The probability P_{LL} of exceeding a threshold $\mathcal{L}$ is defined as

$$P_{\mathrm{LL}}(t_0+\tau, x, \mathcal{L}, k) = \mathrm{Prob}\{\psi(t_0, x-\alpha+r) \geq \mathcal{L} | (x+r) \in \omega_k\}, \quad (1)$$

where ψ is the observed field of radar reflectivity, $\mathcal{L}$ the threshold reflectivity (19 dBZ), ω_k the search area centred on the point of interest x, chosen to be a square of side length k. The scale parameter k (side length of search area) depends on the forecast lead time τ. The probability value is extrapolated using the displacement vector α defined at the point of interest x. It was not considered necessary to implement a more complex algorithm that would allow curved trajectories, since it is expected that extrapolation errors will be smaller than errors associated with changes in the structure and amplitude of the precipitation field (Germann and Zawadzki, 2004).

Probabilities are computed and extrapolated at every grid point with reflectivities larger than 0 dBZ in the evaluation domain. Other grid points are omitted in order to save computational costs. This has the effect that some low probabilities in the area around a precipitation feature are missed. Sensitivity tests (not shown) show that the quality measures discussed in the next section are only slightly affected by omitted regions of low probability. Finally, a smoothing based on Delaunay triangulation (Sugihara and Inagaki, 1995) is applied to the probability field P_{LL} to eliminate possible gaps resulting from divergent displacement vectors. Since the spatial structure of the probability field is relatively smooth to start with, this procedure has little effect except to fill in the gaps. If instead of probability, the reflectivity field itself was extrapolated, the presence of gaps or the application of smoothing to remove them would have a substantial effect on the resulting probability field. Forecasts are provided up to 8 h lead time in 15 min time steps.

The size of the search area ω_k is chosen to depend on the forecast lead time τ and increases with lead time in the first 4 forecast hours as the uncertainty of the temporal evolution increases. Following Germann and Zawadzki (2004), the side length of the search area is assumed to grow linearly at a rate of 1 km per minute. From forecast hours 4–8, the size of the search area is kept constant with a maximum side length of 240 km. This value should represent the distance over which convective cells share the same synoptic environment, and is expected to be related to the Rossby radius of deformation, which is the length over which significant temperature gradients can be maintained by geostrophic balance. Over larger areas the environment varies and the frequency of occurrence across the entire area is no longer representative of the probability at the point of interest.

Typical forecasts derived with this probability technique are illustrated in Figure 2 for 12 August 2007, 2315 UTC for different lead times. Additionally, the reflectivity fields that are the basis for the respective forecasts are displayed in the background. The 15 min forecast provided at 2300 UTC is very sharp and reflects the low uncertainty for short lead times ($\tau = 15$ min, Figure 2(a)). The forecast calculated on basis of the reflectivity observations one hour earlier ($\tau = 60$ min, Figure 2(b)) already shows increased uncertainty, having a smoother P_{LL} field with lower probability maxima. The small-scale structure of the observed field cannot be represented at this lead time. The comparison with the corresponding observation field (cf. Figure 1) reveals that the forecast still has skill concerning the position of the probability field. The forecast based on observations two hours earlier ($\tau = 120$ min, Figure 2(c)) shows a further smoothed probability field. The position of the field in comparison with the observations is still meaningful. However, as the probability field covers a larger area there are some false alarms.

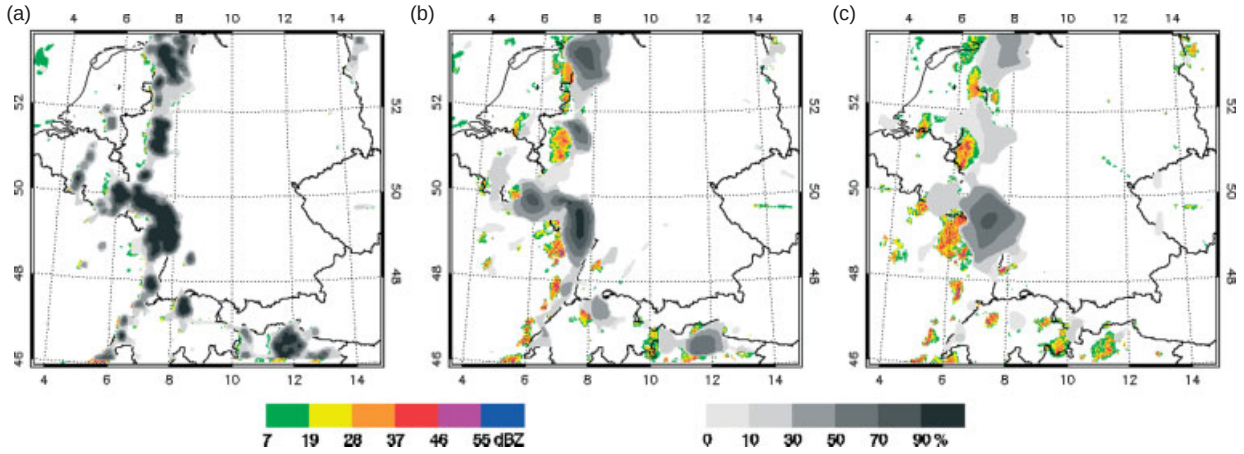

Figure 2. Probabilistic forecasts P_{LL} of Rad-TRAM for 12 August 2007, 2315 UTC with (a) 15 min forecast from 2300 UTC, (b) 60 min forecast from 2215 UTC and (c) 120 min forecast from 2115 UTC, grey-shaded, and the reflectivity observations at the respective initial time colour-coded in the background. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

Table I. List of parameter perturbations in COSMO-DE-EPS.

Parameter	Description	Perturbed	Default
entr_scv	Entrainment rate of shallow convection	0.002	0.0003
clc_diag	Subscale cloud cover given grid-scale saturation in the turbulence scheme	0.5	0.75
rlam_heat	Scaling factor of the laminar sublayers for scalars	50.0	1.0
rlam_heat	Scaling factor of the laminar sublayers for scalars	0.1	1.0
tur_len	Asymptotic mixing length of turbulence scheme	150.0	500.0

2.3. COSMO-DE-EPS and the derivation of probabilistic forecasts

COSMO-DE-EPS is currently under development at DWD based on the COSMO-DE model (Gebhardt *et al.*, 2011). COSMO-DE (previously known as LM-K: Baldauf *et al.*, 2011) is a non-hydrostatic and convection-permitting weather forecasting model. It has been developed for short-range forecasts in the framework of the Consortium of Small-scale Modeling (COSMO). The horizontal resolution is 2.8 km and 50 vertical levels are used up to 30 hPa. Precipitation processes are explicitly parametrized using a bulk cloud microphysical scheme with five prognostic hydrometeor types (rain, snow, cloud water, cloud ice and graupel). Deep convection is explicitly resolved.

COSMO-DE-EPS consists of 20 members. The different members are created by addressing two sources of uncertainty. Firstly, uncertainties in model physics are considered by changing five different parameters of the physics scheme in a non-stochastic approach (Table I). These parameters are chosen in order to maximize the variability of convective precipitation in the physical parametrizations (Gebhardt *et al.*, 2011). Secondly, uncertainties due to the lateral boundary conditions are considered by nesting COSMO-DE into four members of COSMO Short-Range Ensemble Prediction System (COSMO-SREPS, resolution 10 km). The four members are driven by different global models (Marsigli *et al.*, 2008).

From COSMO-DE-EPS, the fields of synthetic radar reflectivity at the 850 hPa pressure surface are used to calculate probabilistic forecasts $P_{\mathrm{EPS}}(x, \mathcal{L})$ of exceeding the threshold $\mathcal{L} = 19$ dBZ. Synthetic reflectivities are calculated with a forward operator using information from the distribution of the hydrometeors rain, snow and graupel at every grid point (Seifert and Beheng, 2006).

Probabilistic forecasts are derived from the ensemble by means of three different approaches (cf. Schwartz *et al.*, 2010). Firstly, as traditionally applied to ensembles, the fraction of members with values above the threshold (here $\mathcal{L} = 19$ dBZ) is determined at every grid point. These probabilities depend on the number of ensemble members. In the following, this method will be called the *fraction method*. Secondly, every member is treated as a deterministic solution and the fraction of precipitation pixels ($\mathcal{L} \geq 19$ dBZ) in a predefined area (neighbourhood) around each precipitating grid point is computed for each member separately. This results in 20 different probabilistic forecasts and is called the *neighbourhood method*. As a third approach, the mean of these 20 different probability fields derived with the neighbourhood method is calculated. In Schwartz *et al.* (2010) this approach is referred to as neighbourhood ensemble probability, here as the *mean method*.

A critical parameter of the neighbourhood method is the size of the search area. Theis *et al.* (2005) and Schwartz *et al.* (2010) varied this parameter systematically, but could not identify an optimal size or shape of the neighbourhood. Here, in contrast to the local Lagrangian method for observation-based forecasts, the size of the neighbourhood is fixed for all lead times as a square of side length 75 km. Sensitivity tests were carried out but revealed no further improvement of the skill scores for larger neighbourhood sizes (not shown). Smaller search areas have been investigated as well and result in sharper probabilities, but lower skill scores.

Altogether, 22 different probabilistic forecasts are available at each forecast time. Both the generation of COSMO-DE-EPS and our analysis consider three sources of uncertainty: the spatial variability around each grid point and, implicitly, timing errors, the imperfectness of model physics and the variability of the lateral boundary conditions. The method providing the mean of the neighbourhood probabilities considers all of them.

Figure 3 illustrates examples of the three approaches applied to COSMO-DE-EPS forecasts for 12 August 2007, 2315 UTC. For the neighbourhood method, only member 1 has been chosen as a typical representative of the ensemble. All forecasts predict a probability of precipitation greater than zero in the area where the front was observed at 2315 UTC (Figure 1). The location and intensity of the probability fields differs between the methods. The fraction method predicts a large and broad probability field. Embedded in this spatially coherent field are scattered probability maxima. In contrast, the probability field of member 1 covers small areas with isolated probability maxima. Note that the gradients are sharp at some locations, relative to the size of the search area, since only the precipitating pixels are used in the computation. In comparison with the fraction method and member 1, the mean of the 20 neighbourhood probabilities is a smooth field with low probability values. The variability in size and spatial distribution of the probability fields in the different forecasts covers a reasonable range given the meteorological situation. The large areas with probabilities larger than zero given by the fraction and mean methods reflect the high variability among the ensemble members. The probabilistic forecasts of the mean and fraction methods are very similar in location and size, since both consider the variability of the entire ensemble. In contrast the mean field is smoother, with lower probability values, since the variability of the reflectivity field around each grid point is additionally considered.

As the ensemble is experimental and still under development, only one forecast period with the same configurations in physical perturbations was available for this study: 9 days from 8–16 August 2007. The model runs started once each day at 0000 UTC and forecast 24 h. Instantaneous synthetic radar reflectivities are available every 30 min starting at 0015 UTC. With a time period of 9 days, the size of the domain and the high temporal and

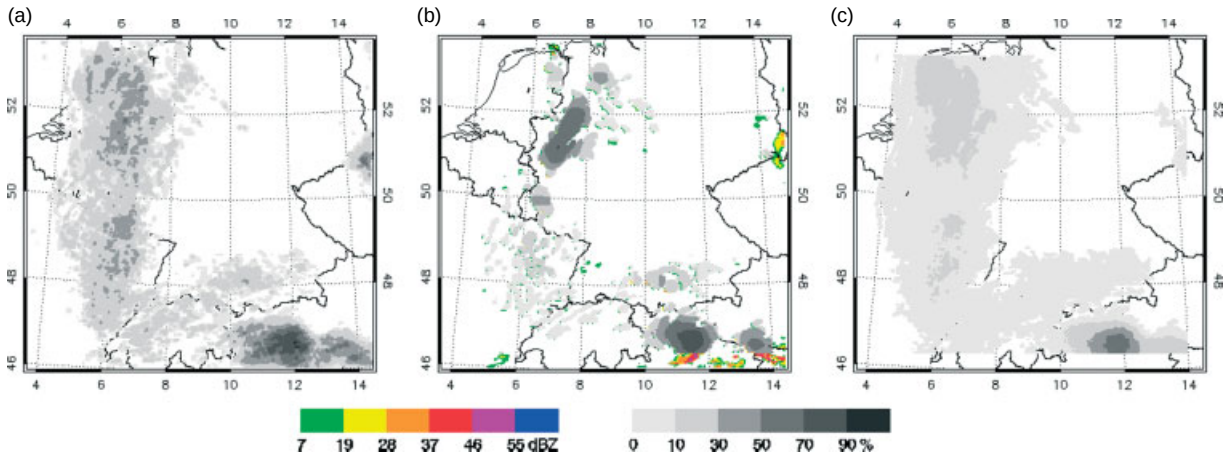

Figure 3. Probabilistic COSMO-DE-EPS forecasts for 12 August 2007, 2315 UTC for (a) the fraction, (b) member 1 as representative for the 20 neighbourhood forecasts and (c) the mean of the ensemble (grey-shaded). In the background of member 1 the synthetic radar reflectivities are colour-coded. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

spatial resolution imply that a large set of data is available, containing around 21 million data points of synthetic radar reflectivity for every member.

The probabilities based on COSMO-DE-EPS forecasts are calibrated with the reliability diagram statistics method (Zhu *et al.*, 1996). For the calibration, the available data are divided into training and testing data sets. The reliability diagram statistics method suggests that if the probability category i (where i indexes 11 equal-sized categories between 0.0 and 1.0) is predicted in the testing subsample, the calibrated probability is the frequency with which the event is observed in the training subsample under the condition that the sample forecast category i is predicted. The training period comprises the period from 8–11 August (around 8 million data points per member) and the testing set the period from 12–16 August (around 10 million data points). The first three hours of each run are not included in the calibration, as the spread among the ensemble members is small. This is done for all three methods of deriving probabilistic forecasts from COSMO-DE-EPS. For the calibration of the neighbourhood probabilities, all members are calibrated together (around 365 million data points in total). The fraction method and mean method probabilities cover 1/20 of the data points. Consistent results for the calibration functions are obtained if the testing and training periods are interchanged. Tests with other definitions of the testing and training period (not shown) revealed similar results, showing that to a first approximation there is no dependence on the choice of periods.

The reliability component of the decomposed Brier score (Brier, 1950; Murphy, 1973) can be used as a measure for a successful calibration (Atger, 2003). In this study, only the domain reliability is calculated due to the limited period of the forecasts. Table II shows the mean and the standard deviation of the reliability component for all three methods separately, over the entire period, with and without calibration. The values reveal that the calibration is successful since both the mean and the standard deviation of the reliability component of the Brier score are reduced for the neighbourhood probabilities, the fraction method and the mean method by at least a factor of 2. In the following sections, only the calibrated probabilities are used.

Table II. Reliability component of Brier score, mean and standard deviation. All grid points are considered together. For the neighbourhood method the total of all single members is calculated.

Method	Mean reliability	Standard deviation
Neighbourhood raw	5.8×10^{-1}	6.6×10^{-1}
Neighbourhood calibrated	2.4×10^{-1}	3.2×10^{-1}
Fraction raw	3.2×10^{-2}	3.2×10^{-2}
Fraction calibrated	0.9×10^{-2}	1.1×10^{-2}
Mean raw	1.9×10^{-2}	2.3×10^{-2}
Mean calibrated	0.9×10^{-2}	1.1×10^{-2}

3. Quality of the probabilistic forecasts

The most important aspects of quality for probabilistic forecasts are reliability, resolution and sharpness (Murphy and Winkler, 1987). The basis for a skilful combination of the probabilistic forecasts from Rad-TRAM and COSMO-DE-EPS will be knowledge of their individual forecast quality. In particular, the evolution of quality measures with lead time is important to define weighting functions to blend the two methods. The quality of probabilistic forecasts of discrete predictands is assessed here using standard measures: the Brier score and its decomposition, together with Relative Operating Characteristic (ROC) curves and the area underneath them (Wilks, 2006). Additionally, a simplified version of the conditional square root of ranked probability score (CSRR) is calculated following Germann and Zawadzki (2004). The CSRR is originally defined for multicategory forecasts. Since only a single threshold (19 dBZ) is used in this study, the CSRR simplifies to

$$CSRR(\tau) = \left\{ \frac{1}{\tilde{\Omega}_{t_0+\tau}} \int_{\Omega} [P(t_0+\tau, x) - \dot{P}(t_0+\tau, x)]^2 \, dx \right\}^{0.5}, \quad (2)$$

where $\tilde{\Omega}_{t_0+\tau}$ is the size of the observed rain domain ($\mathcal{L} >$ 0 dBZ), Ω the entire domain, P the probabilistic forecast and

$\hat{P}$ the observation. Due to weighting with the size of the rain domain, the CSRR is independent of the observed frequency of the event. Therefore, the magnitude of the score reflects skill and is comparable even in different meteorological situations. In contrast to the CSRR, the Brier score is sensitive to correct negatives. In the case of rare events, low values of the Brier score give the illusion of very good performance.

In the following, the skill of the probabilistic Rad-TRAM and COSMO-DE-EPS is evaluated individually in time series over a selected time period. Subsequently, the evolution of skill of both forecast methods with lead time over the entire period is compared.

3.1. *Performance of probabilistic Rad-TRAM*

The Brier score, the CSRR and the area under the ROC curve (ROC area) are presented to illustrate Rad-TRAM's forecast skill for the period 1200–2400 UTC, 12 August 2007 (Figure 4). On this day there was no convective activity around noon, but in the afternoon a cold front from the west propagated into the region and forced deep convection ahead of it. The appearance of the front in the evaluation domain changed the properties of the nowcasts, as actual radar observations are included. This marks a regime transition around 1400 UTC. Figure 4 shows the skill of the probabilistic forecasts based on different lead times from $n = 1$ (15 min) to $n = 32$ (8 h). The first forecast within each hour is highlighted in black (15 min, 75 min, 135 min, ...).

The Brier score shows almost perfect skill, with very low values for all lead times from 1200–1400 UTC, as radar reflectivities of at least 19 dBZ almost never occurred. During the course of the day, the skill of the Brier score decreases as the observed frequency of the event increases. The number of distinguishable lead times increases as well. This reflects the fact that in the second part of the day forecasts are only skilful to the extent that they correspond to radar observations of pre-frontal convective precipitation inside the domain. Forecasts based on earlier observations are very similar to each other and not skilful, as no precipitation was observed yet. During the latter part of the day, the Brier score varies over a large range, with very small values for short lead times and larger values for the forecasts based on older observations.

At noon, the CSRR generally shows a similar behaviour to the Brier score, with small values and low variability between the lead times. In the afternoon, the differences between the lead times increase and a larger number of forecasts can be distinguished. In addition, the CSRR shows that the skill of the forecasts increases within the respective lead times during the day (e.g. for the 15 min forecast CSRR is 0.6 at 1400 UTC and 0.4 at 2200 UTC).

The values of the area under the ROC curve vary over the entire range of possible values (0.5–1.0) with very high skill in the first forecast hour to very low skill for longer lead times. Again, the evolution of the ROC area over the period reflects the meteorological regime with higher skill in the advection-dominated frontal passage (e.g. for the 15 min forecast the ROC area is 0.92 at 1400 UTC and 0.98 at 2200 UTC).

In general, the three skill scores provide a consistent judgement of the forecast quality. If the quality of forecasts based on the different lead times is distinguishable, the scores are ranked according to their lead time. Short lead times (based on the latest observations) have significantly higher skill than longer lead times. Hence, negatively oriented scores (Brier score and CSRR) increase with lead time and the positively oriented ROC area decreases. Differences between the lead times become smaller with increasing lead time (cf. the differences between the black lines in Figure 4). For longer lead times the ranking is not clearly identifiable. The ROC area extends the number of distinguishable forecast hours by about one hour. After the regime change, the CSRR and the ROC area reveal an increase in skill of the forecasts within the respective lead times.

3.2. *Performance of COSMO-DE-EPS*

The calibrated probabilistic forecasts derived from COSMO-DE-EPS output are evaluated over the same period as Rad-TRAM on 12 August 2007, 1200–2400 UTC (Figure 5). Here, the different lines denote the different methods (fraction method, 20 members based on neighbourhood method and mean of neighbourhood members) that were applied to the COSMO-DE-EPS output.

Although Figure 5 depicts just a 12 h section of the entire 9 day period, the results reveal that the temporal variability in the Brier score and the CSRR is larger than the variability between the 22 different methods (spread). Only the area under the ROC curve shows a variability within the different methods that is significant in comparison to the temporal variations. Generally, the values of the scores in this 12 h period indicate low skill. For example, the area under the ROC curve exceeds 0.7 only for three forecast hours and only in two of the 22 solutions. This threshold value is sometimes considered as an indicator for useful forecasts (Buizza *et al.*, 1999).

In the frontal regime, all three scores agree that the fraction method and member 1 have more skill than the other solutions (Figure 5). Their skill remains higher for several forecast hours in this regime. The skill of the mean method in all scores is within the range of the skill of the neighbourhood members. The highest values in ROC area of the neighbourhood members during 1800 UTC and 2100 UTC are found for those members where the entrainment rate of shallow convection is perturbed (Table I). In other meteorological regimes during the 9 day period, the members are grouped with respect to the boundary conditions of the global models (not shown).

3.3. *Dependence of forecast quality on lead time*

A second possibility to evaluate the performance of the probabilistic forecasts from Rad-TRAM and calibrated COSMO-DE-EPS is the evolution of skill with lead time. Rad-TRAM forecasts are evaluated every two hours beginning at 0200 UTC (0200, 0400, 0600, 0800, 1000, 1200, 1400, 1600 UTC) for 8 h lead time for every day. Due to the smaller temporal resolution of COSMO-DE-EPS output (30 min: *hh*15 and *hh*45), Rad-TRAM is analysed at the same times. This means that probabilistic forecasts are analysed for lead times of 15, 45, 75 min and so on up to 465 min. As the model is started once a day, an actual lead time-dependent evaluation of COSMO-DE-EPS is not possible.

The evaluation of the time series of all days (not shown for the entire period but exemplified in Figure 5) revealed that to a first approximation the skill of the forecasts does not depend on lead time within a 24 h period. Within the 8 h

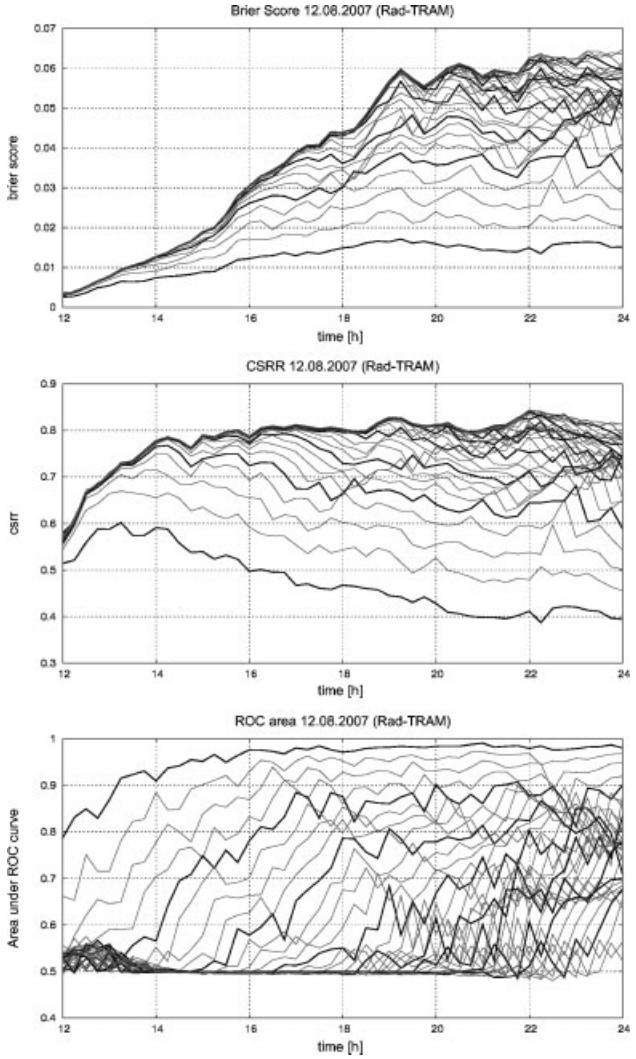

Figure 4. Evolution of Brier score, CSRR and area under ROC curve for Rad-TRAM forecasts on 12 August 2007 from 1200–2400 UTC. Black lines denote the first forecast for each of the eight forecast hours (*hh*15), grey lines the three other forecasts within each forecast hour.

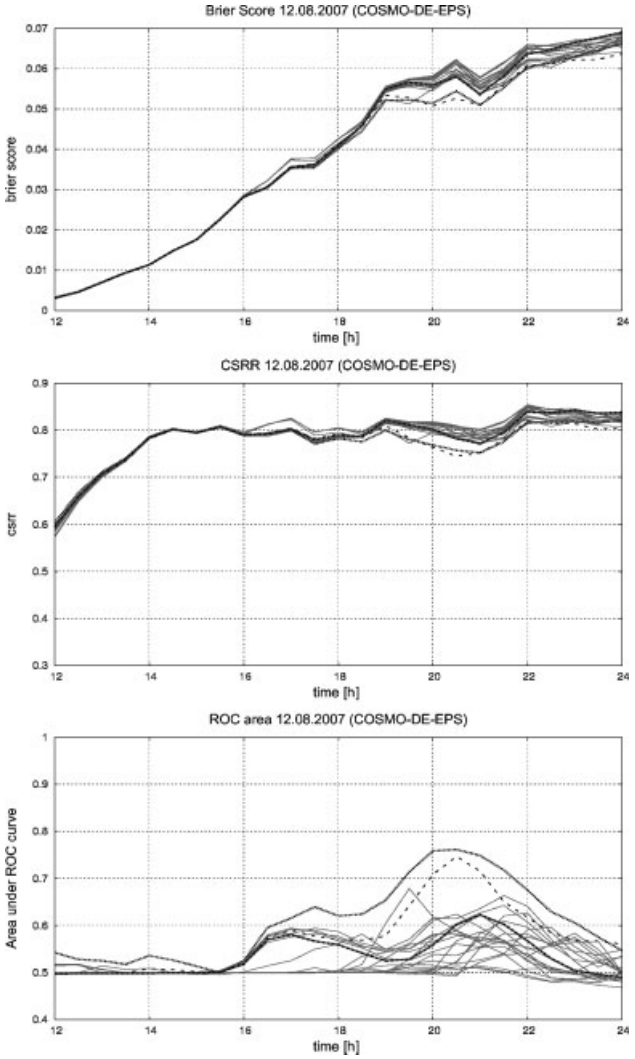

Figure 5. Evolution of Brier score, CSRR and area under ROC curve for the calibrated COSMO-DE-EPS forecasts on 12 August 2007 from 1200–2400 UTC. Dotted line: fraction method; dashed: member 1; solid grey: all other neighbourhood members; dot–dashed: mean of neighbourhood probabilities.

time frames defined above for Rad-TRAM (0200–1000 UTC, 0400–1200 UTC, . . . , 1600–2400 UTC) for each of the nine days, the mean model skill is calculated as a temporal average. Finally, the mean and the standard deviation over the entire period (8–16 August 2007) are derived for Rad-TRAM as well as COSMO-DE-EPS forecasts (Figure 6).

Rad-TRAM's mean skill (thick black solid line) decreases with lead time for each score. In the first three hours the decrease is faster than it is later (e.g. CSRR decreases after 3 h to 66% of the initial skill and after 8 h to 58%). The standard deviation (thin black solid lines) as a measure for the variability of the mean value is very large for the Brier score at all lead times (80%). It is larger than the variability or decrease of the mean values with lead time. The CSRR and the area under the ROC curve have smaller standard deviations (around 10 and 20% respectively).

Obviously, the mean values of the COSMO-DE-EPS forecasts have smaller skill than the Rad-TRAM forecasts for short lead times (Figure 6). As there is only one ensemble run each day, the mean values cannot depend on lead time and therefore they are constant. The spread and the ranking among the methods applied to the COSMO-DE-EPS output varies in the different scores. Differences between the skill of the methods are very small in the Brier Score. The CSRR shows that the neighbourhood members have more skill (lower values) than the mean method and the fraction method. The area under the ROC curve shows the fraction method slightly better than the mean method and the others (Figure 6, right). As already seen in the time series for the case study (Figure 5), the area under the ROC curve is the only score that shows significant spread between the different neighbourhood members. Their scores are grouped according to the driving global models.

Each score shows a crossover point that identifies the lead time after which a model method has more skill than Rad-TRAM. Here, as several methods are applied to the ensemble, a time frame is identified corresponding to the interval in which the score of Rad-TRAM's mean is worse that the best ensemble method, but better than the worst ensemble method. This time frame ranges between forecast hours $\tau = 5, \ldots, 6.75$ h (Brier score), $\tau = 5.25, \ldots, 6.75$ h (CSRR) and $\tau = 4, \ldots, 6.75$ h (ROC area). Importantly, all three scores show similar crossover time frames. However, the large standard deviations of Rad-TRAM and the

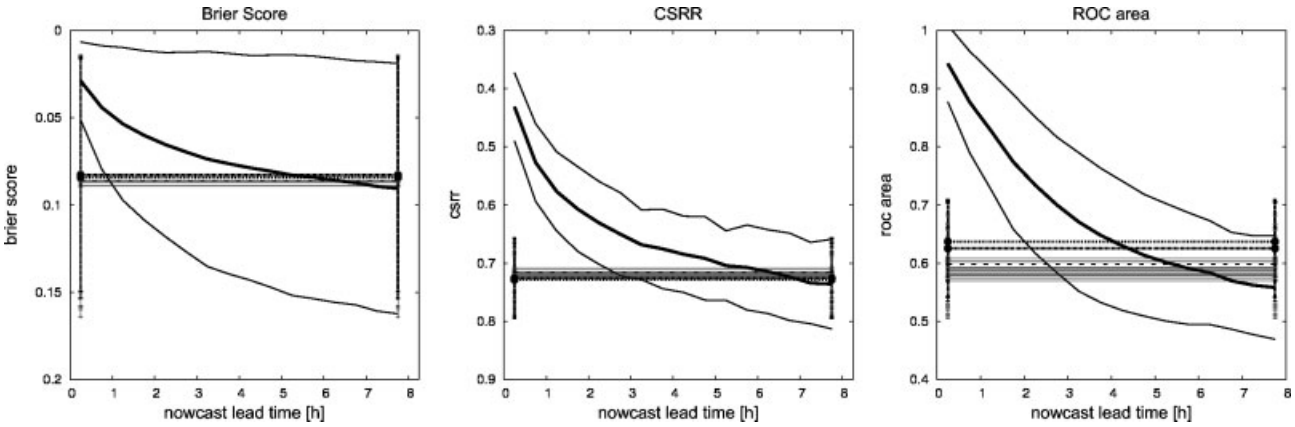

Figure 6. Evolution of Brier score, CSRR and area under ROC curve with lead time for Rad-TRAM and calibrated COSMO-DE-EPS forecasts from 8–16 August 2007. Solid black thick: Rad-TRAM mean; solid black thin: Rad-TRAM standard deviation; dotted line: fraction method; dashed: member 1; solid grey: all other neighbourhood members; dot–dashed: mean of neighbourhood probabilities. Error bars on the COSMO-DE-EPS based forecasts indicate standard deviations.

model methods demonstrate that there is still considerable variability in crossover points, especially for the Brier score. To investigate the influence of the calibration over the entire period, the mean and standard deviation of COSMO-DE-EPS in Brier score, CSRR and area under the ROC curve are displayed in Figure 7 for uncalibrated and calibrated probabilities. The change of mean values from left to right in Figure 7 shows the effect of calibration on the respective method. The Brier score and the CSRR show a reduction of spread and an increase of skill: the range of values in the Brier score is reduced from [0.091, 0.111] to [0.0828, 0.0892] and in CSRR from [0.746, 0.835] to [0.709, 0.729]. The various methods are affected to different degrees by calibration (e.g. CSRR of the fraction method shows a reduction of mean value of 13%, compared with 5% for the mean method). This results in a change in the order of the methods in CSRR: the mean method has the lowest values before calibration and the largest afterwards. For all scores, the fraction method is mostly affected by calibration. In the ROC area, the fraction method is the only method that is changed. The standard deviations decrease clearly for CSRR (fraction method: 17–9% of the mean value) but retain the same magnitude in comparison with the mean values for the Brier score (fraction method: 80%). For the area under the ROC curve, no effect on standard deviation (10% before and after calibration) can be identified.

4. Blending of Rad-TRAM and COSMO-DE-EPS probabilities

4.1. *Method*

The basis for the combination of the probabilistic forecasts provided by Rad-TRAM and the calibrated probabilities derived from COSMO-DE-EPS output is knowledge of the behaviour of their forecast quality with lead time. In section 3.3, this behaviour was evaluated with the Brier score, the CSRR and the area under the ROC curve (Figure 6). The skill of Rad-TRAM forecasts as evaluated with the CSRR is chosen to be the basis for the derivation of the weighting functions for the combination. The weighting function for Rad-TRAM, w_r, is defined in an analogous manner to Kilambi and Zawadzki (2005), with a dependence on lead time τ given by

$$w_r(\tau) = 2.11 - \frac{1}{1 - CSRR(\tau)^{2.8}}, \tag{3}$$

and normalized to unity at the first available lead time ($\tau = 15$ min). The exponent 2.8 is chosen such that the weight crosses 0.5 in the time interval between 5 and 6 h. As the weights of both forecast methods should sum to unity, the weight for all COSMO-DE-EPS based forecasts, w_c, is

$$w_c(\tau) = 1 - w_r(\tau). \tag{4}$$

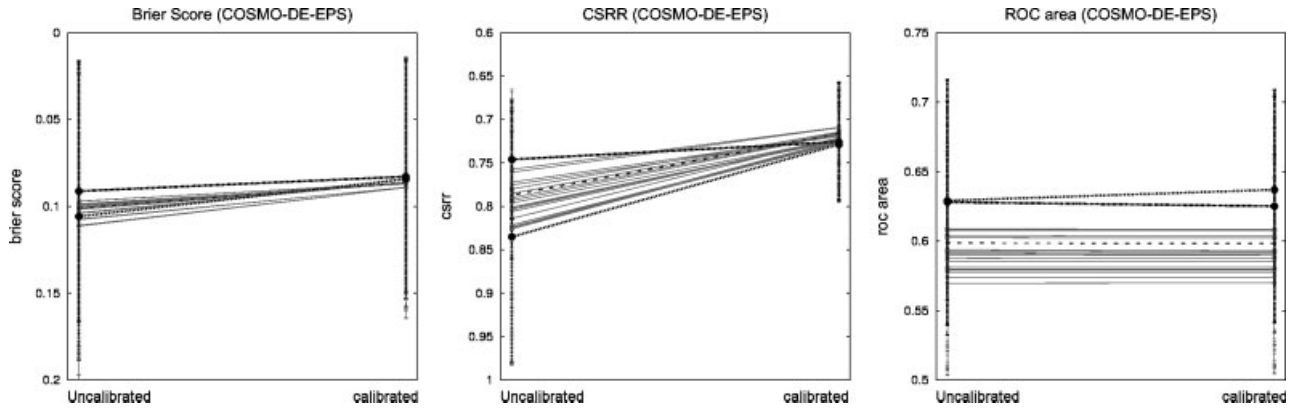

Figure 7. Effect of calibration on mean skill of COSMO-DE-EPS probabilities in Brier score, CSRR and area under ROC curve (left: uncalibrated; right: calibrated). Dotted line: fraction method; dashed: member 1; solid grey: all other neighbourhood members; dot–dashed: mean of neighbourhood probabilities and error bars indicating the standard deviations.

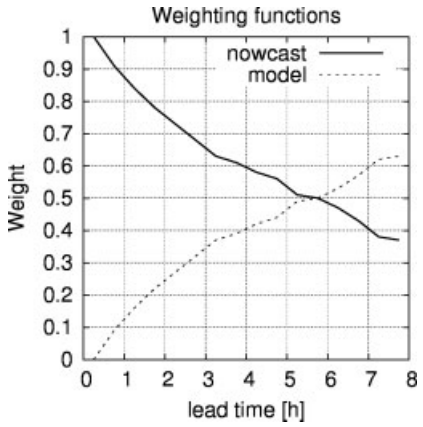

Figure 8. Weighting functions for Rad-TRAM and COSMO-DE-EPS forecasts in the blending procedure.

The resulting weighting functions are displayed in Figure 8. The crossover point is at 5.75 h. That means that after 5.75 h more weight is given to the model forecasts. Note that the maximum weight for the model is 0.63, reflecting the fact that Rad-TRAM forecasts with long lead times can also have skill and the differences between Rad-TRAM and COSMO-DE-EPS are smaller for long than for short lead times.

The weighting functions are applied to the single forecasts according to

$$P_{\mathrm{blend},i} = w_{\mathrm{r}}(\tau)P_{\mathrm{LL}}(\tau) + w_{\mathrm{c}}(\tau)P_{\mathrm{EPS},i}. \quad (5)$$

This combination of the probabilities based on Rad-TRAM (P_{LL}) and COSMO-DE-EPS ($P_{\mathrm{EPS},i}$) at each time in the respective 8 h interval results in blended probabilities $P_{\mathrm{blend},i}$, with i being the 22 respective COSMO-DE-EPS forecasts. All forecasts derived from COSMO-DE-EPS are treated with the same weight w_{c}, as differences between the methods turned out to be small in the evaluation (Figure 6).

Figure 9 displays two examples of the components, P_{LL} and $P_{\mathrm{EPS},i}$, and the resulting blended probability field $P_{\mathrm{blend},i}$ for two different lead times valid on 12 August 2007, 2315 UTC. The upper row (Figure 9(a)–(c)) represents forecasts at a lead time $\tau = 1.25$ h. At this lead time, the Rad-TRAM forecast (Figure 9(a)) is multiplied by a larger weight w_{r} than the COSMO-DE-EPS forecast (fraction method, Figure 9(b)). Therefore, the combined probability field (Figure 9(c)) reflects the high probabilities from the Rad-TRAM forecast. Nevertheless, the influence of the forecast with the fraction method is visible in additional small probabilities. The lower row (Figure 9(d)–(f)) displays forecasts at a lead time $\tau = 7.25$ h. The model forecast is the same as in Figure 9(b), as only one model run per day is available. At this lead time, the weight for the model w_{c} is larger than for Rad-TRAM. Therefore, the blended probability field (Figure 9(f)) is dominated by the fraction method forecast. The probabilities of both components are low and therefore the combined probability is low as well.

Comparing Figure 9(c) and (f), it is not possible to deduce which component leads to which pattern in the blended probability field. This illustrates that the blended forecasts deliver a seamless combination of Rad-TRAM and COSMO-DE-EPS based probabilistic forecasts.

4.2. *Quality of blended probabilities*

A quality evaluation based on the various scores is conducted for the combined probabilities $P_{\mathrm{blend},i}$ in the same way as in section 3.3. The skill of the blended forecasts should be at least as high as that of the respective best single forecast at each lead time. The Brier score of the blended probabilistic forecasts very well reflects the high skill of Rad-TRAM forecasts at short lead times (Figure 10, left). The decrease of skill with lead time for short lead times is comparable to that of Rad-TRAM alone. For long lead times the rate of decrease becomes smaller, as the COSMO-DE-EPS forecasts have a larger weight. The variability in terms of the standard deviation remains high for the combined probabilities. The differences between the methods applied to COSMO-DE-EPS are small with and without the combination. Therefore, no ranking of the methods can be identified.

Likewise, the CSRR of the combined probabilities reproduces the high skill of Rad-TRAM at short lead times and the decrease with increasing lead time (Figure 10, middle). For longer lead times, however, the blended probabilities have higher skill than Rad-TRAM alone (compare the middle panels of Figures 6 and 10). The variability of the mean values is within the magnitude of that for the single forecasts and smaller than the decrease of the mean value with lead time.

The area under the ROC curve also shows a steady decrease with lead time (Figure 10, right). There is a large variability between the solutions based on different COSMO-DE-EPS methods, starting at the second forecast hour. The ranking of the methods is consistent with the evaluation in Figure 6, where the fraction method and the mean method outperform the neighbourhood members. However, an increase in comparison to each component's skill alone can be seen for lead times around the crossover time (compare the right panels of Figures 6 and 10). For example, after four hours, Rad-TRAM only has 68% of its initial skill, which is then the same as the fraction method's, but the blended probabilities based on the fraction method still have 73%.

For all scores, Figure 6 showed that Rad-TRAM is superior for short lead times and COSMO-DE-EPS for longer lead times. At each lead time the blended forecasts perform at least as well as the single forecasts. For lead times around the crossover time, an improvement in the mean performance through the blending procedure is seen for all methods in all scores.

5. Discussion

As demonstrated in the previous section, the combination of the probabilistic nowcasting method Rad-TRAM with COSMO-DE-EPS facilitates a seamless prediction of convective precipitation for lead times from 0–8 h. For this purpose, Rad-TRAM has been extended to consider the intrinsic uncertainty in extrapolation forecasts. The output of COSMO-DE-EPS is post-processed with three different methods to derive probabilistic forecasts. The quality of the combined probabilistic forecasts is evaluated by means of three different skill scores. The skilful blending of both methods maintains the overall predictive skill for the entire forecast range. Although developed in the context of a particular nowcaster and EPS, this approach can be applied

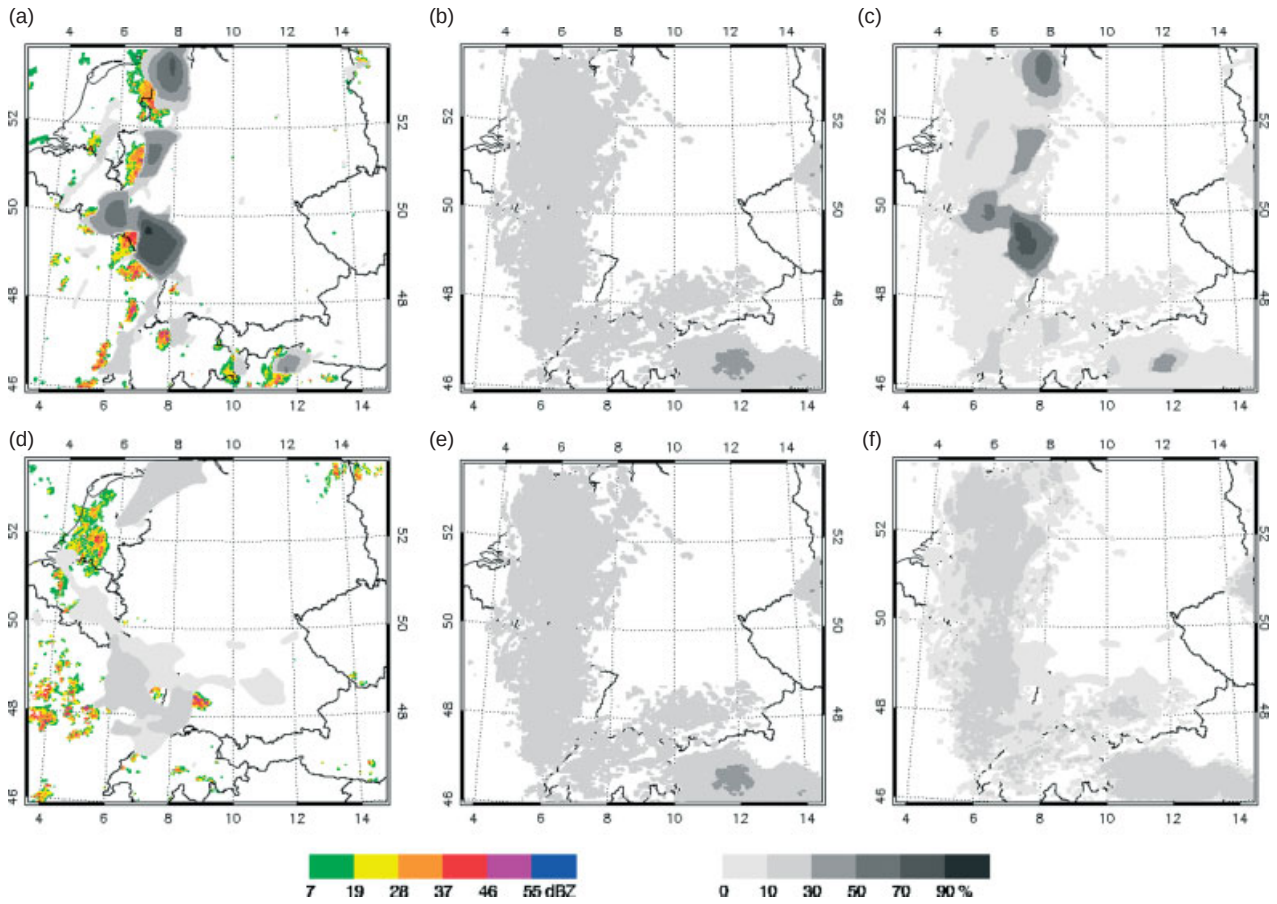

Figure 9. (a), (d) Components from Rad-TRAM and (b), (e) calibrated COSMO-DE-EPS fraction method and (c), (f) combined probabilities for 12 August 2315 UTC, for (a)–(c) $\tau = 1.25$ h and (d)–(f) $\tau = 7.25$ h, grey-shaded. Observations used to initialize the Rad-TRAM forecasts are shown in colour in the background of panels (a) and (d). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

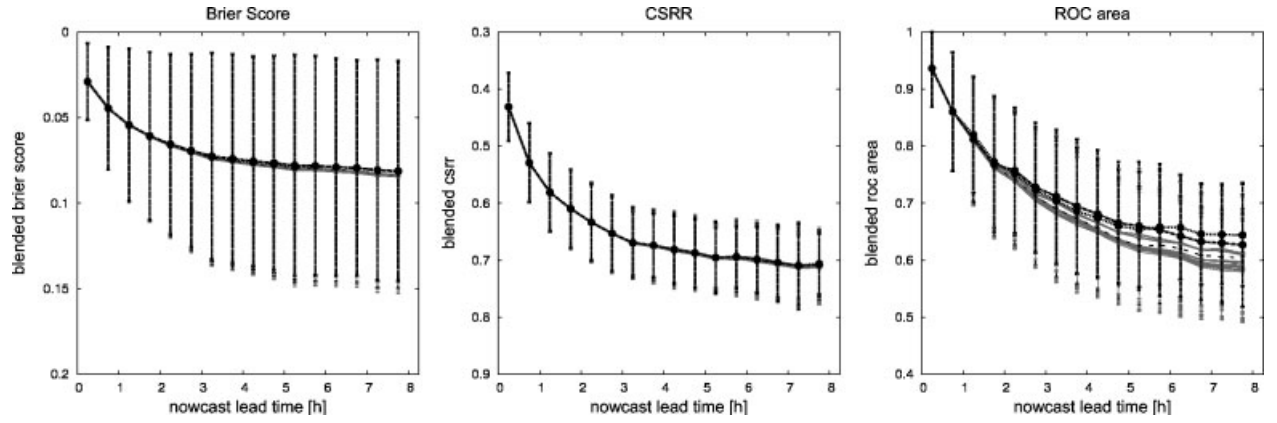

Figure 10. Evolution of Brier score, CSRR and area under ROC curve with lead time for blended probabilities from 8–16 August 2007. Dotted line: fraction; dashed: member 1; solid grey: all other neighbourhood members; dot–dashed: mean of neighbourhood probabilities and error bars indicating the standard deviations.

for forecasts of events that require the combination of any probabilistic extrapolation and NWP method.

For our study, we have chosen the probabilities of exceeding a reflectivity threshold of 19 dBZ. This corresponds to a precipitation rate of about 1 mm h^{-1}, but for the summertime period under consideration virtually all precipitation is associated with convective storms. The use of a higher threshold value would focus attention on more intense convective cores, however sensitivity studies applying a threshold of 37 dBZ resulted in far fewer events. In this case, the statistical evaluation of the probabilistic forecasts fails due to the limited amount of data. For other meteorological situations (mesoscale convective systems) a higher threshold might be appropriate. A lower threshold of 7 dBZ was not chosen, as observations from the European radar composite often contain outliers at this value. It is known that the choice of precipitation threshold influences the forecast quality (Bowler *et al.*, 2006), so it would be of interest to explore other thresholds in future work.

The deterministic nowcast tool Rad-TRAM (Kober and Tafferner, 2009) is extended by considering the variability in the precipitation field around each grid point (Germann and

Zawadzki, 2004). The fraction of precipitation pixels in a predefined search area is extrapolated with the displacement vectors. This fraction is highly dependent on the size of the search area. In this study, the side length of the search area is increased linearly during the first 4 h of the forecast (in agreement with Germann and Zawadzki, 2004). After this time, the search area is kept constant. This means that for lead times from 4–8 h the difference between forecasts at different lead times can only be due to the length of the displacement vector (i.e. position of the probability field).

The choice of the growth rate of the search area is certainly problem-dependent. As Rad-TRAM has higher skill in situations in which the evolution of the precipitation field is dominated by advective processes, the growth of the search area could, for instance, be modified for frontal situations. Here, the precipitation field is more coherent and the search area could grow more slowly compared with purely convective situations. This would result in higher probabilities for longer lead times, as the temporal variability of the precipitation field is smaller.

The calculation of the displacement vector field also impacts the forecast quality. We apply the pyramidal image matcher developed by Zinner *et al.* (2008). This implies that reliable displacement vectors can only be calculated in a neighbourhood of a grid point where precipitation actually occurred. Therefore, our approach is not semi-Lagrangian as in Germann and Zawadzki (2002), but we extrapolate linearly with the vector defined at the point of interest centred in the search area (called in Germann and Zawadzki (2002) *constant vector*). Inclusion of rotational motion as performed by Germann and Zawadzki (2004) would require a change in the derivation of the displacement vectors. For the domain of this study, which is significantly smaller than the domain used by Germann and Zawadzki (2004), this effect is likely to be small and thus has little influence on the results. For the model forecasts, the search area appears in the neighbourhood method (Theis *et al.*, 2005) and implicitly in the mean method. As discussed in section 2.3, the fraction of precipitation pixels is computed for a square region of a fixed side length of 75 km.

A distinct ranking of the different methods applied to generate probabilistic forecasts from the COSMO-DE-EPS output cannot be established, as differences in the overall forecast quality among them are very small. This is in contrast to the results of Schwartz *et al.* (2010), who found that the neighbourhood method with different sizes of neighbourhood outperformed their fraction method. The results of our study suggest that the fraction method is preferred, since the computational effort is significantly smaller than for the neighbourhood and mean methods.

Interestingly, time series of the ROC area show some persistent differences in the skill of neighbourhood forecasts derived from individual ensemble members (Figure 5). If such a ranking were found to occur reliably, and if the meteorological regime were sufficiently steady, the best members could be identified or the relative skill of different members could be used in deriving forecast probabilities. It must be emphasized, however, that a much larger data base would be required to demonstrate a useful degree of persistence in the relative skill of ensemble members.

An important feature of our approach is the calibration of the NWP-derived probabilities. Calibrating a relatively rare event in an inhomogeneous precipitation field is an active field of research (Hamill *et al.*, 2008). The calibration here is conducted in a simple and straightforward way using the reliability diagram statistics method (Zhu *et al.*, 1996). All neighbourhood members are calibrated with the same calibration function. A larger amount of data would allow the derivation of more refined calibration functions for each member separately. Other more advanced approaches for the definition of probability bins are possible as well. For example, they could be defined in such a way that they are equally populated to avoid ill-sampling (Atger, 2003). However, this was not possible for this study as the data were limited. It is not yet established that more advanced approaches to calibration (e.g. Raftery *et al.*, 2005; Hamill *et al.*, 2008) will result in significant improvements to the skill of the probabilistic forecast.

Our calibration reduces the reliability component of the Brier score (Table II) and the sharpness. With a single calibration function, the spread of the neighbourhood members is reduced. Furthermore, the various methods differ marginally in their skills after calibration, as their calibration functions are similar as well. Hence, the main difference between the forecasts based on COSMO-DE-EPS with calibration and COSMO-DE-EPS without calibration is not the magnitude but the location of the probability fields.

The weighting functions are the basis for the combination of probabilistic forecasts. Here, we restricted ourselves to a single function w_r that is determined by the evolution of Rad-TRAM's forecast skill in CSRR. This score is chosen because its general decrease of skill with lead time was similar to the other scores but the standard deviations were smaller (Figure 6). It would be desirable to have a similar lead-time-dependent weighting function for the COSMO-DE-EPS output. However, due to the set-up of the ensemble runs, such a quantity was not available. We have shown that COSMO-DE-EPS based forecasts in a first approximation do not depend on lead time (section 3.2) for lead times larger than 3 h. Several model runs starting every day (Kilambi and Zawadzki, 2005), or a time-lagged ensemble, could provide the model performance as a function of lead time. This is planned in future work.

The application of the weighting functions results in blended probabilistic forecasts. The evaluation of their forecast skills with all of the quality measures used in this study shows consistently at least the same skill as the best respective single forecasts. In all scores, the skill is even improved for lead times around the crossover time. Nevertheless, the combination of Rad-TRAM and COSMO-DE-EPS could be further advanced for special meteorological regimes. For example, in advection-driven situations one could assign a larger weight for longer lead times to Rad-TRAM.

As a first attempt to construct a blending of probabilistic nowcasts and high-resolution NWP ensemble forecasts, the methods here have been chosen to be as simple as possible. An important factor that has been neglected is the dependence of forecast skill on weather regime. If this is different for nowcasts and ensemble forecasts, it may be possible to optimize the blending for different situations, provided that a robust and objective method is available to identify the relevant regimes. One parameter that has considerable potential for this application is the convective time-scale introduced by Done *et al.* (2006), which measures the degree to which cumulus convection is controlled by larger-scale dynamical processes. This parameter has been

shown to be a good predictor of certain aspects of forecast performance in high-resolution numerical models, e.g. Craig *et al.* (2011) and Zimmer *et al.* (2011). It could be used to construct more optimal calibration and weighting functions for short and long time-scale regimes.

In the long run, one might expect that blending of nowcasts with numerical forecasts could be replaced by direct assimilation of radar and other data into the numerical model, and indeed modern data assimilation methods have significantly improved precipitation forecasts within the first few hours. However, a significant obstacle may be posed by systematic errors in the model treatment of microphysical and other cloud processes, which will lead to forecast deficiencies even with perfect initial conditions. Another, more practical, factor is the computation time required to compute a numerical forecast. It may be some time before any model forecast that could be provided within an hour of the observation time exceeds the skill of a simple nowcasting method. The blending of nowcasts and numerical forecasts is likely to produce the best results for the foreseeable future.

6. Conclusion

In this study, the skill of a new probabilistic version of the radar tracker Rad-TRAM is compared systematically with probabilistic forecasts based on the high-resolution NWP ensemble COSMO-DE-EPS. Three techniques are introduced to derive probabilistic information from COSMO-DE-EPS. After calibration, no significant difference between the skill of the solutions is found. The probabilities based on the two forecast methods, nowcasting and NWP, are combined based on the lead-time-dependent evaluation of their skill such that a meaningful seamless probabilistic forecast is provided.

The results of this investigation are robust in terms of the applied probabilistic quality measures. In addition, the separate lead-time-dependent evaluation and evaluation of the blended probabilities reveals the same qualitative results for all three quality measures. Most importantly, the skill of the blended forecast is equal to, or even exceeds, that of the individual methods at all lead times.

Acknowledgements

We gratefully acknowledge the Deutscher Wetterdienst (DWD) for providing the European radar composite and the COSMO-DE-EPS forecasts. Matthias Steiner made helpful comments on an earlier version of this article. The reviewer's comments improved this work significantly.

References

Anandan P. 1989. A computational framework and an algorithm for the measurement of visual motion. *Int. J. Comput. Vision* **2**: 283–310.

Andersson T, Ivarsson KI. 1991. A model for probability nowcasts of accumulated precipitation using radar. *J. Appl. Meteorol.* **30**: 135–141.

Atger F. 2003. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Mon. Weather Rev.* **131**: 1509–1523.

Baldauf M, Seifert A, Förstner J, Majewski D, Rauschendorfer M. 2011. Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. *Mon. Weather Rev.* DOI: 10.1175/MWR-D-10-05013.1.

Barron J, Fleet D, Beauchemin S. 1994. Performance of optical flow techniques. *Int. J. Comput. Vision* **12**: 43–77.

Bowler N, Pierce C, Seed A. 2006. STEPS: a probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Q. J. R. Meteorol. Soc.* **132**: 2127–2155.

Brier G. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**: 1–3.

Bright DR, Mullen SL. 2002. Short-range ensemble forecasts of precipitation during the southwest monsoon. *Weather and Forecasting* **17**: 1080–1100.

Buizza R, Hollingsworth A, Lalaurette F, Ghelli A. 1999. Probabilistic predictions of precipitation using ECMWF ensemble prediction systems. *Weather and Forecasting* **14**: 168–189.

Craig GC, Keil C, Leuenberger D. 2011. Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection. *Q. J. R. Meteorol. Soc.* DOI: 10.1002/qj.929.

Dixon M, Wiener G. 1993. TITAN: Thunderstorm Identification, Tracking, Analysis and Nowcasting – a radar-based methodology. *J. Atmos. Oceanic Technol.* **10**: 785–797.

Dixon M, Li Z, Lean H, Roberts N, 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. Weather Rev.* **137**: 1562–1584.

Done J, Davis CA, Weisman M. 2004. The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmos. Sci. Let.* **5**: 110–117.

Done JM, Craig GC, Gray SL, Clark PA, Gray MEB. 2006. Mesoscale simulations of organized convection: Importance of convective equilibrium. *Q. J. R. Meteorol. Soc.* **132**: 737–756.

Fritsch JM, Carbone RE. 2004. Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Am. Meteorol. Soc.* **85**: 955–965.

Gebhardt C, Theis S, Paulat M, Bouallegue ZB. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atm. Res.* **100**: 168–177.

Germann U, Zawadzki I. 2002. Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Weather Rev.* **130**: 2859–2873.

Germann U, Zawadzki I. 2004. Scale-dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *J. Appl. Meteorol.* **43**: 74–89.

Germann U, Berenguer M, Sempere-Torres D, Zappa M. 2009. REAL – Ensemble radar precipitation estimation for hydrology in a mountainous region. *Q. J. R. Meteorol. Soc.* **135**: 445–456.

Golding BW. 1998. Nimrod: A system for generating automated very short range forecasts. *Meteorol. Appl.* **5**: 1–16.

Golding BW. 2000. Quantitative precipitation forecasting in the UK. *J. Hydrology* **239**: 286–305.

Hamill TM, Hagedorn R, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Mon. Weather Rev.* **136**: 2620–2632.

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

Keil C, Craig GC. 2007. A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Weather Rev.* **135**: 3248–3259.

Kilambi A, Zawadzki I. 2005. 'An evaluation of ensembles based upon MAPLE precipitation nowcasts and NWP precipitation forecasts'. In *32nd Conference on Radar Meteorology, Albuquerque, New Mexico, 24–29 Oct 2005*; 3 pp.

Kober K, Tafferner A. 2009. Tracking and nowcasting of convective cells using remote sensing data from radar and satellite. *Meteorol. Z.* **1**: 75–84.

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

Lewis JM. 2005. Roots of ensemble forecasting. *Mon. Weather Rev.* **133**: 1865–1885.

Li L, Schmid W, Joss J. 1995. Nowcasting of motion and growth of precipitation with radar over a complex orography. *J. Appl. Meteorol.* **34**: 1286–1300.

Lin C, Vasic S, Kilambi A, Turner B, Zawadzki I. 2005. Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophys. Res. Letters* **32**: 1–4.

Marsigli C, Montani A, Paccagnella T. 2008. 'The COSMO-SREPS ensemble for the short-range: system analysis and verification on the

MAP D-PHASE DOP'. In *Joint MAP D-PHASE Scientific Meeting-COST 731 Mid-term Seminar, Bologna, Italy, 19–22 May 2008*; pp 9–14.

Megenhardt DC, Mueller C, Trier S, Ahijevych D, Rehak N. 2004. 'NCWF-2 probabilistic nowcasts'. In *11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis, Massachusetts, 4–8 Oct 2004*; 23 pp.

Mittermaier MP. 2007. Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Q. J. R. Meteorol. Soc.* **133**: 1487–1500.

Murphy AH. 1973. A new vector partition of the probability score. *J. Appl. Meteorol.* **12**: 595–600.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Mon. Weather Rev.* **115**: 1330–1338.

Pierce CE, Ebert E, Seed AW, Sleigh M, Collier CG, Fox NI, Donaldson N, Wilson JW, Roberts R, Mueller CK. 2004. The nowcasting of precipitation during Sydney 2000: an appraisal of the QPF algorithms. *Weather and Forecasting* **19**: 7–21.

Pinto J, Mueller C, Weygandt S, Ahijevych D, Rehak N, Megenhardt D. 2006. 'Fusion observation- and model-based probability forecasts for the short term prediction of convection'. In *12th Conference on Aviation, Range, and Aerospace Meteorology, Atlanta, Georgia, 30 Jan–2 Feb 2006*; 5 pp.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**: 1155–1174.

Roebber P, Schultz DM, Colle BA, Stensrud DJ. 2004. Toward improved prediction: High-resolution and ensemble modeling systems in operation. *Weather and Forecasting* **19**: 936–949.

Rossa A, Haase G, Keil C, Alberoni P, Ballard S, Bech J, Germann U, Pfeifer M, Salonen K. 2010. Propagation of uncertainty from observing systems into NWP: COST-731 Working Group 1. *Atmos. Sci. Lett.* **11**: 145–152.

Schmid W, Mecklenburg S, Joss J. 2000. Short-term risk forecasts of severe weather. *Phys. Chem. Earth Part B* **25**: 1335–1338.

Schwartz CS, Kain JS, Weiss SJ, Xue M, Bright DR, Kong F, Thomas KW, Levit JJ, Coniglio MC, Wandishin MS. 2010. Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small scale ensemble membership. *Weather and Forecasting* **25**: 263–280.

Seifert A, Beheng KD. 2006. A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 2: Maritime vs. continental deep convective storms. *Meteorol. Atmos. Phys.* **92**: 67–82.

Sokol Z, Rezacova D. 2006. Assimilation of radar reflectivity into the LM COSMO model with a high horizontal resolution. *Meteorol. Appl.* **13**: 317–330.

Stensrud DJ, Bao JW, Warner TT. 2000. Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Weather Rev.* **128**: 2077–2107.

Stephan K, Klink S, Schraff C. 2008. Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. *Q. J. R. Meteorol. Soc.* **134**: 1315–1326.

Sugihara K, Inagaki H. 1995. Why is the 3D Delaunay triangulation difficult to construct? *Inform. Process. Lett.* **54**: 275–280.

Theis S, Hense A, Damrath U. 2005. Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorol. Appl.* **12**: 257–268.

Weusthoff T, Ament F, Arpagaus M, Rotach M. 2010. Assessing the benefits of convection-permitting models by neighborhood verification: Examples from MAP D-PHASE. *Mon. Weather Rev.* **138**: 3418–3433.

Wilks D. 2006. *Statistical methods in the atmospheric sciences*. Academic Press: San Diego.

Wilson J, Xu M. 2006. 'Experiments in blending radar echo extrapolation and NWP for nowcasting convective storms'. In *4th Conference on Radar in Meteorology and Hydrology, Barcelona, Spain, 18–22 Sep 2006*; 4 pp.

Wilson JW, Crook NA, Mueller CK, Sun J, Dixon M. 1998. Nowcasting thunderstorms: A status report. *Bull. Am. Meteorol. Soc.* **79**: 2079–2099.

Wilson JW, Ebert EE, Sleigh M, Pierce CE, Saxen TR, Roberts RD, Mueller CK, Seed A. 2004. Sydney 2000 forecast demonstration project: convective storm nowcasting. *Weather and Forecasting* **19**: 131–150.

Wong W, Yeung L, Wang Y, Chen M. 2009. 'Towards the blending of NWP with nowcast: Operation Experience in B08FDP'. In *World Weather Research Program Symposium on Nowcasting, Whistler, BC, Canada, 30 Aug–4 Sep 2009*; 14 pp.

Zhu Y, Iyengar G, Toth Z, Traclon S, Marchok T. 1996. 'Objective evaluation of the NCEP global ensemble forecasting system'. In *15th Conference on Weather Analysis and Forecasting, Norfolk, Virginia, 19–23 Aug 1996*; J79–J82.

Zimmer M, Craig GC, Wernli H, Keil C. 2011. Classification of precipitation events with a convective response time-scale. *Geophys. Res. Lett.* **38**: L05 802. DOI:10.1029/2010GL046 199.

Zinner T, Mannstein H, Tafferner A. 2008. Cb-TRAM: Tracking and monitoring severe convection from onset over rapid development to mature phase using multi-channel Meteosat-8 SEVIRI data. *Meteorol. Atmos. Phys.* **101**: 191–210.

