Probabilistic forecasting of convective precipitation by combining a nowcasting method with several interpretations of a high resolution ensemble

Dissertation

Fakultät für Physik
Ludwig-Maximilians-Universität
München

Dipl.-Met. Kirstin Kober
München

München, August 2010

Gutachter der Dissertation:

1. Gutachter: Prof. Dr. G. C. Craig
2. Gutachter: Prof. Dr. U. Schumann

Tag der mündlichen Prüfung: 30.07.2010

Contents

Zusammenfassung

Qualitativ hochwertige Vorhersagen für konvektiven Niederschlag im Bereich von 0 bis 8 Stunden Vorhersagezeit sind nur durch die Kombination verschiedener Ansätze möglich. In dieser Arbeit werden Vorhersagen einer Nowcastingmethode mit Vorhersagen, die aus einem konvektionserlaubenden Ensemble abgeleitet wurden, zu nahtlosen probabilistischen Niederschlagsvorhersagen zusammengefügt. Diese kombinierten Vorhersagen sollen die Güte des jeweils besten Vorhersageverfahrens zu den verschiedenen Vorhersagezeiten erhalten. Probabilistische Vorhersagen werden erzeugt, um sowohl die inhärente Unsicherheit beider Verfahren als auch die stochastische Natur des Phänomens Konvektion zu berücksichtigen. Zum ersten Mal werden Vorhersagen eines hochaufgelösten Ensembles, das explizit Konvektion berechnet, mit beobachtungsbasierten Vorhersagen auf ähnlicher Skala kombiniert, so dass die Darstellung des physikalischen Phänomens vergleichbar ist. Die Kombination im Wahrscheinlichkeitsraum erlaubt einen glatten Übergang von einer Vorhersagequelle zur anderen mit eindeutiger Bedeutung der kombinierten Größe.

Zur Berechnung der probabilistischen Vorhersagen mit Beobachtungsdaten wird das existierende deterministische Extrapolationsverfahren Rad-TRAM um die 'Local Lagrangian' Methode erweitert. Diese Methode berechnet die Wahrscheinlichkeit, mit der ein bestimmter Schwellenwert in der Radarreflektivität überschritten wird. Zur Berechnung der numerischen Wettervorhersagen wird das experimentelle, hochaufgelöste Ensemble COSMO-DE-EPS verwendet. Mit drei verschiedenen Verfahren werden aus den Feldern der instantanen synthetischen Radarreflektivität probabilistische Vorhersagen abgeleitet. Diese Vorhersagen werden mit der 'Reliability diagram statistics' Methode kalibriert. Die Güte der Vorhersagen des Nowcastingverfahrens und des Ensembles wird mit verschiedenen probabilistischen Qualitätsmaßen in unterschiedlichen Konfigurationen evaluiert. Die Entwicklung der Vorhersagegüte mit der Vorhersagezeit definiert die Wichtungsfunktionen für die additive Kombination der beiden Vorhersagequellen.

Die Untersuchung der Entwicklung der Vorhersagegüte von Rad-TRAM und COSMO-DE-EPS mit der Vorhersagezeit zeigt, dass die 'Cross-over' Zeit, d.h. die Zeit ab der das Modell eine höhere Güte als das Nowcastingverfahren hat, etwa im Bereich von 5 bis 7 Stunden liegt. Die Unterschiede in der Qualität der drei auf das Ensemble angewendeten Verfahren zur Ableitung der probabilistischen Vorhersagen sind gering. Deswegen werden alle Modellvorhersagen mit der gleichen Wichtungsfunktion mit Rad-TRAM verbunden. Durch die Kombination wird eine nahtlose probabilistische Vorhersage von konvektivem Niederschlag erzeugt. Die Untersuchung der Güte der kombinierten Vorhersagen zeigt, dass das Hauptziel dieser Arbeit erreicht wurde. Die Kombination die Vorhersagegüte zu den verschiedenen Vorhersagezeiten ist in sofern optimiert, als dass die Qualität der jeweils besten Methode zu den verschiedenen Vorhersagezeiten reproduziert wird. Für Vorhersagezeiten im Bereich der 'Cross-over' Zeit wird die Güte durch die Kombination sogar erhöht.

Innerhalb dieser Arbeit wurden Verfahren entwickelt und angewandt, die es ermöglichen, probabilistische Vorhersagen der Überschreitung eines Schwellenwertes in der beobachteten oder simulierten Radarreflektivität zu berechnen und diese zu kombinieren. Obwohl dies die Güte optimiert und somit differenzierte Entscheidungsfindungen unter Berücksichtigung der Zuverlässigkeit der Vorhersage erlaubt, zeigt die Studie auch, dass es Verbesserungsmöglichkeiten gibt.

Abstract

A meaningful prediction of convective precipitation for a continuous range of lead times from 0 to 8 hours requires the application of different approaches. In this work, a nowcasting method and a high-resolution ensemble are combined to provide seamless probabilistic precipitation forecasts. The overall goal of this study is to provide blended probabilistic forecasts that maintain the skill of the respective best individual forecast at the different lead times. Probabilistic forecasts are chosen to consider the intrinsic uncertainty of both methods as well as the stochastic nature of convection. The innovative aspect of this work is that for the first time high-resolution ensemble forecasts that explicitly simulate convection are combined with observations on a similar horizontal scale so that the representation of the physical phenomena is comparable. A new method is developed to perform the combination in probability space, enabling a smooth transition from one forecast source to the other without ambiguity in the meaning of the blended quantity.

Concerning the nowcast, the existing deterministic extrapolation technique Rad-TRAM is modified by the Local Lagrangian method to calculate the probability of exceeding a threshold value in radar reflectivity. Secondly, the experimental high-resolution ensemble COSMO-DE-EPS provides 20 different deterministic forecasts of synthetic radar reflectivity. Probabilistic information is derived by three different approaches from the ensemble output. These probabilistic forecasts based on the ensemble were calibrated with the reliability diagram statistics method. Various probabilistic quality measures were applied to evaluate different aspects of the forecast skill of both forecasts in different evaluation setups. The development of forecast skill with forecast lead time determines the weighting functions for the additive combination of the nowcasting and ensemble methods as function of lead time.

The evaluation of the development of skill with lead time reveals that the cross-over point, the time when the model starts to have higher skill than the nowcaster, is found in mean between 5 and 7 hours. The variability between the three approaches applied on the ensemble output is small. Therefore, all model forecasts are combined with the same weighting function to the nowcasts. The combination of both approaches through the respective weighting functions provides a seamless and skillful probabilistic forecast of convective precipitation. The evaluation of the skill of the blended probabilities reveals that the goal of this study is reached. The skill of the blended forecasts is at least as high as the one of the respective best individual forecast. For lead times around the cross-over time, the skill is even improved through the blending procedure.

This thesis provides the combination and further development of methods for the calculation of probabilistic forecasts of exceeding a threshold in observed or simulated reflectivity and a blending of both. Although this already enables a differentiated decision making with respect to the confidence of the forecast, the study yields as well that there is still room for improvement.

Chapter 1

Introduction

An accurate forecast of the future atmospheric state at different forecast lead times is of great societal and economical significance. Convective precipitation forecasts affect daily life in various sectors including aviation, construction industry, and leisure activities, but their utility may be limited by uncertainty. The quantification of forecast uncertainty in a probabilistic forecast enables a more precise decision making considering 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) models. Both show different forecast skills depending on the forecast lead time (Fig. 1.1).
Nowcasts are short-term forecasts initialised with observed patterns in remote-sensing data, e.g. areas of high radar reflectivity. These patterns represent convective elements with their own characteristic life times. Usually, the forecasts are spatio-temporal extrapolations for a lead time of maximum two hours. For very short lead times with respect to the mean life time of an observed pattern, linear extrapolation shows very high forecast skill (Fig. 1.1, dotted). However, as only advective transport is considered in most nowcasting methods, the continuous temporal evolution of precipitation fields cannot be taken into account. Attempts to include lifecycle 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 up-to-date data assimilation techniques the initial humidity fields deviate from the observed state. Furthermore, the parameterised model physics limit the predictive skill of the precipitation forecasts. Most essentially, convective elements develop during the model integration from initially small-scale cells to larger patterns. Their evolution and the turbulent character of the flow limit the predictability in the first integration hours. Together with the gradual development of precipitable water, the first forecast hours are characterised by low skill (Fig. 1.1, dashed). Nevertheless, as with increasing integration time larger scales become better resolved, NWP model forecast skill outperforms nowcasting methods after some lead time (about 6 hours in the study of Lin et al., 2005).

The intrinsic uncertainty of both methods as well as the stochastic nature of convection requires a probabilistic approach. In this work, an existing deterministic nowcasting method

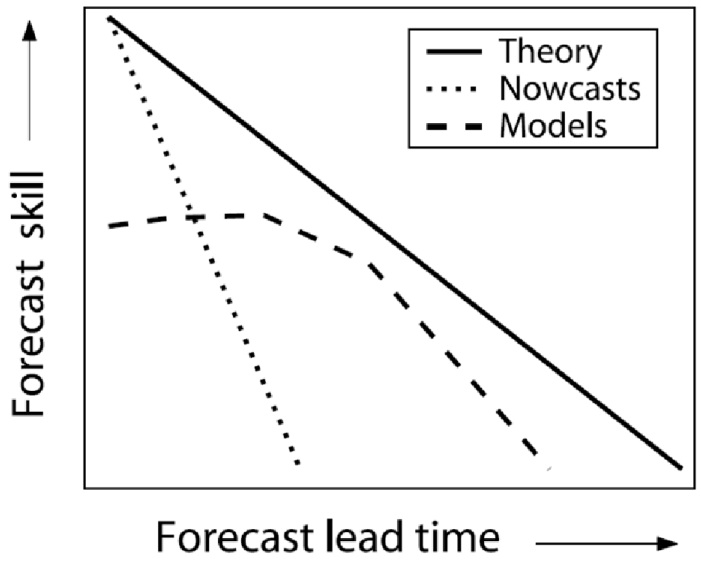

Figure 1.1: Schematic representation of the loss of forecast skill as a function of forecast lead time. The solid line represents the theoretical limit of predictability. The dashed and dotted lines correspond to numerical weather prediction models and nowcasting methods respectively (from Lin et al. (2005), following Golding (1998)).

is extended by a module considering the spatial variability and movement of the precipitation field. Additionally, high-resolution ensembles are applied to specify the variability of the precipitation fields in NWP. A seamless prediction of convective precipitation for a continuous range of lead times from 0 to 8 hours calls for the combination of both methods.
Numerous traditional nowcasting methods forecast deterministically 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 the deterministic forecasts probabilistic approaches predict the probability of exceeding a threshold in the observed field. The most straight-forward 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, 2002; Germann and Zawadzki, 2004). Germann and Zawadzki (2004) introduced and compared four methods to provide probabilistic forecasts based on continental radar observations. They concluded that the most skillful method was the Local Lagrangian method. Therefore, this method has also been adapted by others, (e.g. Megenhardt et al., 2004). These approaches do not incorporate precipitation forecasts from NWP models.
Several studies on forecasting convective precipitation with mesoscale models exist (e.g. Rotach et al. (2009)). The skill of the deterministic forecasts depends on the respective model configurations. For example, the horizontal resolution determines if convection is explicitly resolved or parameterised (Done et al., 2004). Not only the model set up impacts the forecast quality (Gebhardt et al., 2010), but also the representation of the initial fields and their discrepancies to observations. Data assimilation methods have been found to have an important influence on the behaviour of numerical forecasts (e.g. Sokol and Rezacova (2006),

Stephan et al. (2008), Dixon et al. (2009)).
In order to quantify the variability of model predictions, ensemble methods have been developed at weather prediction centres and matured to a well-established approach (reviewed by Lewis (2005)). Several approaches exist to design ensembles. Usually, perturbations of the initial or boundary conditions or perturbations of model physics in a linear or stochastic way can be applied to create different forecasts (Bright and Mullen, 2002). Furthermore, different forecast models (multi model ensemble), runs of the same forecast model starting at different times (time-lagged ensemble), and combinations thereof can be combined (Roebber et al., 2004).
Although ensemble prediction systems (EPS) have advanced to a standard technique on large- and mesoscales, only a few operational convection-permitting ensembles exist, e.g. Gebhardt et al. (2010). The design of high resolution and, therefore, convection permitting ensembles differs from that of mesoscale ensembles with parameterised convection, because the error grows in a different way at smaller scales (Hohenegger and Schär, 2007). The EPS of the Deutscher Wetterdienst (DWD) is based on the 2.8 km grid space configuration of the COSMO[1] deterministic forecast model (COSMO-DE-EPS, Gebhardt et al, 2010). In this experimental high resolution ensemble, only boundary conditions and physical parameterisations are varied to maximise the spread in precipitation forecasts at short lead times, see Stensrud et al. (2000).

The skillful combination of nowcasting methods and NWP models to forecast precipitation has the potential to maintain the overall predictive skill for the continuous range of lead times from 0 to 8 hours. Usually, the combined prediction is the weighted sum of both methods. The weighting functions are determined by calculating the skill of the forecasts with suitable quality measures. Several studies identify the forecast skill of nowcasting and NWP models using deterministic (e.g. Golding, 1998; Kilambi and Zawadzki, 2005) or probabilistic (Bowler et al., 2006) quality measures. The evaluated quantity is either radar reflectivity (Wilson and Xu, 2006), rainfall rate (Golding, 1998), or probability of precipitation (Pinto et al., 2006). The combination is 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). Most of the mentioned methods use coarse resolution models (larger 10 km) where convection is parameterised.

The aim of this work is to create an optimal forecast of precipitation for the range from 0 to 8 hours by combining extrapolated radar observations with numerical weather prediction. The key elements are the probabilistic approach for a seamless integration taking into account errors of each method at different lead times and the newly available 'cloud-resolving' ensemble that represents a level of detail comparable to the radar. And finally, the careful quantitative formulation of the methods and the evaluation of their performance. The combination of these elements is new, and has potential for significant advance towards above goal. 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 (real 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

[1]Consortium of Small-scale Modeling (COSMO)

the other, and reflects the increasing uncertainty of the prediction with increasing lead time. The goal of this dissertation is to describe how the probabilistic nowcasts and forecasts are created, combined, and then evaluated. The goal of this evaluation is to demonstrate that the combined forecast matches or exceeds the performance of the individual components at all lead times.

As summertime precipitation is mainly based on convection, the first section of Chapter 2 gives a brief introduction about the physics of convective precipitation. Radar as one method to measure precipitation is explained and the quantity that is used in this work is introduced. Two methods to forecast convective precipitation are introduced: nowcasts and NWP models. For both, existing approaches are briefly reviewed and the approaches used in this work are introduced. The third section of this chapter explains different quality measures to quantify the quality of the probabilistic forecasts. Finally, the probabilistic forecasts of COSMO-DE-EPS are calibrated. Chapter 3 investigates the two forecasts‘ quality in three different case studies representing different meteorological situations in time series and lead time dependent. Chapter 4 repeats this evaluation for the entire investigated period. The knowledge about the development of forecast skill with lead time is the basis for the definition of the weighting functions in the additive combination (Chapter 5). Here, an overview is given over existing approaches before the blending method applied in this work is defined. Finally, the quality of the blended probabilistic forecasts is evaluated for the case studies of Chapter 3 and the entire period. These findings are discussed in the context of current knowledge in Chapter 6. Chapter 7 summarises this study and suggests possible future work.

Chapter 2

Probabilistic forecasts of convective precipitation

Summertime precipitation is mainly caused by convective processes. Therefore, the first section will introduce the main physical principles of convection and its appearance in the atmosphere. There are several possibilities to measure precipitation. Amongst these, Radar is the most suitable technique for this work, since it provides data with high temporal and spatial resolution. In this study, radar reflectivity will be used as predictor and hence, this measured quantity and its relation to precipitation will be explained briefly. One possibility to forecasts precipitation are methods based on observation data, also referred to as nowcasting methods. Existing approaches will be reviewed in a literature overview and the method applied in this work will be explained (section 2). The new probabilistic extension of the already existing radar tracker Rad-TRAM (Radar Tracking and Monitoring) will be introduced. The second possibility to forecast precipitation are numerical weather prediction (NWP) models that solve the equations of motion (section 3). Ensembles enable to quantify the variability in the numerical solutions. The experimental ensemble of the Deutscher Wetterdienst (DWD), COSMO-DE-EPS, is the model used in this study. With three methods probabilistic forecasts will be derived from the output of this ensemble. The quality of the probabilistic forecasts provided from both the nowcaster and the model can be quantified with quality measures. Therefore, in section 4, different aspects of quality and the respective probabilistic quality measures will be introduced. In the last section, one of these methods is applied in order to calibrate the forecasts based on COSMO-DE-EPS.

2.1 Convective precipitation and its Observation

This section will introduce the basic physical concept of convection and different realisations and organisations of convective precipitation in the atmosphere. Furthermore, one method to observe precipitation, Radar, will be described. As in this work the main concern is forecasting convective precipitation, only the main principles will be explained briefly. A more detailed discussion can be found in the literature cited respectively.

2.1.1 Convective Precipitation

Definition of Convection

Generally, convection is the transport of a physical property like energy, momentum, or mass in fluids or gases. It is next to conduction and radiation a mechanism to transport latent heat. In the atmosphere, convection is the thermally driven current that is initiated if gravity balances an instable vertical mass distribution. To illustrate stability in moist air, the ascent of an isolated air parcel is regarded. Buoyancy per unit mass of an air parcel mainly depends on differences in density and temperature of the parcel and the environment and is calculated with Archimedes' principle (e.g. Holton (2004), Emanuel (1994), or Smith (1997)):

$$B = g\frac{T_{vp} - T_v}{T_v} \tag{2.1}$$

with g the acceleration due to gravity, T_{vp} being the virtual temperature[1] of the parcel, and T_v the virtual temperature of the environment. An unsaturated parcel cools while ascending with the dry adiabatic lapse rate. Through this cooling, the parcel can reach saturation so that condensation of water vapour begins. This is accompanied by a release of latent heat. If the parcel still has positive buoyancy, it ascends with a moist adiabatic lapse rate that is smaller than the dry adiabatic lapse rate. The stability of an air mass can be classified with the (measured) lapse rate in comparison to the dry or moist adiabatic lapse rate to stability, instability, and conditional instability. Convection can develop in instable and conditionally instable conditions, together with the availability of moisture.
These two conditions, moisture and sufficient instability, are combined in the convective available potential energy (CAPE)

$$CAPE = R_d \int_{p(z_2)}^{p(z_1)} T_{vp} - T_v \; d(\ln p), \tag{2.2}$$

where R_d is the individual gas constant of dry air, p is pressure, and the heights z_1 and z_2 limit the region where free ascent occurs. Normally, parcels have to be lifted until saturation occurs. At this height, condensation and the formation of clouds begin (lifted condensation level, LCL). The energy needed to lift the parcel through the stable region is the convective inhibition (CIN). If this energy is available, the free ascent of the parcel starts at the level of free convection (LFC) up to the level of neutral buoyancy (LNB) where the parcel is in a stable environment.
Therefore, convection can only occur if parcels have enough energy to overcome the stable layer at the ground (CIN). Possible trigger mechanisms for deep convection are convergence near the ground, forced lifting along a frontal zone, orography, or local heating.

Forms of Atmospheric Convection

The convective ascent of air is characterised by updraughts. Once these updraughts have formed, first condensation and later perhaps precipitation develops. Depending on the loca-

[1]The virtual temperature $T_v = T(1 + 0.608r)$ with r the water vapour mixing ratio, considers the water vapour dependence of density in moist air to modify the definition of temperature, see Emanuel (1994).

tion of up- and downdraughts, systems of different intensity and life time can evolve. Mainly, three different types of convective clouds are distinguished (Smith, 1997):

1. Cumulus clouds: they have a horizontal and vertical dimension of around 1 km and are not precipitating. Nevertheless, they contribute to the vertical exchange of latent heat.
2. Deep convective storms: their vertical dimension is significantly larger, ranging possibly to the tropopause. Intense showers or thunderstorms are possible and lead to efficient vertical exchange in the troposphere.
3. Mesoscale convective systems: are organised complexes of several single thunderstorm cells with a horizontal scale of several 100 km and long life times leading to large amounts of heavy precipitation or even hail.

2.1.2 Observation of Precipitation with Radar

Radar[2] is an object detection remote sensing system that uses electromagnetic waves to identify the range, altitude, direction, or speed of both moving and fixed objects such as aircraft, terrain, or weather formations. Here, only the application of Radar systems to observe precipitation is of interest as it is able to capture the structure of the entire system. A radar emits a pulse of an electromagnetic wave that is then reflected by the hydrometeors in a cloud. Weather radar works with frequencies within 3 and 30 GHz (in wavelength: 10 to 1 cm). The intensity of the backscattered signal is measured as a function of time. The signal itself is a function of the particle, the drop size distribution, and the thermodynamic phase of the scattering particles (Höller, 1994). Therefore, the signal allows to draw conclusions to the precipitation intensity. The temporal delay of the signal enables to locate the precipitation field.

The relation of the emitted and the received energy after scattering on hydrometeors is described with the radar equation. Due to the long wave length of the emitted wave, the backscattering volume is large. Therefore, the beam is scattered on numerous hydrometeors and the equation has to be formulated for volume objects. The relation between the transmitted power P_t and the reflected power P_r is (Battan, 1973)

$$P_r = \frac{P_t g^2 \lambda^2 \theta_0^2 h}{1024 ln(2) \pi^2 r^2} \sum_{Vol} \sigma_i. \tag{2.3}$$

Here, g is the antenna gain, λ the wavelength of the transmitted electromagnetic wave, θ_0 the opening angle, h the pulse length, σ_i the backscatter cross-section of the single scatterer i, and r the distance from the target to the radar. Eq. (2.3) is the basis for quantitative precipitation estimation (Rinehart, 1997). In this estimation, the loss of power in the radar system and attenuation in the atmosphere is not considered.
The cross-section in Eq. (2.3) is defined as

$$\sigma = \frac{\pi^5}{\lambda^4} |K|^2 D^6, \tag{2.4}$$

with the particle diameter D and the dielectric constant K (Doviak and Zrnic, 1984). In the formulation of the backscatter cross-section, Rayleigh scattering is assumed (valid for

[2]Radio detecting and ranging

particles that have a smaller size than the wavelength). The dielectric constant K is defined as

$$|K|^2 = |\frac{m^2 - 1}{m^2 + 2}|, \tag{2.5}$$

where the refraction index m depends on the phase, temperature, and wavelength.
To relate the measurement to the physical characteristics of precipitation, the radar reflectivity factor z is introduced as

$$z = \sum_{vol} D_i^6. \tag{2.6}$$

z is a cloud physical quantity and the 6th moment of the drop size distribution. Its advantage is that it is independent of wavelength and therefore, measurements of different wavelengths can be compared.
If the radar equation (Eq. 2.3) is formulated with the reflectivity factor (Eq. 2.6), a relationship between received power and a cloud physical quantity is formulated:

$$P_r = \frac{C|K|^2 z}{r^2}, \tag{2.7}$$

with C denoting the constants in Eq. (2.3). The reflectivity factor ranges over several magnitudes and therefore, the logarithmic formulation is used:

$$Z = 10log_{10}(\frac{z}{1\ mm^6 m^{-3}}). \tag{2.8}$$

Normally, displays of radar data show Z. Due to brevity, the radar reflectivity factor is often denoted simply as radar reflectivity or reflectivity. In this work, Z is the quantity that is used.

2.2 Forecasts based on Observations

2.2.1 Brief review of existing forecasting methods

One approach to accurately forecasting convective precipitation are so-called nowcasting methods. Nowcasting denotes very short term forecasts based on observation data. Two main approaches are distinguished: extrapolation methods and storm generating methods (Barillec and Cornford, 2009; Pierce et al., 2004).
Extrapolation methods predict the evolution of the observed rainfall field using object tracking and advection-based techniques (Wilson et al., 1998). Therefore, they rely on the assumption that the temporal evolution of the rain field is governed by motion. The objects are identified in radar, satellite or lightning data by applying one or a combination of several thresholds. Most nowcasting methods are radar-based. Notable examples are TITAN (Thunderstorm Identification, Tracking, Analysis, and Nowcasting, Dixon and Wiener (1993)), NIMROD (Golding, 1998, 2000), TREC (Tracking Radar Echoes by Correlation, Rinehart

and Garvey (1978))/COTREC (Continuity of TREC, Li et al. (1995), and Rad-TRAM (Radar Tracking and Monitoring, Kober and Tafferner (2009)). Examples for tracking algorithms based on satellite data are MASCOTTE (Maximum Spatial Correlation Tracking Technique, Carvalho and Jones (2001)), RDT (Rapid Developing Thunderstorms, Morel and Senesi (2002)), and Cb-TRAM (Cumulonimbus Tracking and Monitoring, Zinner et al. (2008)). As physical processes are not included in these methods, the temporal thermodynamical evolution of the precipitation field cannot be considered. Some of the algorithms monitor the development of the field over several timesteps and can therefore, nowcast an expected trend. More advanced systems combine a variety of data sources (radar, satellite, lightning, wind profiles, NWP) in order not only to predict the storm's track but also areas where convection initiation is very likely (Auto-Nowcaster, Mueller et al. (2003)) or include a conceptual life cycle (GANDOLF, Pierce et al. (2000)).

The second category, the storm generating methods, focus on the birth, growth, and dissipation of storms and can be divided into point process models and multifractal models. Point process models estimate the internal dynamics of precipitation fields using statistical representations where the occurrence of precipitation objects is governed by Poisson point processes (Barillec and Cornford, 2009). Multifractal approaches are based on the finding that precipitation fields show statistical invariance with respect to the scale at which they are observed (Lovejoy and Mandelbrot, 2010). Examples for systems based on this approach combined with dynamics are S-PROG (SPROG, Seed (2003)) and STEPS (Short-Term Ensemble Prediction System, Bowler et al. (2006)).

Furthermore, it has to be distinguished if the above mentioned nowcasting systems provide single point forecasts or probabilistic forecasts. Probabilistic forecasts consider that the observations as well as the models cannot exactly represent the true process but contain errors and approximations and are therefore, a measure for the prediction's uncertainty (Barillec and Cornford, 2009). Most existing studies calculate a probability of exceedance of a threshold based on the fraction of precipitation pixels near the point of interest (Bowler et al., 2006). Examples are Andersson and Ivarsson (1991), Schmid et al. (2000) and Schmid et al. (2002), and Germann and Zawadzki (2004).

The following section will describe briefly the originally deterministic DLR radar tracker Rad-TRAM, and the European radar composite on which it is based (Fig. 2.1). The emphasis will be put on the calculation of the motion field as this is essential for the probability operator. More details of the algorithm can be found in Kober and Tafferner (2009) and Zinner et al. (2008).

2.2.2 European Radar Composite

The databasis used for Rad-TRAM is the European radar composite issued by the Deutscher Wetterdienst (DWD) (Fig. 2.1). It consists of radar reflectivities given in six dBZ classes with a horizontal resolution of 2 $km \times 2$ km and encompasses an area of 1800 $km \times 1800$ km (Weigl et al., 2005). This spatial coverage is unique for radar data in Europe and therefore chosen in this study. Fig. 2.2 shows the area that is covered by the European radar composite together with the evaluation domain of this study.

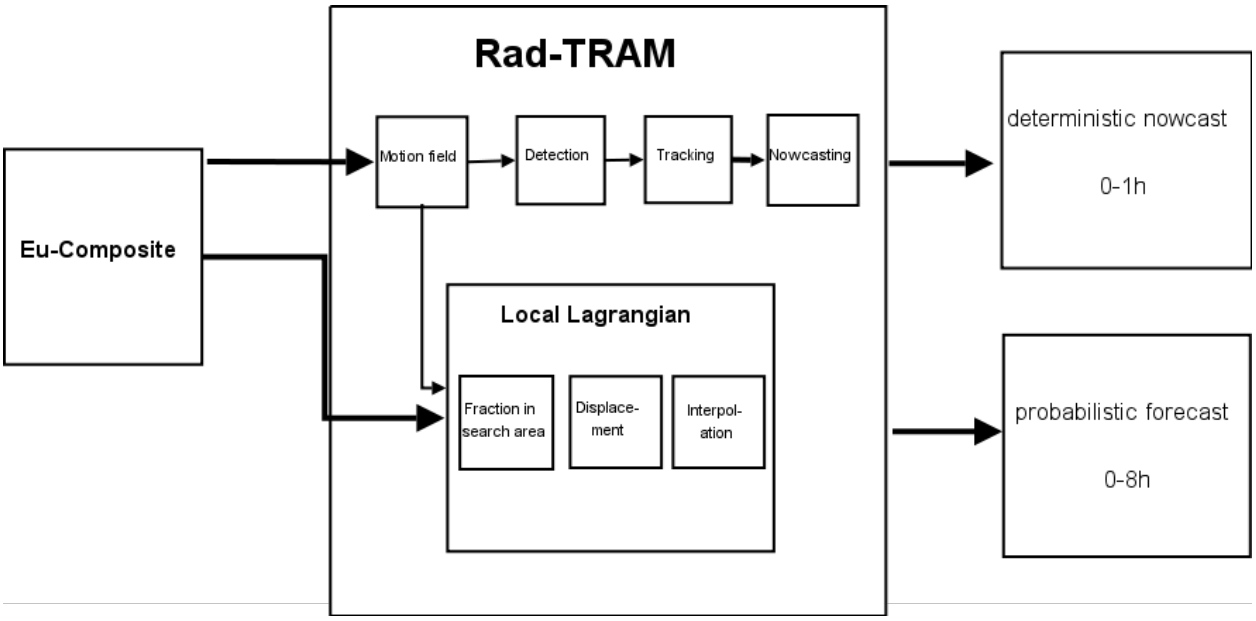

Figure 2.1: Schematic overview over Rad-TRAM.

The radar reflectivity values are observations received from 3-dimensional radar scans of various radars across Central Europe. As there is no common scan strategy for all countries not every value at every pixel represents the value of the lowest scan in the vertical where an echo is found (as it is defined in Germany), but also the highest values as representatives are possible. In case observations from two or more radars overlap at one point, the maximum of these observations is chosen (Weigl et al., 2005). Also, the national constituents do not have identical level boundaries but are adjusted during the composite procedure to the six reflectivity classes used by DWD (Tab. 2.1).

The radar composite provides a synopsis of the weather situation with regard to precipitation over a large domain, but individual pixel values are not representative of the actual microphysical process in place. Several factors influence the measurements (Rinehart, 1997). For example, the scanning mode changes from radar to radar. The lowest scan in mountainous regions might see the core precipitation processes of the cloud while over flat land it might see the precipitation falling out of the cloud. Due to beam blocking or clutter there can be inconsistencies in the composite, i.e. pixels with no or wrongly identified radar measurement. Data quality and resolution change with distance from the radar (attenuation, overshooting, evaporation, anomalous propagation) and are dependent on the physical processes (bright band: melting snow leads to higher reflectivities). Aside from this, data processing is not standardised among the radars in Europe. E.g. sometimes white spots appear in the composite over France which are indeed observations of very high reflectivity. These facts resulting in an inhomogeneous field must be considered when interpreting the radar composite in a quantitative sense.

2.2.3 Radar Tracker Rad-TRAM

The tracking algorithm Rad-TRAM (Radar TRAcking and Monitoring) originally consists of 4 parts (Fig. 2.1): the extraction of the motion field by solving the optical flow equation,

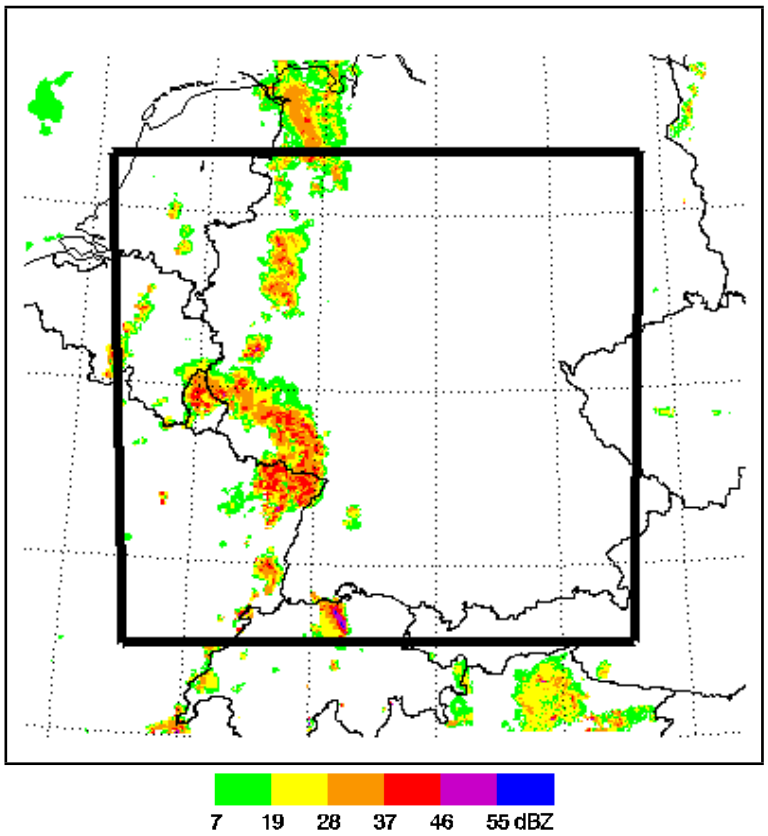

Figure 2.2: 12 August 2007, 23:15 UTC: Observed radar reflectivities in the European radar composite (DWD) with the evaluation domain of this study (black).

the detection of convective cells, the tracking and the nowcasting part (Kober and Tafferner, 2009). Rad-TRAM is upgraded to also produce probabilistic forecasts by implementing the Local Lagrangian method (Germann and Zawadzki, 2004). The algorithm description is concentrated on the extraction of the motion field as only this part is important for the implementation of the probability operator. The detection, tracking, and nowcasting parts are described briefly for the sake of completeness.

Table 2.1: Reflectivity classes of European radar composite (Schreiber (2000)).

reflectivity (dBZ)	colour	approx. precipitation types
> 55.0	blue	very heavy rain and hail, large hail possible
46.0 - 55.0	violet	very heavy rain, hail possible
37.0 - 45.5	red	moderate to heavy rain
28.0 - 36.5	orange	moderate rain
19.0 - 27.5	yellow	light rain
7.0 - 18.5	green	very light rain

Extraction of the motion field

In the first part of the algorithm, the displacement vector field is derived by solving the optical flow equation of a pair of consecutive radar images. In contrast to feature-based matchers which select a certain pattern in one image and search for it within a target area of the second one, here an area-based matcher is used: for each pixel position, a displacement vector is calculated by minimising the local squared difference between both images, or optionally, maximising the local correlation. In order to take into account that small-scale motions in precipitation fields are often superposed on the large scale flow, the 'pyramidal image matcher' has been developed which handles this scale dependency (Zinner et al., 2008). In a stepwise procedure, lower resolution images representing larger scales are created by averaging over 2^n pixels with n being the number of successive iterations. For every pixel location, a displacement vector is computed by shifting one image within the range of $+/-$ 2 pixel elements in both horizontal dimensions to calculate the best fit to the other image. After each step, the displacement vector field is interpolated to the full resolution grid and the image to be matched by this vector field is advanced. These steps are repeated at successively finer scales with decreasing n. Finally, the displacement vector field is the sum of the displacement vectors derived at the different resolutions.

Detection

In Rad-TRAM, convective cells are identified as areas reaching or exceeding a certain threshold. In the original version of Rad-TRAM, it was fixed at 37 dBZ, but in this study the threshold is 19 dBZ, because the focus is set on convective precipitation. The first threshold was chosen with the focus on thunderstorms. A cell must consist of at least three neighbouring pixels. Due to the applied circular smoothing in the detection algorithms, such small cell elements are extended to at least 21 contiguous pixels.

Tracking

Detected convective cells are tracked in consecutive images by using the displacement vectors derived in the pyramidal image matcher together with the method of maximum overlap. The short refresh cycle of 15 minutes makes this approach feasible. Based on the detected cells at time $t-1$ the motion field is used to estimate the position of the cells at observation time t (first guess patterns). The extrapolated cells are overlaid with the observed cells at time t. The observed cell which shows the maximum overlap with the first guess pattern adopts the cell's history. If no overlap is found, a new cell is created and the old one disappears. Lines connecting the cell's centre of gravity at consecutive times display the cell's track. Here, the centre of gravity is the intensity weighted centre of the cell.

Nowcasting

A further application of the displacement vector field is the generation of deterministic very short range forecasts. Extrapolating the pixel positions of the detected cells for four time steps provides nowcasts up to one hour. Using the displacement vector field for every cell pixel instead of translating the cell as a whole enables the cell to change size and shape, thereby taking into account the trend.

Probabilistic Rad-TRAM

Rad-TRAM is upgraded to produce probabilistic forecasts by implementing a concept similar to Local Lagrangian (Germann and Zawadzki, 2004). This part of the algorithm is independent of the objects described before but also makes use of the pixel-based scale-dependent displacement vector field (Fig. 2.1). Germann and Zawadzki (2004) identify two main error sources in forecasts based on persistence: wrong displacement and processes other than advection which cannot be described. The thermodynamical evolution of the precipitation field (growth or decay) is not represented. They assume that the error based on the not represented temporal evolution dominates. This assumption was later proven in Bowler et al. (2006) who showed that the error of wrong displacement is 10 % of the total error. Furthermore, Germann and Zawadzki (2004) assume that the rate of temporal evolution is related to the spatial variability of the field. Therefore, the spatial variability around each grid point in the domain can be used in order to quantify the uncertainty resulting from the precipitation field's development.

Thus, the probabilistic forecast P_{LL} of exceeding a threshold $\mathcal{L}$ is provided by deriving the probability distribution via the variability in the precipitation field ψ in the search area of the point of interest (ω_k) with the scale parameter k (side length of search area). The value is extrapolated with the displacement vector α defined on location x. The length of the vector depends on the forecast lead time τ.

$$P_{LL}(t_0 + \tau, x, \mathcal{L}, k) = Prob\{\psi(t_0, x - \alpha + r) \geq \mathcal{L} | (x + r) \in \omega_k\} \tag{2.9}$$

$x + r$ describes every position within the search area ω_k. This procedure is applied on every grid point in the domain. Finally, smoothing based on Delaunay triangulation (Sugihara and Inagaki, 1995) is applied on the probability field P_{LL} to eliminate possible gaps resulting from divergent displacement vectors. The threshold $\mathcal{L}$ is in this study fixed at 19 dBZ. Probabilistic forecasts are created up to 8 hours lead time in 15 minutes time steps. The size of the search area increases with lead time in the first 4 forecast hours as the uncertainty coming from the not represented temporal evolution increases. Following Germann and Zawadzki (2004), the side length of the search area ω_k is assumed to grow linearly with 1 km per lead minute. From hour 4 to 8 the size of the search area is kept constant as it is assumed that the correlation distances are saturated. Therefore, the maximum side length is 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.

A typical result of the applied probability technique for 12 August 2007, 23:15 UTC based on different lead times is illustrated in Fig. 2.3. Additionally, the reflectivity fields at the initial time which are the basis for respective forecast are displayed. The 15 minutes forecast provided at 23:00 UTC is very sharp and reflects the low uncertainty for short lead times (τ = 15 min, Fig. 2.3a). The forecast calculated on basis of the reflectivity observation one hour ago (τ = 60 min, Fig. 2.3b) already demonstrates the increased uncertainty by 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 related observation (cf.

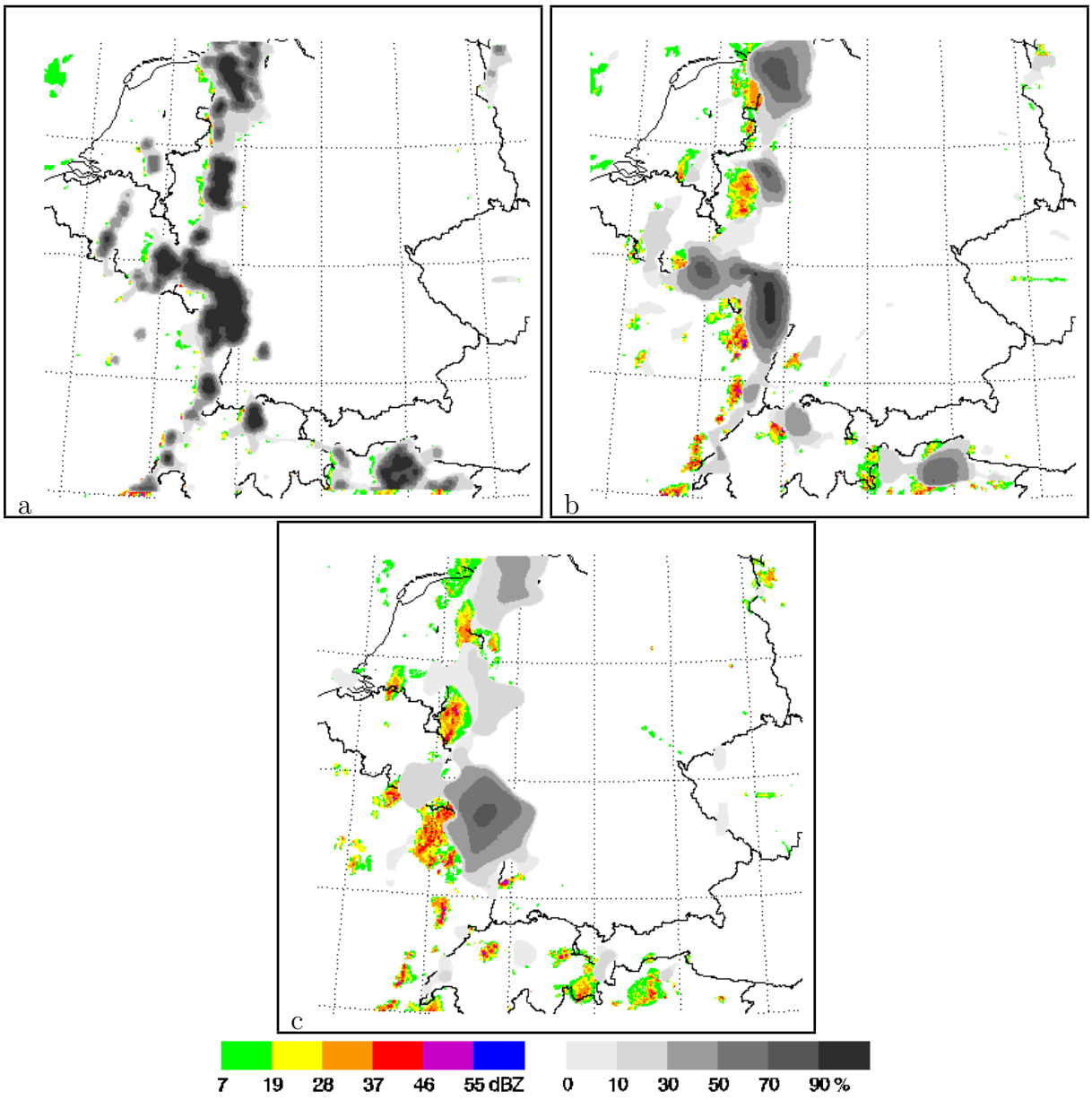

Figure 2.3: Probabilistic forecasts P_{LL} of Rad-TRAM for 12 August 2007, 23:15 UTC based on (a) 15 min forecast from 23:00 UTC, (b) 60 min forecast from 22:15 UTC, and (c) 120 min forecast from 21:15 UTC grey-shaded. In the background are colour-coded the reflectivity observations at the respective initial time.

Fig. 2.2) reveals that the forecast still has skill concerning the position of the probability field. The forecast based on the observation two hours before (τ = 120 min, Fig. 2.3c) shows a further smoothed probability field. The position of the field in comparison with the observation is still meaningful. However, as the probability field covers a larger area, there are some false alarms where a probability of exceeding 19 dBZ was predicted, but not exceeded in the observations.

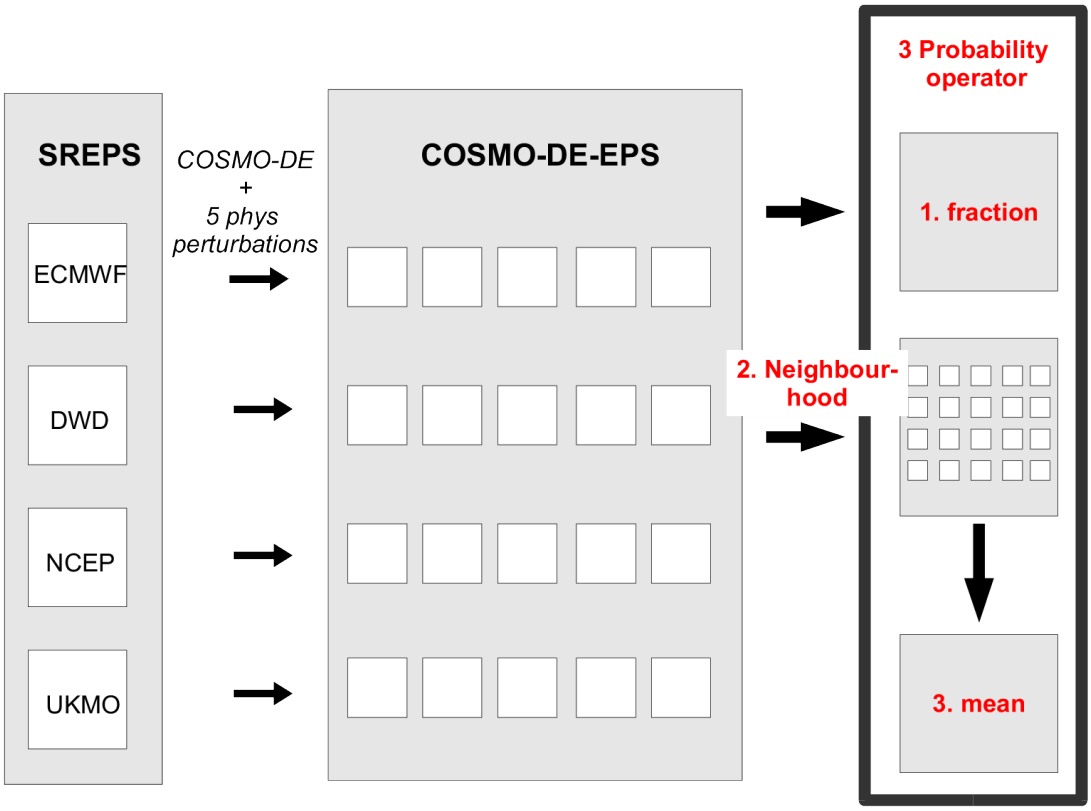

Figure 2.4: Schematic overview over COSMO-DE-EPS and the derivation of probabilistic forecasts from the ensemble.

2.3 Forecasts based on Numerical Weather Prediction

A second possibility to address the problem of forecasting convective precipitation is using forecasts based on numerical weather prediction (NWP). High-resolution meso- and storm-scale models with largely explicit precipitation physics are of special interest as they have the highest potential for forecasting convective precipitation. It has to be distinguished if radar data is used to initialise the model or not (Wilson et al., 1998; Bowler et al., 2008). Applying the ensemble approach to high resolution models enables the characterisation of model prediction uncertainty (Lewis, 2005). There are several sources for model uncertainty and therefore, several approaches in designing ensembles exist. Usually, perturbations of the initial or boundary conditions or perturbations of model physics in a linear or stochastic way are applied to create different forecasts. Furthermore, different forecast models (multi model ensemble), runs of the same forecast model starting at different times (time-lagged ensemble), and combinations thereof are possible (Roebber et al., 2004).

2.3.1 COSMO-DE-EPS

The COSMO-DE-EPS is the experimental ensemble based on the COSMO-DE (Gebhardt et al., 2010). Fig. 2.4 summarises schematically how the ensemble is created. The COSMO-DE (formally known as LM-K (Baldauf et al., 2006)) is a non-hydrostatic and convection-

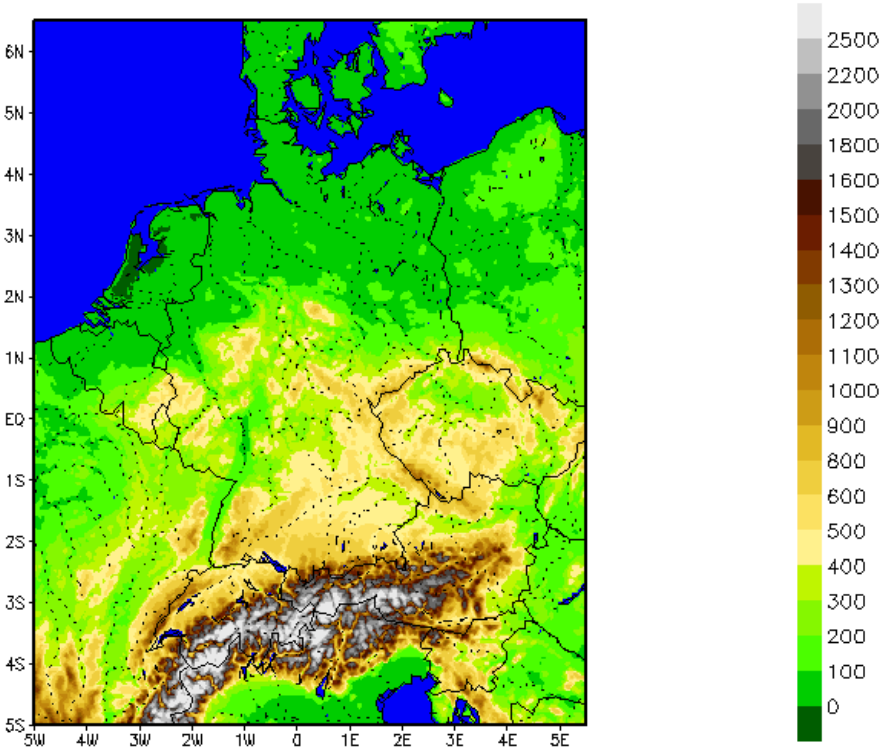

Figure 2.5: Orography (height in m) of the operational domain of the COSMO-DE at DWD.

permitting weather forecasting model for very short-range forecasts of the DWD. Fig. 2.5 shows the domain and orography on which COSMO-DE is run. COSMO-DE has been developed in the framework of the **Co**nsortium of **S**mall-scale **M**odeling (COSMO). The horizontal resolution is 2.8 km and 50 vertical levels are used up to 30 hPa. Precipitation processes are explicitly described using a bulk-type cloud micro-physical scheme containing five prognostic hydrometeor types (rain, snow, cloud water, cloud ice, and graupel). There is no parameterisation of deep convection, whereas shallow convection is still parameterised with the Tietdke scheme (Tiedtke, 1989).

COSMO-DE-EPS consists of 20 members and is created by addressing several sources of uncertainty. First, the uncertainties due to the lateral boundaries conditions are considered by nesting the COSMO-DE into four different members of COSMO-SREPS (Short-Range Ensemble Prediction System, resolution: 10 km) (Fig. 2.4). These members result of a further nesting technique and finally represent different global models from different national weather services: ECMWF[3], DWD, NCEP[4] and UKMO[5] (Fig. 2.6a) (Marsigli et al., 2008).

Second, uncertainties in model physics are considered by perturbing five different parameters of the physics scheme in a non stochastic approach (Fig. 2.6b and Tab. 2.2). The parameters in the physical parameterisations are chosen and perturbed such that variability in the precipitation forecasts is maximised.

[3]European Centre for Medium-Range Weather Forecasts
[4]National Center for Environmental Prediction
[5]United Kingdom Meteorological Office

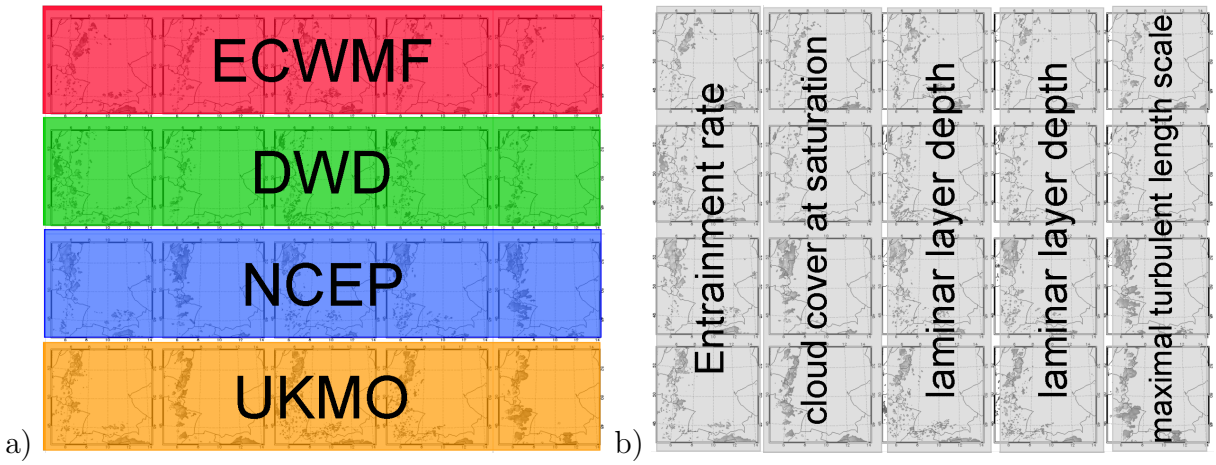

Figure 2.6: Creation of COSMO-DE-EPS: (a) different lateral boundary conditions and (b) different perturbed physical parameters.

The entrainment rate (first perturbed parameter in Tab. 2.2) is part of the parameterisation of shallow convection. In contrast to deep convection, shallow convection is still parameterised in COSMO-DE and especially needed to transport moisture out of the boundary layer. Therefore, it prevents the occurrence of too many clouds on top of the boundary layer (Doms et al., 2007). It is parameterised with Tiedtke (1989), but only for cloud heights smaller than 2 km. Entrainment describes the lateral transport across cloud boundaries via turbulent exchange of mass. The perturbed value of the entrainment rate (Tab. 2.2) is a factor 10 higher than the default value so that the exchange of mass is stimulated. Potentially, the amount of moisture transported out of the boundary layer is enlarged. Hence, the budget of moisture that is potentially available for precipitation forming processes is enlarged as well. Therefore, as a result of a smaller entrainment more precipitation is possible.
In the parameterisation of small-scale turbulence, two parameters are perturbed. This parameterisation links the resolvable scales and the nonresolvable fluctuating scales of motion. Turbulent fluxes may contribute to the exchange of momentum, heat, and humidity between the surface and the atmosphere (Doms et al., 2007). The subgridscale turbulent fluxes are parameterised on basis of the K-theory where they are assumed to be proportional to the diffusion coefficient K of the respective quantity. The main part in the closure of a turbulence parameterisation is the calculation of the diffusion coefficients. For the closure, K is calculated as product of the vertical mixing and the turbulent kinetic energy (TKE). The vertical mixing is derived following Blackadar (1962) depending on the asymptotic mixing length. This value is reduced in COSMO-DE-EPS (fourth perturbed parameter in Tab. 2.2). Therefore, the diffusion coefficient K is reduced as well. A smaller length scale results in

Table 2.2: List of parameter perturbations.

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	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	500

more dissipation. The lower TKE limits the rate of vertical transport leading to larger vertical gradients. This increased local instability enables the formation of convection (personal communication Baldauf).
The subgrid scale cloud cover (second perturbed parameter in Tab. 2.2) is parameterised in dependence of the saturation deficit. A smaller cloud cover reduces the production of TKE. Again, a smaller TKE limits the vertical transport and can therefore, enhance the triggering of convection (Doms et al., 2007).
The development of atmospheric convection is as well sensitive to the surface fluxes of momentum, heat, and moisture (Doms et al., 2007). The surface fluxes affect the atmospherics part of the model as lower boundary condition and are the coupling between the soil and the atmospheric part of the model. The scaling factor for the laminar sublayers is perturbed two times in COSMO-DE-EPS (third perturbed parameter in Tab. 2.2): it is enlarged about a factor 50 and decreased with a factor 10.

From COSMO-DE-EPS, the fields of synthetic radar reflectivity at the 850 hPa pressure surface are used to calculate probabilistic forecasts $P_{EPS}(x, \mathcal{L})$ 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 assuming a Rayleigh relationship (Seifert and Beheng, 2006). An example of the fields is displayed in Fig. 2.7.

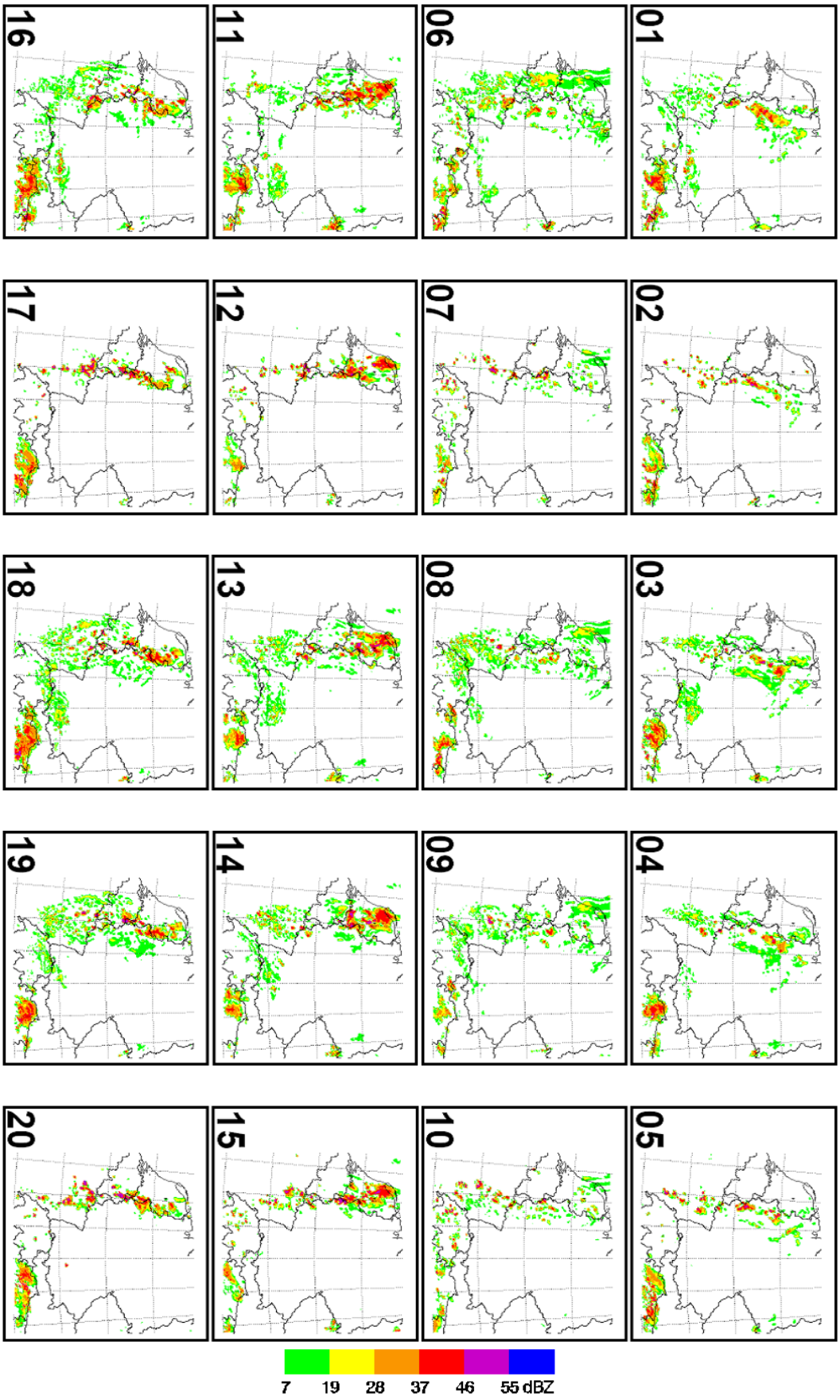

Figure 2.7: COSMO-DE-EPS fields of synthetic radar reflectivity in 850 hPa for 12 August 2007, 23:15 UTC for each single member (1-20).

2.3.2 Probabilistic Forecasts with COSMO-DE-EPS

In agreement with Schwartz et al. (2010), three different approaches are applied on the ensemble to calculate probabilistic forecasts of exceeding a precipitation threshold based on the synthetic reflectivity fields (Fig. 2.4).
First, as traditionally applied on ensembles, at every grid point the **fraction** of members with values above the threshold ($\mathcal{L} = 19\ dBZ$) is determined (Fig. 2.8, fraction). These probabilities depend on the number of ensemble members. Schwartz et al. (2010) named this method 'traditional ensemble probability'. Here, it will be called fraction method.
Second, every member is treated as a deterministic solution and a method similar to the **neighbourhood method** (Theis et al., 2005) and to Local Lagrangian (Chapter 2.2.2) is applied (Fig. 2.8, 1 denoting member 1 as representative of the ensemble). Theis et al. (2005) introduced a pragmatic approach to derive probabilistic precipitation forecasts from a deterministic model by looking into the spatio-temporal neighbourhood at each grid point. This method is in this study only applied in the spatial sense (comparable to Megenhardt et al. (2004)). This means, as in the probabilistic module of Rad-TRAM, the fraction of pixels above the threshold in a search area around each grid point is determined. In contrast to Rad-TRAM, the size of the search area is not increased with lead time. It is fixed at a side length of 75 km. This equals the search area of Rad-TRAM at a lead time of 75 min. Sensitivity studies with 15 km and 120 km showed that the differences between forecasts decrease with increasing side length. Therefore, and in order to produce sharp probabilities, 75 km is chosen. The size of the search area is almost equal to the medium size of 84 km chosen by Theis et al. (2005). In contrast to nowcasting methods, there is no clear correlation of the spatial variability to the motion of the precipitation field. Now, the latter is determined by the prognostic dynamic equations. In fact, the spatial variability around each grid point considers the uncertainty coming from timing and location errors. This uncertainty is in a first approximation assumed to be independent of lead time and therefore, the size of the search area is constant. The application of the neighbourhood method results in 20 probabilistic forecasts.
In the third approach, the **mean** of the probabilities derived with the neighbourhood method is calculated. In Schwartz et al. (2010), this approach is referred to as 'neighbourhood ensemble probability', here as mean method.

Altogether, 22 different probabilistic forecasts are available at each forecast time. Both the generation of COSMO-DE-EPS as well as the analysis consider three sources of uncertainty:

- the spatial variability around each grid point and, implicitly, timing errors,
- the imperfectness of model physics,
- the variability of the lateral boundary conditions.

The method providing the mean of the neighbourhood probabilities considers all of them.

Figure 2.8 illustrates examples of the three approaches applied on COSMO-DE-EPS forecasts for 12 August 2007, 23:15 UTC. For the neighbourhood method, only member 1 has been chosen as representative of the ensemble. All forecasts predict a probability of precipitation larger than zero in the area where the front was observed at 23:15 UTC (cf. Fig. 2.2). Location and intensity of the probability fields differ between the methods. The fraction

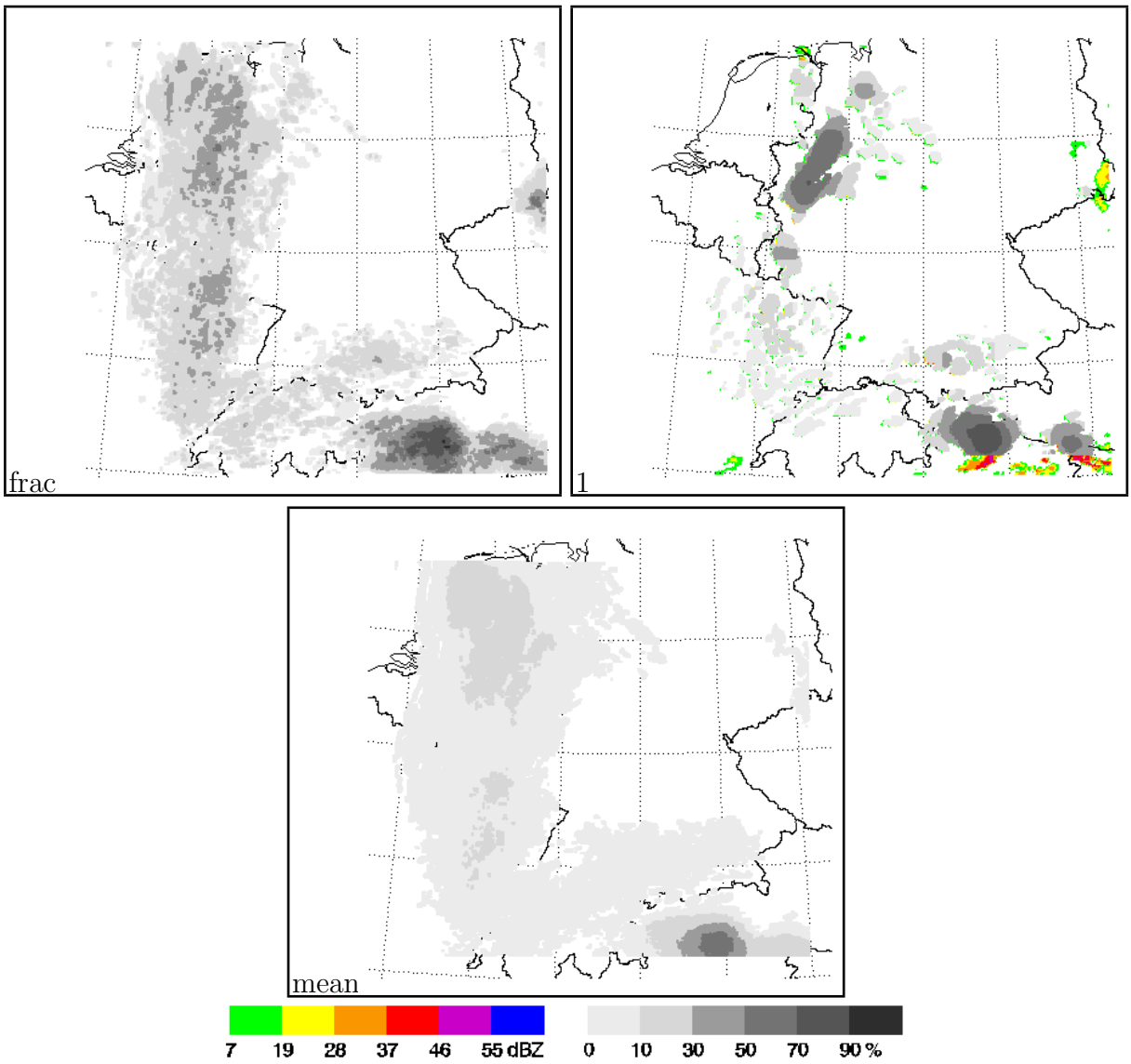

Figure 2.8: Probabilistic COSMO-DE-EPS forecasts for 12 August 2007, 23:15 UTC for the fraction (frac), member 1 as representative of the ensemble, and the mean of the neighbourhood members (mean)(grey-shaded). In the background of member 1, colour-coded the synthetic radar reflectivities in 850 hPa.

method predicts a large and broad probability field (Fig. 2.8, frac). Embedded in this spatially coherent field are spotted probability maxima. In contrast, the probability field of member 1 (Fig. 2.8, 1) covers small areas with isolated probability maxima. In comparison to the fraction method and member 1, the mean of the 20 neighbourhood probabilities is a smooth field with low probability values (Fig. 2.8, mean). The variability of size and spatial distribution of the probability fields in the different forecasts is in a reasonable range given the meteorological situation. The large areas of the fraction method and the mean method reflect the high variability in the solutions of the ensemble. The probabilistic forecasts of the mean and fraction method are very similar concerning location and size but the mean field is smoother with lower probability values.

2.4 Quality of Probability forecasts of discrete predictands

2.4.1 Aspects of quality

To assess the quality of probabilistic forecasts, the term 'goodness' has to be defined for probabilistic forecasts in general. It is distinguished between three types of forecast goodness: consistency, value, and quality (Stanski et al., 1989; Murphy, 1993). Consistency is the degree to which the forecast corresponds to the forecaster's best judgement about the situation, based upon his knowledge. Value is the benefit the forecaster gained through the use of the forecast. Quality is the degree to which the forecast corresponds to what actually happened. Murphy (1993) described nine different aspects (attributes) that contribute to the quality of forecasts.

Probabilistic forecasts mean a probability of an event occurring is forecasted with a value ranging between 0 and 1. They are verified against observations in which this event either occurred (o_j=1) or not (o_j=0). An accurate probability forecast system basically needs three attributes: reliability, resolution and sharpness (Wilks, 2006). Reliability is the agreement between forecast probability and mean observed frequency. Resolution describes the ability of the forecast to resolve the set of sample events into subsets with different outcomes. Sharpness is the tendency of the forecast system to describe probabilities near 0 or 1.

2.4.2 Quality measures

The aspects of forecast quality described above are calculated for a forecast system with different quality measures (Wilks, 2006). Relevant information for the verification of probabilistic forecasts of discrete predictands is contained in the joint distribution of forecasts and observations. The simplest setting is that dichotomous predictands (occurrence o_1 - nonoccurrence o_2) are combined with $i = 11$ possible forecasts y_i ranging from $y_1 = 0.0$ to $y_{11} = 1.0$ so that finally 22 probabilities $p(y_i, o_j)$ with $j = 1, 2$ are available.

It is possible to describe the joint distribution with factorisations containing a conditional and a marginal distribution (Murphy and Winkler, 1987). In the calibration-refinement factorisation, this involves the conditional distribution of the observations given the forecasts and the marginal distribution of the forecasts and can be written as

$$p(y_i, o_j) = p(o_j|y_i)p(y_i), \tag{2.10}$$

with the conditional probability $p(o_j|y_i)$ that the event o_j occurred given the forecast y_i. The marginal distribution of the forecasts (also named refinement) $p(y_i)$ consists of the relative frequency of the event in each probability bin i and can be calculated with

$$p(y_i) = \frac{N_i}{n}, \tag{2.11}$$

where n is the total number of forecast-event pairs and N_i the number of times each forecast y_i is used. n is the sum of the subsample N_i

$$n = \sum_{i=1}^{I} N_i. \tag{2.12}$$

The relative frequency is the sum of the two conditional probabilities

$$p(y_i) = p(y_i, o_1) + p(y_i, o_2), \tag{2.13}$$

with $\sum_j p(y_j) = 1$.

Brier Score and its algebraic decomposition

The most common scalar score for the verification of probabilistic forecasts is the Brier Score (BS) (Brier, 1950)

$$BS := \frac{1}{n}(y_k - o_k)^2, \tag{2.14}$$

where k denotes a numbering of n forecast - event pairs, y_k the forecast probability and o_k the subsequent observations. The Brier Score[6] is the mean squared error of the forecast and negatively oriented. Its values range between [0,1].

The algebraic decomposition of the Brier Score relates to the calibration refinement factorisation of the joint distribution (Wilks, 2006) and was introduced by Murphy (1973). The calculation of the relative frequencies of the forecasts (Eq. 2.10), the relative frequencies of occurrence in each subsample $\bar{o_i}$

$$\bar{o_i} = p(o_1|y_i) = \frac{1}{N_i} \sum_{k \epsilon N_i} o_k \tag{2.15}$$

and the overall frequency of the event $\bar{o}$,

$$\bar{o} = \frac{1}{n} \sum_{k=1}^{I} o_k = \frac{1}{n} \sum_{i=1}^{I} N_i \bar{o_i} \tag{2.16}$$

results in a new formulation of the Brier Score:

$$BS = \frac{1}{n} \sum_{i=1}^{I} N_i (y_i - \bar{o_i})^2 - \frac{1}{n} \sum_{i=1}^{I} N_i (\bar{o_i} - \bar{o})^2 + \bar{o}(1 - \bar{o}). \tag{2.17}$$

The three different terms in the decomposition are known as reliability, resolution and uncertainty. Reliability measures the ability of the system to forecast accurate probabilities. In other words, it summarises the calibration or the conditional bias of the forecast. Therefore, it should ideally be small. Resolution indicates the ability of the forecast system to correctly separate the different categories, whatever the forecast probability. You could also say, that resolution summarises the ability of the forecasts to discern different relative frequencies of the event, independent if they are right or wrong in the sense of reliability. This term is negative in the sum and should be large in a forecast with skill in resolution. The uncertainty term is independent of the forecast system and depends only on the variability of the observations. Uncertainty describes the intrinsic difficulty of forecasting the event during the evaluation period.

[6] It is common use to name this score Brier Score, but actually it is comparable to the method of least squares first introduced by C.F. Gauß (1777-1855). Therefore, the score could as well be named Gauß Score.

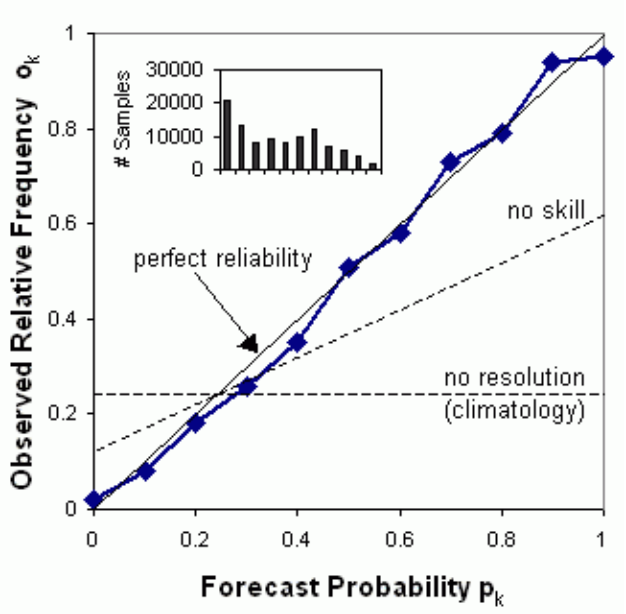

Figure 2.9: Example for Reliability diagram with the calibration function (blue) and the refinement distributions (histogram) (http://www.cawcr.gov.au/projects/verification).

Reliability diagram

Reliability diagrams show the full distributions of forecasts and observations in terms of the calibration refinement distribution (Eq. 2.10). Therefore, they can be understood as a graphical representation of the decomposed Brier score (Eq. 2.17). Hence, a reliability diagram consists of two elements: the calibration function and the refinement distributions (Fig. 2.9). The calibration function shows the distribution of observations given each of the allowable values of the forecast $p(o_1|y_i)$. The refinement distribution expresses the frequency of use of each of the possible forecasts y_i (Fig. 2.9, histogram in the upper left corner). In well calibrated forecasts, $p(o_1|y_i)$ equals the probability category y_i (high reliability) and is therefore, near the 'perfect reliability' diagonal. Normally, a line called 'no resolution' that is the sample climatology $\bar{o}$ and a 'no skill' line is included in the graph. The 'no skill' line indicates that on these points, the reliability component of the Brier score exactly matches the resolution component which means skill goes to 0. If these two lines are included, the reliability diagram may also called attributes diagram. The resolution component of the Brier score (Eq. 2.17) can be identified as the weighted squared difference between the 'no resolution' line and the calibration probabilities $p(o_1|y_i)$. Whereas the reliability component of the Brier score (Eq. 2.17) is the weighted squared difference between the 'no reliability' diagonal and the calibration probabilities $p(o_1|y_i)$. The histogram reflects the confidence of the forecasts and gives information about their sharpness. If the extreme values are frequently used the forecasts are sharp and have high confidence.

The example in Fig. 2.9 shows a forecast with high reliability as the distance between the 'perfect reliability' line and the calibration function is small. The forecasts have high skill concerning the aspect of resolution as well seen in the large difference between the calibration function and the 'no resolution' line. Summarising resolution and reliability to more general skill, it is seen that the forecast as well has skill as the calibration function is found within the 'no skill' and the 'perfect reliability' line. From the histogram it is learnt that the forecasts are not sharp although the $y_1 = 0.0$ bin is highly populated as the $y_{11} = 1.0$ is

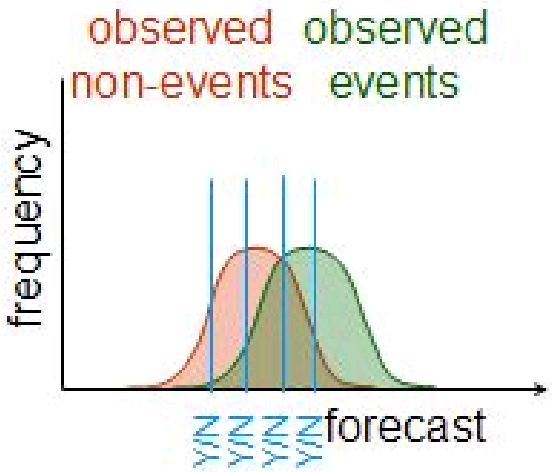

Figure 2.10: Concept of discrimination in a likelihood diagram (http://euromet.meteo.fr/resources/ukmeteocal/verification).

only sparsely used.

Relative Operating Characteristics (ROC)

In comparison to reliability diagrams, Relative Operating Characteristics (sometimes also referred to as Receiver Operating Characteristics, both abbreviated as ROC; e.g. Stanski et al. (1989)) contain not the entire information of the joint distribution of forecasts and observations. They are based on the likelihood base factorisation. Therefore, the joint distribution is described with conditional probability $p(y_i|o_j)$ of the forecast y_i given the observation o_j. Since there are only two possible observations o_j (event occurred and the event did not occur), the sample is divided into two groups. The distribution of the probability forecast values for each group can be plotted and compared (Fig. 2.10).

The graph of the two distributions is sometimes called a likelihood diagram. Higher forecast probabilities should be associated with occurrences of the event and lower forecast probabilities with non-occurrences. If so, then the two conditional distributions are well-separated with minimum overlap. The separation of these distributions is a measure of discrimination.

Contingency tables (Tab. 2.3) can be constructed (one for each probability threshold) by counting probabilities higher than the threshold as event (indicated with blue lines in Fig. 2.10).

Table 2.3: Contingency table.

		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

Finally, ROC curves are constructed by evaluating from each of the $I-1$ contingency tables the hit rate H

$$H = \frac{a}{a+c} \tag{2.18}$$

and the false alarm rate F

$$F = \frac{b}{b+d}. \tag{2.19}$$

The connection line of the point pairs (H, F) is plotted after adding the point (0,0) for the case 'never forecast' and the point (1,1) for the case 'always forecast' (Fig. 2.11).

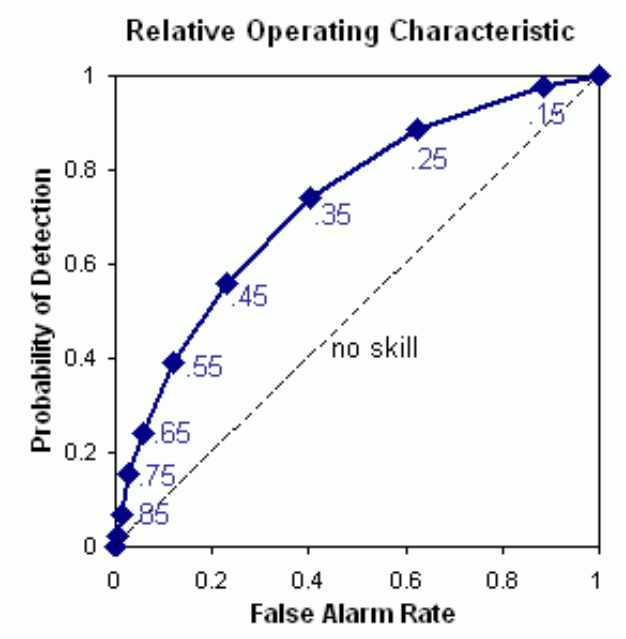

Figure 2.11: Example for Relative Operating Characteristics (http://www.cawcr.gov.au/projects/verification).

In a perfect forecast, the ROC curve travels from the lower left corner to the upper left and to the upper right, so that the area under the curve is nearly 1. An useless forecast is the diagonal where the hit rate equals the false alarm rate. This would mean that no discrimination between occurrence and non occurrence of the event can be made by the forecasts. Then, area under the curve would be 0.5 ('no skill' line in Fig. 2.11). The ROC curves are insensitive to conditional or unconditional bias and therefore, reflect the potential skill which could be achieved if the forecasts were perfectly calibrated. As they make the contrary assumption (conditioned the event was observed, what was forecast), they are a good enhancement for Reliability diagrams or the decomposed Brier score. In the following, only the areas under the ROC curve are evaluated.

Conditional square root of Ranked Probability Score (CSRR)

Another scalar score is the conditional square root of ranked probability score ($CSRR$) that was suggested by Germann and Zawadzki (2004). The $CSRR$ is defined for multicategorial forecasts and is basically the mean squared error of the probabilistic forecast P and the binary observation $\hat{P}$. In contrast to the Brier score (Eq. 2.14), this difference is weighted with $\tilde{\Omega}_{t_0+\tau}$, the size of the rain domain ($\mathcal{L} > 0$ dBZ) and the square root of this value is taken

$$CSRR(\tau) = \left\{ \frac{1}{\tilde{\Omega}_{t_0+\tau}(\mathcal{L}_{max} - \mathcal{L}_{min})} \int_{\Omega} \int_{\mathcal{L}_{min}}^{\mathcal{L}_{max}} [P(t_0+\tau, x, \mathcal{L}) - \hat{P}(t_0+\tau, x, \mathcal{L})]^2 d\mathcal{L} dx \right\}^{0.5}, \tag{2.20}$$

where Ω is the space domain, $\mathcal{L}$ is the threshold and $\mathcal{L}_{min}$ and $\mathcal{L}_{max}$ are the respective maximum and the minimum thresholds applied.

As in this study only one threshold ($\mathcal{L} = 19\ dBZ$) is applied, the $CSRR$ simplifies to

$$CSRR(\tau) = \left\{ \frac{1}{\tilde{\Omega}_{t_0+\tau}} \int_{\Omega} [P(t_0+\tau, x, \mathcal{L}) - \hat{P}(t_0+\tau, x, \mathcal{L})]^2 dx \right\}^{0.5}. \tag{2.21}$$

The advantage of the $CSRR$ in comparison to the Brier score is that different cases can be compared more fairly. The Brier score is very sensitive to the correct negatives and especially in case of rare events, low scores pretend a very good performance.

In this work, mainly the following quality measures are used: the Brier score, the CSRR, and the area under the ROC curve. They are all scalar measures and therefore, they can be displayed efficiently as well for a low number of events in case studies as over a longer period in time series. The three scores are chosen as they show different aspects of quality of the forecasts. The Brier score and CSRR measure the mean error, the Brier score in general and the CSRR relative to the number of observed events at the predicted time step. Furthermore, the Brier score is able to give information about the reliability and the resolution of the forecasts if compared to the uncertainty component of its decomposition. The area under the ROC curve summarises the ability of the forecasts to discriminate between the occurrence and the non-occurrence of the event and is hence a good companion to the Brier score.

2.5 Calibration of COSMO-DE-EPS forecasts

Calibration of probabilistic NWP forecasts is necessary as the observed frequency of an event normally differs from the relative frequency in the forecast. Therefore, calibration refers to the statistical correction of numerical forecasts to produce calibrated probabilistic forecasts that are still as sharp as possible while remaining reliable (Hamill et al., 2008). Calibrating probabilistic precipitation forecasts is a special challenge as there are many zero events in strongly intermittent fields. Several techniques like linear or logistic regression (Hamill and Colucci, 1998; Hamill et al., 2008) or Bayesian model averaging (Raftery et al., 2005) exist

to calibrate precipitation forecasts.

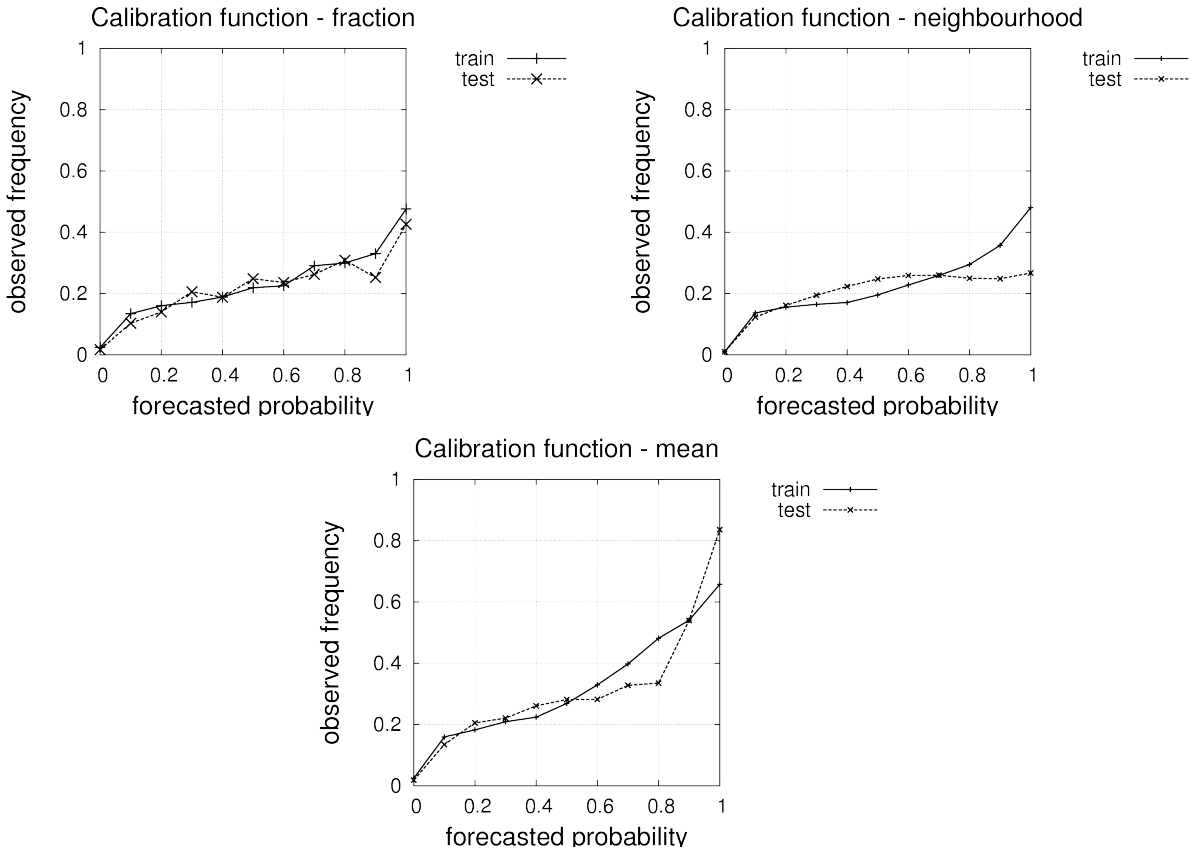

Figure 2.12: Calibration functions for the three different methods (fraction, neighbourhood, and mean): comparison of functions derived from testing and training data set.

In this study, the probabilistic forecasts derived from COSMO-DE-EPS are calibrated with the reliability diagram statistics method (Zhu et al., 1996). The method suggests that when the probability category i is forecast in the testing subsample, the calibrated probability is the frequency with which the event is observed in the training subsample when the sample forecast category i is forecast. The forecast categories chosen in this study contain $i = 11$ bins from 0.0 to 1.0 probability values. For the calibration, the available data is divided into a training and a testing data set. The training period comprises the 08, 09, 10, and 11 of August (around 8 mio grid points per member) and the testing set the 12, 13, 14, 15, and 16 of August (around 10 mio grid points). For all three methods, the first three hours of each run are not included as the spread between the members is small. For the calibration of the neighbourhood probabilities, all members are calibrated together (total around 370 mio grid points). The fraction and mean method based probabilities of course cover 1/20 of the values.

Figure 2.12 shows the calibration functions for the three methods applied on the ensemble output. For each method, the calibration functions derived from the testing and from the training period are displayed. All three calibration functions are robust in the sense, that there are no large differences between the functions based on the testing and training period. But the differences between the three methods are small as well. Generally, the observed frequencies are smaller than the forecasted probabilities. Only for high bins, differences can

be seen. There, the mean method shows a higher observed frequency than the fraction and the neighbourhood method. Nevertheless, in all three methods, the calibration results in a reduction of the probabilities to lower values.

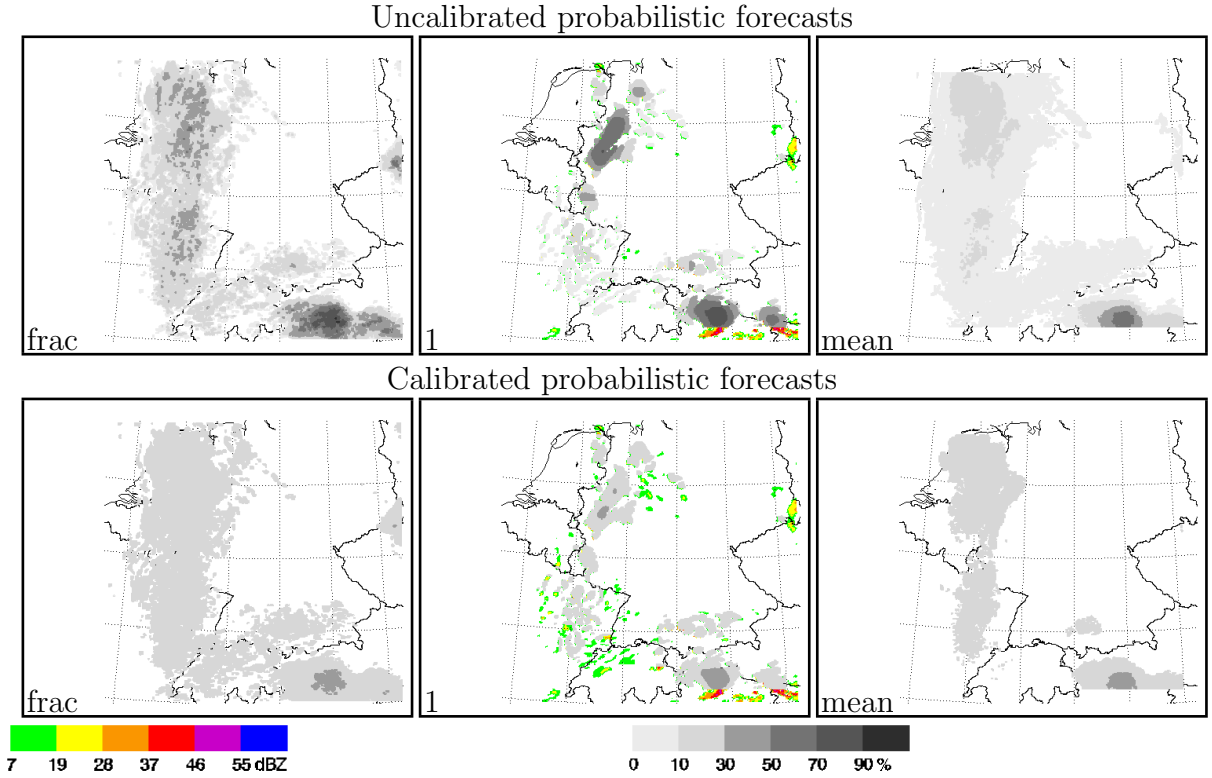

Figure 2.13: Effect of calibration on COSMO-DE-EPS forecasts (grey-shaded) on 12 August 2007, 23:15 UTC for the fraction method (frac), member 1, and the mean of the neighbourhood members (mean). In the top row, the uncalibrated probabilistic forecasts and in the bottom row, the calibrated. In the background of member 1, colour-coded the synthetic radar reflectivities in 850 hPa.

Figure 2.13 displays the effect of the calibration on the probability fields on 12 August 2007, 23:15 UTC of the three methods fraction, neighbourhood, and mean. Member 1 is chosen as representative of the 20 neighbourhood members as on all the same calibration function is applied. It is seen in all methods that the maximum values are lower after the calibration. The maximum probabilities on this day were after calibration lower than 30 % whereas the fraction method before calibration even predicted 90 % over the Alps. The position of the probabilities is not changed through calibration. The differences in the mean method are due to the plotting thresholds. The effect of calibration can be summarised is a reduction of the amplitude of the probabilities. Therefore, sharpness is reduced.

The reliability component of the decomposed Brier score (Eq. 2.17) can be used as measure for a successful calibration (Atger, 2003). In this study, only the domain reliability is calculated due to the limited period of forecasts. Tab. 2.4 shows the mean reliability component and their standard deviation for the three methods, respectively before and after calibration. The calibration is successful as both the mean and the standard deviation are reduced for the neighbourhood probabilities, the fraction, and the mean method.

Table 2.4: Reliability component of Brier score, mean and standard deviation. All gridpoints are considered together. For the neighbourhood members 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}

Chapter 3

Quality of probabilistic forecasts - Selected case studies

The previous chapter focused on the calculation of probabilistic forecasts based on observations and NWP. In addition, quality measures to calculate the quality of probabilistic forecasts were introduced to quantify the value of the forecasts. In this chapter, these quality measures will be applied on the probabilistic forecasts of Rad-TRAM and COSMO-DE-EPS and their skill will be evaluated in three case studies representing different meteorological regimes. The skill of forecasts based on Rad-TRAM and COSMO-DE-EPS is evaluated considering:

- the development of skill with time in time series of the Brier score, the CSRR, and the area under the ROC curve
- the development of skill with lead time with Brier score, CSRR, and area under the ROC curve with consideration of the effect of calibration on the model forecasts.

In the evaluation of the development of skill with lead time, the different set ups of Rad-TRAM and COSMO-DE-EPS have to be considered (Fig. 3.1). COSMO-DE-EPS is started at 0:00 UTC and runs for 24 hours. In 30 minutes intervals, fields of synthetic radar reflectivity are available, starting at 0:15 UTC. In contrast, Rad-TRAM is started beginning at 0:00 UTC every 15 minutes with a lead time of eight hours and available forecasts every 15 minutes. To unify the evaluation, some adoptions are made. Rad-TRAM is evaluated every two hours beginning at 2:00 UTC (2, 4, 6, 8, 10, 12, 14, 16 UTC) for 8 hours lead time every day (Fig. 3.1 for one day). Within the resulting eight time frames, the mean skill is calculated for same lead times. As this evaluation is not possible for the COSMO-DE-EPS forecasts, the mean model skill within the respective eight Rad-TRAM time frames is calculated. Due to the coarser temporal resolution of COSMO-DE-EPS, Rad-TRAM is only analysed at times when COSMO-DE-EPS forecasts are available. Therefore, probabilistic forecasts are analysed for lead times of 15 to 465 minutes in 30 minutes increments. Finally, the mean and the standard deviation of the respective quality measures is calculated for single days as well for the forecasts based on Rad-TRAM and on COSMO-DE-EPS.

Furthermore, an eyeball investigation of the probabilistic forecasts under consideration of the meteorological situation will be deduced.

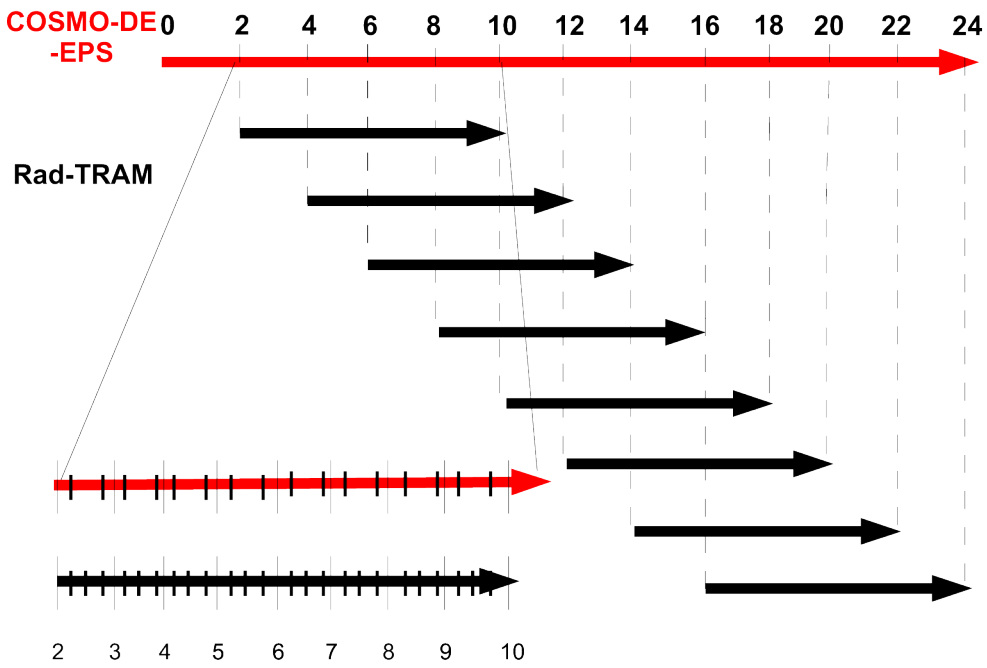

Figure 3.1: Schematic overview over the different setups and the lead time dependent evaluation for one day.

Three specific days within the period are chosen and will be discussed in detail to describe the forecast skill of Rad-TRAM and COSMO-DE-EPS in different meteorological situations with convective precipitation. Within the investigated period, three days were especially interesting as they differed in the meteorological situation causing the precipitation and the nature and amount of precipitation. These three days are simultaneously intensive observation periods (IOPs) of the COPS campaign[1] (Wulfmeyer et al., 2008). This campaign was conducted over southern Germany and north-eastern France in summer 2007 and covered a small part of the evaluation domain in this study (compare Fig. 2.2).
The 9 August 2007 (IOP 14) was characterised by large scale ascent in a strong easterly flow. The amount of precipitation was high over the entire day. Due to the continuously large amount of precipitation, it is expected that the nowcasting method is superior to the model forecasts even for long lead times. On 12 August (IOP 15), some small convective cells developed under the influence of high pressure. In the late afternoon, the regime changed and a cold front with weak embedded convection moved into the evaluation domain from the west. In this case study, it is expected that the nowcaster has high skill during the passage of the cold front in the afternoon and evening as this phenomena can well be described by advection. Whereas the regime change and the small convective cells should be poorly represented. As the amount of precipitation on this day is low, the total error of the forecasts should be small as well. In the third case study, 15 August 2007 (IOP 16), a complete frontal system passed the domain, with a warm front and a cold front. Due to the several phases that occur within a frontal system, the overall skill of both forecasts is expected to be small, because the exact timing and velocity of the system is very important but a large source of

[1] Convective and Orographically-induced Precipitation Study

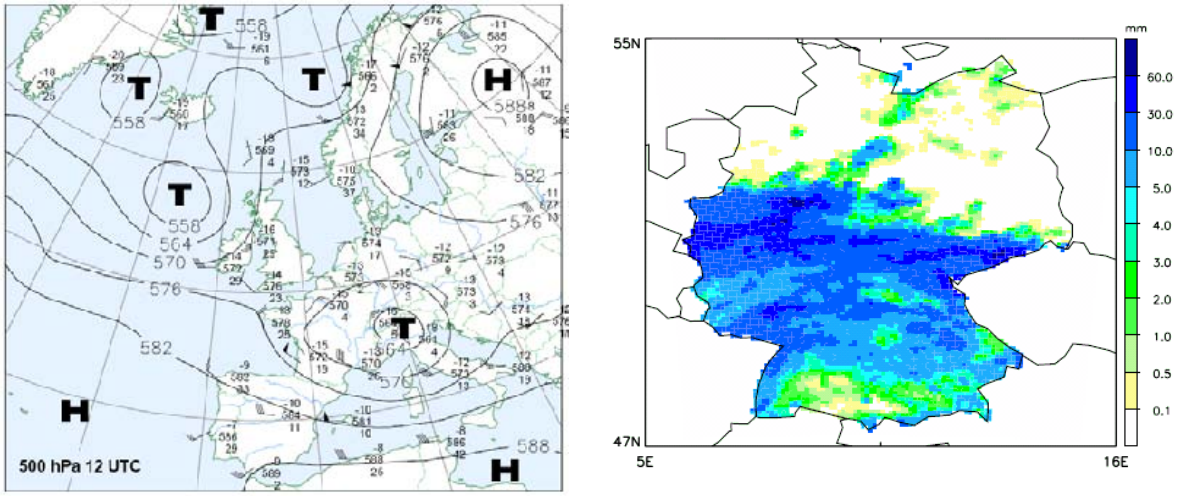

Figure 3.2: Synoptic overview: 500 hPa on 9 August 2007, 12 UTC and disaggregated daily precipitation sum over Germany (Zimmer and Wernli (2008)).

error. Nevertheless, the nowcaster should be superior to the model forecasts during specific phases but poor if the phases change.
Generally, it is expected that Rad-TRAM is superior to COSMO-DE-EPS for short lead times. But the rate of decrease and the general skill of forecasting methods should depend on the meteorological situation. To compare the different cases more easily, the analysis of the case studies will be carried out in the same way. First, a short description of the synoptic situation will be given. The quality of the probabilistic forecast will be evaluated as explained above. The description of the respective case studies will close with a discussion.

3.1 IOP 14: 9 August 2007 (Synoptic scale ascent)

3.1.1 Synoptic overview

On 9 August 2007, a low pressure system in 500 hPa moved eastward and became stationary over northern Italy (Fig. 3.2, left). A cut-off low developed over the Alps. The frontal system divided Germany into a warm North and a cold South. The westerly flow around the low over the Alps led to heavy precipitation (Fig. 3.2, right) in almost the entire evaluation domain, except northern Germany. The observed radar reflectivities show over the entire day a large amount of wide spread precipitation with embedded heavy precipitation cells (Fig. 3.3). Several intense isolated cells were interrupted by phases without precipitation. In the night and morning hours, a very intensive precipitation region over the Alps (Fig. 3.3, 3:00 UTC) and later north eastern France (Fig. 3.3, 8:00 UTC) caused heavy precipitation and flooding. Around noon, precipitation decreased over southern Germany before a new intensive cell moved into the evaluation domain from the Czech Republic (Fig. 3.3, 13:30 UTC). These cells moved over the middle of Germany the following hours (Fig. 3.3, 20:15 UTC) and caused heavy precipitation there.

The uncertainty component of the Brier score only depends on the observations (sec. 2.4.2). Therefore, it is a suitable measure to give an overview over the investigated period with the overall frequency of the event (here: reaching 19 dBZ in radar observations). Uncertainty

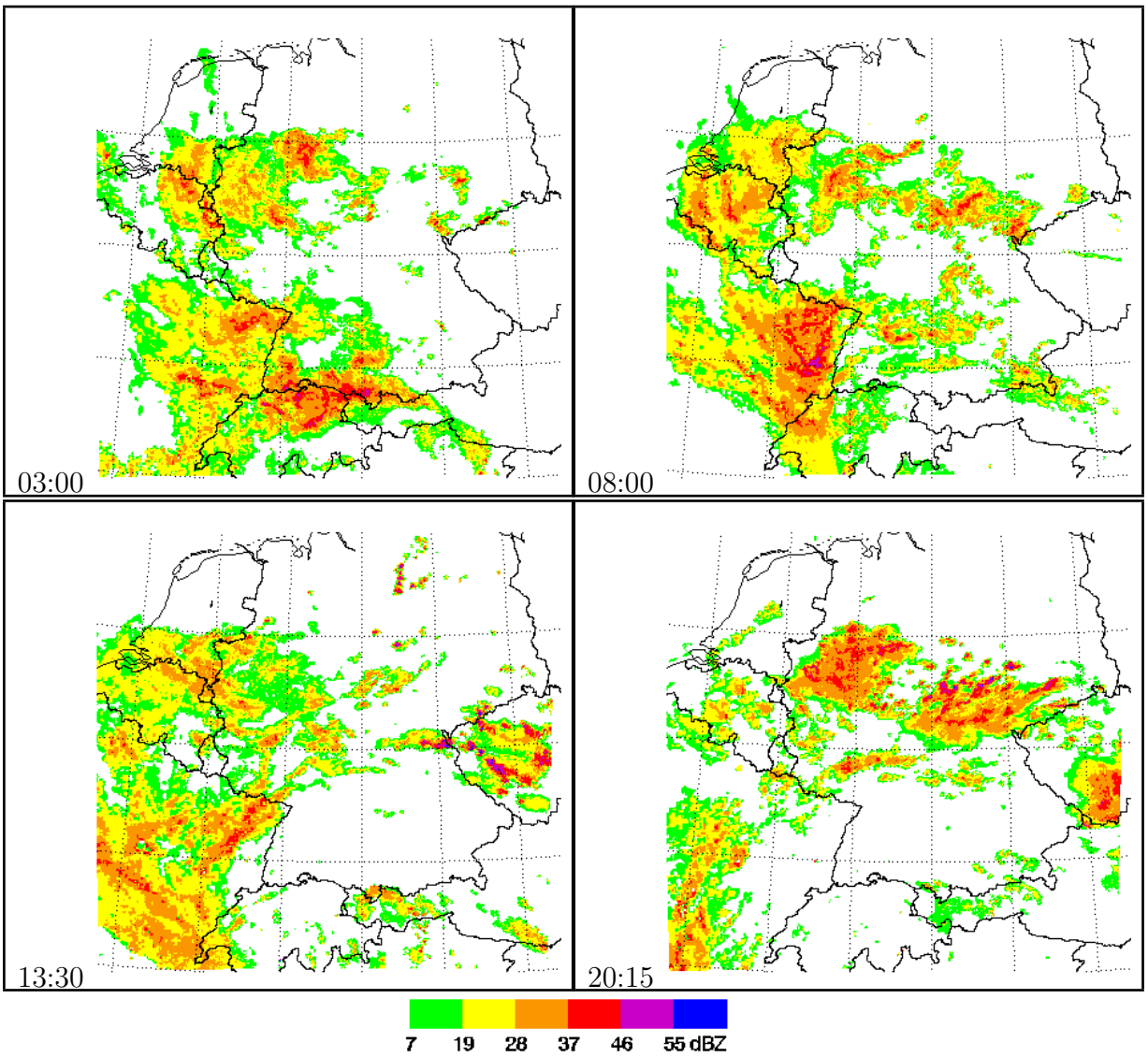

Figure 3.3: Observed radar reflectivity on 9 August 2007, 03:00 UTC, 08:00 UTC, 13:30 UTC, and 20:15 UTC.

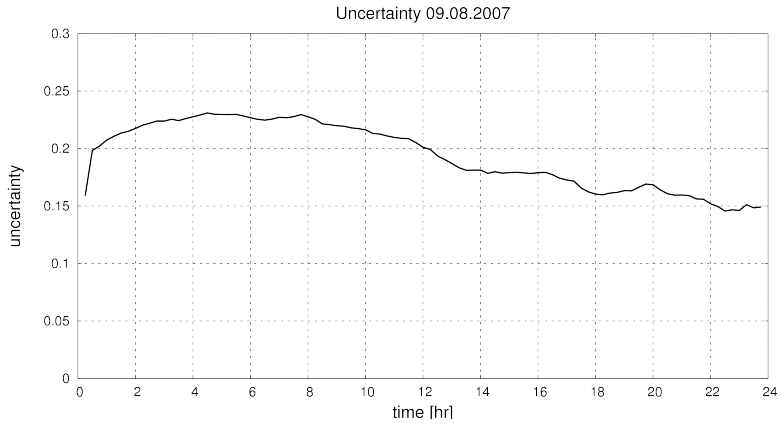

Figure 3.4: Uncertainty component of the Brier score on 9 August 2007.

can be understood as the variance of the observed frequency of reaching 19 dBZ in radar reflectivities. It measures the inherent uncertainty in the event. For binary events, it is at a maximum when the event occurs 50 % of the time and the uncertainty is zero if the event always or never occurs. Uncertainty on 9 August 2007 (Fig. 3.4) is relatively high over the entire day (compare to Fig. 4.1). This means, most of the time, half of the domain is covered with radar reflectivities of 19 dBZ or more. Only at the end of the day, uncertainty decreases. This could theoretically mean that the event happens more or less often. On this particular day, the precipitation in the domain decreased as Fig. 3.3 shows.

3.1.2 Quality of probabilistic forecasts

Figure 3.5 is an example of the probabilistic forecasts based on Rad-TRAM for 20:15 UTC. The observation at this time is displayed in Fig. 3.3. The 15 minutes nowcast (Fig. 3.5a) represents very well the two large convective cells over the middle of Germany with high probabilities. But also the various small less intensive cells are reproduced with high and sharp probabilities. The forecast based on the observation one hour ago is significantly smoother than the 15 minutes nowcast. The two large convective cells are reproduced (Fig. 3.5b). The smaller cells observed are forecasted with a lower probability, but at the right location. In the two hour forecast (Fig. 3.5c) the probabilities of reaching the threshold of 19 dBZ further decreased. The general situation with the two intensive cells is still captured by the forecast, but the small scale structure of the two cells cannot be described. Also the small scale development, for example over southern Germany or Austria, cannot be described in detail. The presence of low probabilities in this area show the best possible skill for this lead time.

Calibrated probabilistic forecasts derived from COSMO-DE-EPS for 20:15 UTC are displayed in Fig. 3.6. For clarity, only three of the 22 forecasts based on COSMO-DE-EPS are selected: the fraction method, member 1 as representative of the neighbourhood members, and the mean method. The fraction method has a large probability field covering large parts of the southern domain. The separation of the dry north and the wet south is reproduced by

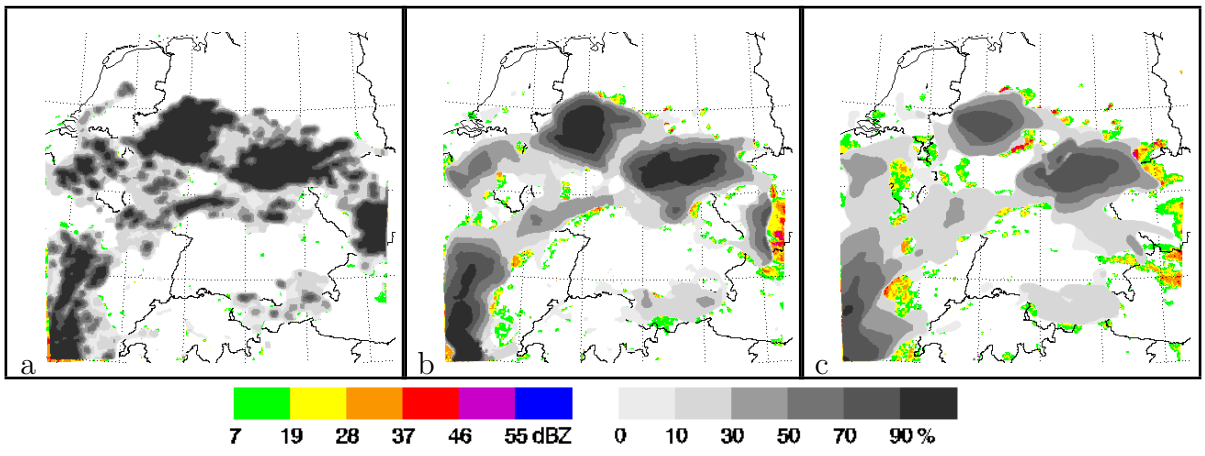

Figure 3.5: Probabilistic forecasts of Rad-TRAM for 9 August 2007, 20:15 UTC based on (a) 15 min forecast from 20:00 UTC, (b) 60 min forecast from 19:15 UTC and (c) 120 min forecast from 18:15 UTC grey-shaded and the reflectivity observations at the respective initial time colour-coded in the background.

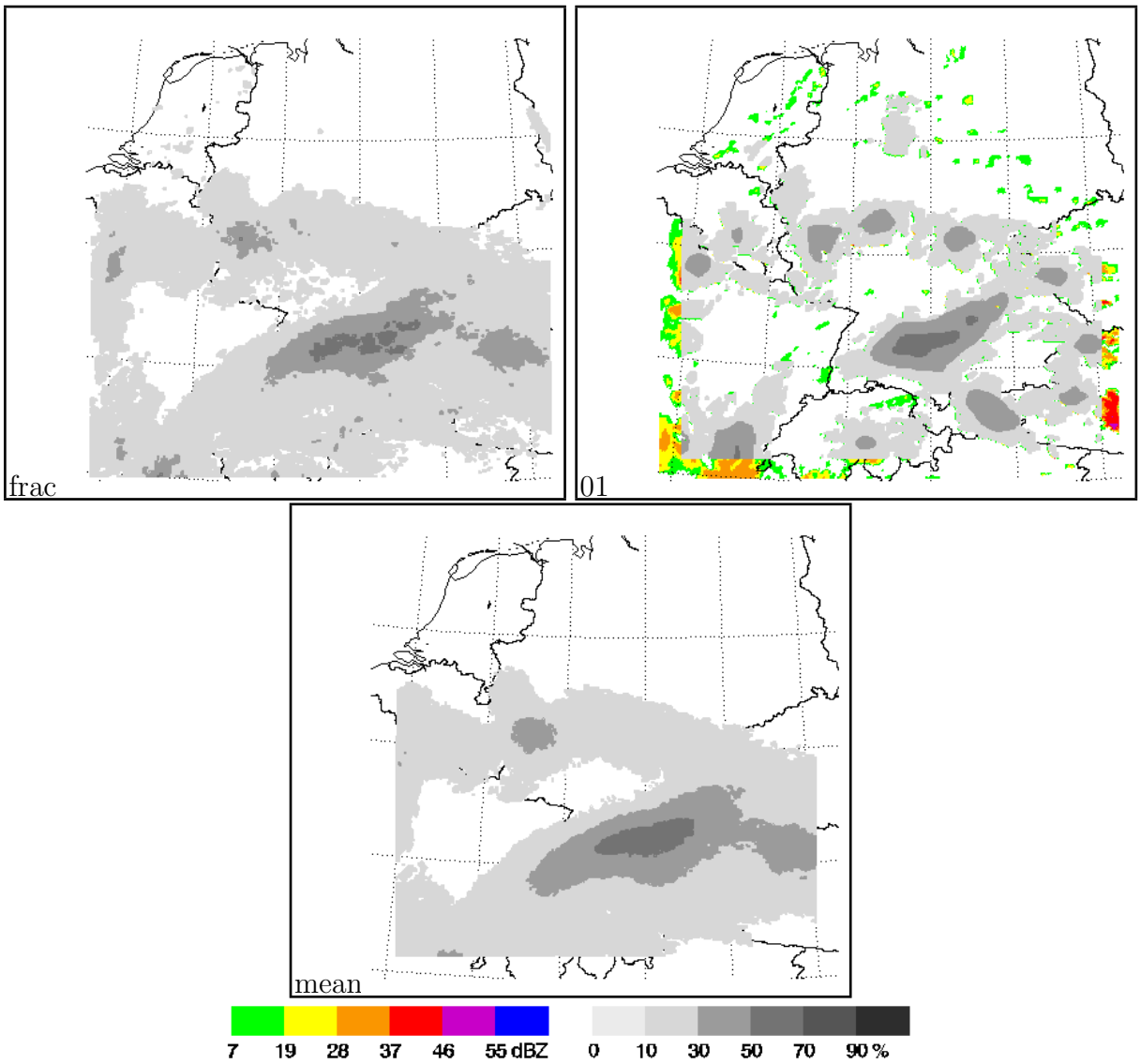

Figure 3.6: Calibrated probabilistic COSMO-DE-EPS forecasts for 09 August 2007, 20:15 UTC (grey-shaded) for the fraction (frac), member 1 as representative for the neighbourhood members, and the mean of the neighbourhood members (mean). In the background of member 1 colour-coded the synthetic radar reflectivities at 850 hPa.

the model. The probabilities are peaked with larger values over southern Germany. In the observations, it is seen that there was no precipitation at this time in the South (Fig. 3.3). The rain here occurred earlier that day. A second area where several members forecasted reflectivities of at least 19 dBZ is found in western Germany. This forecasted probability field also misses the actually observed reflectivity pattern that occurred further north. Member 1 fails as well in capturing the exact location of the precipitation, but shows some skill. The main convective cells over Central Germany are reproduced, but the size is significantly underestimated. The cells over the north of Germany are as well as the cell over southern Germany false alarms. The mean over all neighbourhood members (Fig. 3.6, mean) looks very similar to the fraction method concerning the distribution of the probabilities over the domain but is smoother in amplitude. The general agreement with the dry north and the wet south is seen again, but also the large area over southern Germany and eastern France resulting in false alarms.

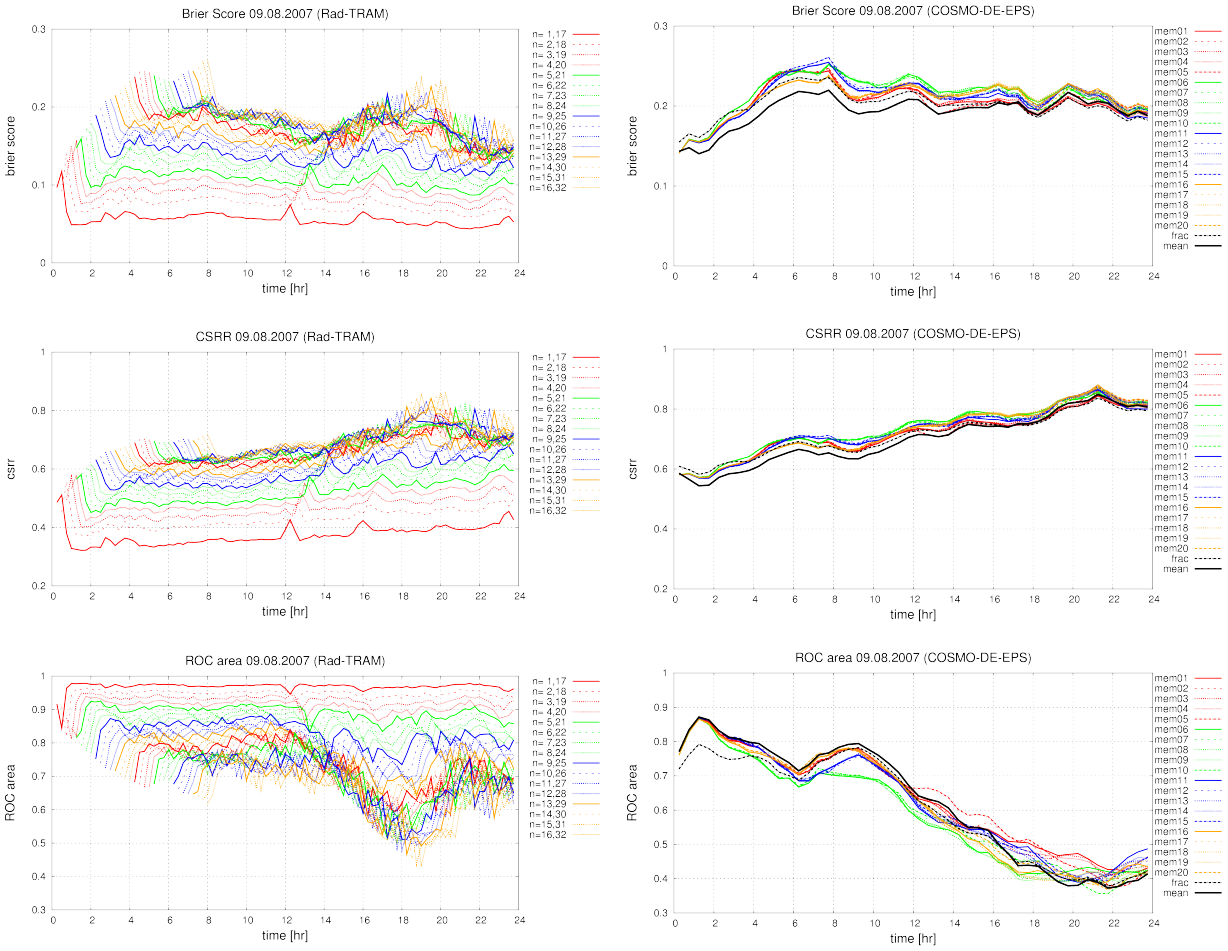

Figure 3.7: Development of Brier Score, CSRR, and area under ROC curve for Rad-TRAM (left) and calibrated COSMO-DE-EPS (right) forecasts from 9 August 2007. Concerning Rad-TRAM, the first forecast hour is displayed in red, the second in green, the third in blue, the fourth in orange, and the next four hours in the same way. The different line styles denote the four lead times within each forecast hour. Concerning COSMO-DE-EPS, the neighbourhood members are colour-coded such that members based on lateral boundary conditions from ECMWF are red, from DWD green, from NCEP blue, and from UKMO orange. The different physical perturbations are displayed with different line styles. The black solid line represents the mean of the neighbourhood members and the black dashed-dotted line is the fraction method.

Fig. 3.7 shows the development of forecast skill over 9 August 2007 for Rad-TRAM (left) and COSMO-DE-EPS (right) with the Brier Score (top), the CSRR (middle), and the area under the ROC curve (bottom). Rad-TRAM is started at 0:00 UTC. Therefore, the first available forecast for each lead time to be evaluated is at 0:00 UTC plus the respective lead time. For example, the first two hour forecast can be evaluated at 2:00 UTC. Furthermore, the quality of the very first forecast of each lead time per day varies from the following. The calculation of the displacement vector field is based on solving the optical flow equation of two consecutive time steps (sec. 2.2). At 0:00 UTC, the previous observation is not available and therefore, displacement vectors to shift the probability fields cannot be calculated.
The quality of Rad-TRAM forecasts within a lead time evaluated with the Brier score varies

hardly over the day (Fig. 3.7, left top). A clear ranking following lead time can be identified. The skill of the forecasts with different lead times decreases steadily with increasing lead time in the first three forecast hours. In the first half of the day, this behaviour is even seen in the first four forecast hours. Longer lead times differ only little from each other. At 12:15 UTC, a peak representing a short but intense decrease in forecast skill can be seen. This peak can be identified in all forecasts based on 12:00 UTC. At this time step there seems to be an error in processing the radar data of the observation. The part of the observed precipitation field located over France remains completely unchanged for one time step whereas the data based on other radars at least changes slightly (not shown).
CSRR (Fig. 3.7, left middle) of Rad-TRAM forecasts generally shows a similar behaviour as the Brier score. The forecasts are ranked following lead time with decreasing differences with increasing lead time. After three or four hours of lead time, the forecasts cannot be distinguished from each other. In comparison to the Brier score, the CSRR shows that forecasts loose skill during the day: at 2:00 UTC CSRR for the 15 minutes nowcast is 0.33 and at 20:00 UTC, it is 0.4. This can be seen for all lead times.
The forecasts based on Rad-TRAM have very high skill concerning discrimination as investigated with the ROC area (Fig. 3.7, left bottom). In the first half of the day, forecasts of all lead times have ROC areas larger than 0.7. The erroneous values based on the very first observation with the above explained shortcoming of the missing displacement vector field are exempted. In the second half of the day, when convection intensifies and gets a more small scale structure, the skill in discrimination decreases for longer lead times. Forecasts longer than four hours hardly have skill during 16:00 and 20:00 UTC. After 20:00 UTC, the skill in discrimination of longer lead times increases again. This is because during 16:00 and 20:00 UTC a new intense precipitation field moves into the domain from the east. The forecasts with longer lead times are based on older observations when this pattern was not observed yet. Therefore, Rad-TRAM cannot predict this pattern.

The Brier score of the calibrated forecasts based on COSMO-DE-EPS (Fig. 3.7, right top) varies over the day. The values are lowest and therefore best at the beginning of the day. Brier score increases to a maximum at 8:00 UTC and decreases over the entire rest of the day. Until 16:00 UTC, the mean method performs better than the other forecasts. The spread of the forecasts decreases over the day and finally they can hardly be distinguished.
The CSRR (Fig. 3.7, right middle) shows in comparison to the Brier score a steadily decrease in forecast skill over the day. The ranking of the members is similar to Brier score: the mean method has higher skill than the other solutions over more than the half of the day. In the morning, the solutions vary around 0.6 and in the night around 0.8.
The ROC area (Fig. 3.7, right bottom) as well as the CSRR shows a decrease in skill over the day. In the morning, the skill in discrimination is high (around 0.85) but at the end of the day the values decrease even below 0.4. The differences between the 22 solutions is the largest of all three scores. The mean method has most skill in the first half of the day and least at the end. The neighbourhood members are ranked following global models. At the beginning, the fraction method is worst, but during the day it is within the neighbourhood members. Generally, the variability between the different methods is much smaller than the variability of the respective score over the course of the day. All scores show for the COSMO-DE-EPS a decrease in skill over the day. Also the ranking is if identifiable most of the time in agreement.

The comparison of the probabilistic forecasts based on Rad-TRAM and on COSMO-DE-EPS in terms of Brier score reveals that Rad-TRAM forecasts based on the latest observations clearly have a smaller mean error than COSMO-DE-EPS (Fig. 3.7, top). To longer Rad-

TRAM lead times, the differences get smaller, but still the nowcaster has more skill. This is generally seen as well with the CSRR (Fig. 3.7, middle). But for long lead times, the skill of Rad-TRAM forecasts is in the same order of magnitude as the one of COSMO-DE-EPS forecasts. The ROC areas of COSMO-DE-EPS forecasts loose skill over the day. This is not seen for the Rad-TRAM forecasts. Therefore, at the end of the day, Rad-TRAM clearly has more skill in discrimination than the forecasts derived from COSMO-DE-EPS.

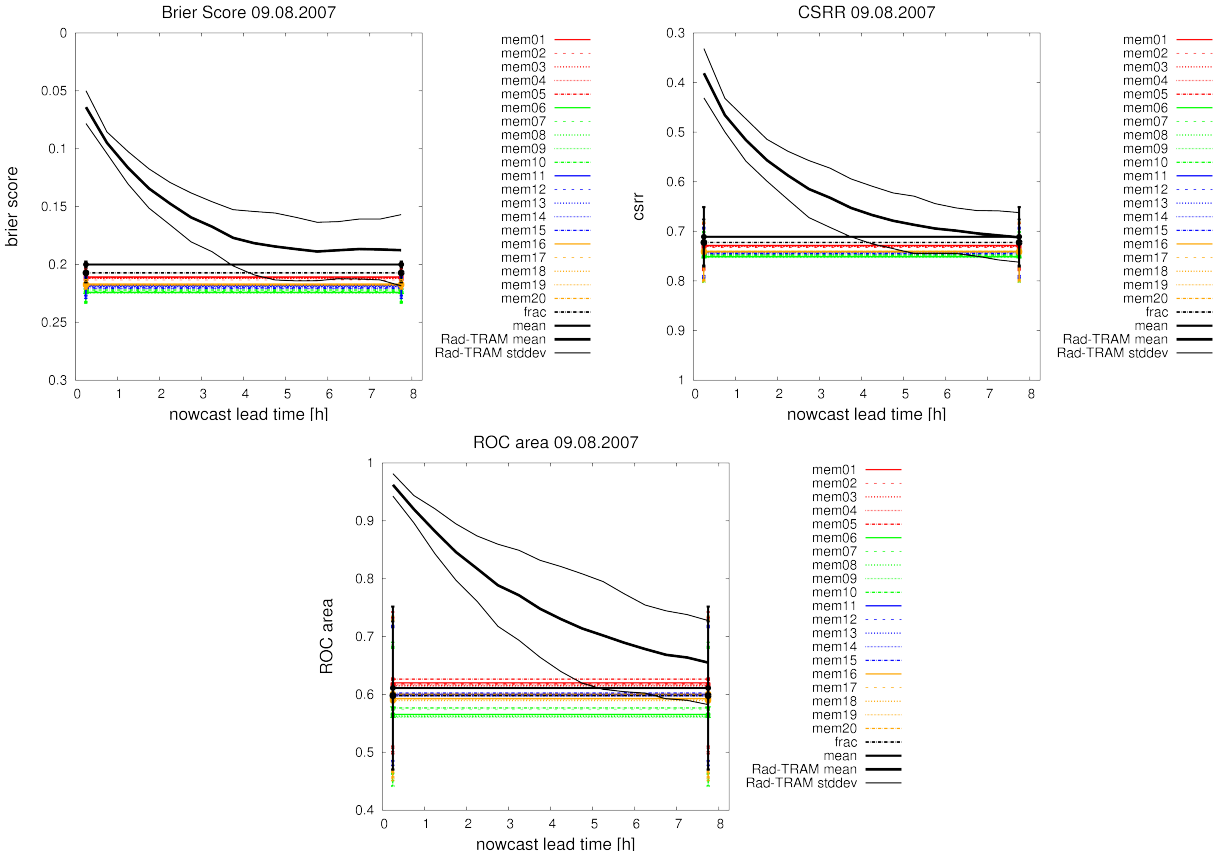

Figure 3.8: Development of Brier score, CSRR, and area under ROC curve with lead time for Rad-TRAM and calibrated COSMO-DE-EPS forecasts on 9 August 2007. Rad-TRAM's mean skill is displayed as thick black solid line and its standard deviation as thin black solid line. The COSMO-DE-EPS forecasts are colour-coded as in Fig. 3.7.

The development of forecast skill with lead time is evaluated for Rad-TRAM and calibrated COSMO-DE-EPS with the Brier score, the CSRR, and the ROC area in Fig. 3.8. The mean Brier score (Fig. 3.8, top left) of Rad-TRAM forecasts decreases with lead time and therefore, the mean skill. But even after eight hours Rad-TRAM has still more skill than the various model forecasts. The standard deviations of COSMO-DE-EPS forecasts are small in comparison to those of Rad-TRAM for long lead times. Taking the variability of the mean Rad-TRAM values into account, a cross-over point could be identified earliest after four hours. Nevertheless, even with consideration of the variability not with all model forecasts a cross-over point can be identified. The skill of the different forecasts based on COSMO-DE-EPS shows the mean method with more skill than the fraction method and the neighbourhood members. These are ranked following global models.
The CSRR (Fig. 3.8, top right) shows a similar behaviour as the Brier score: Rad-TRAM's

mean skill decreases steadily and rapidly during the first three forecast hours. In contrast to the Brier score, a cross-over point between the mean Rad-TRAM and mean COSMO-DE-EPS performance can be identified after eight hours. The standard deviations even cross already after four hours.
The ROC area (Fig. 3.8, bottom) reflects the low skill of the model at the end of the day as seen in the time series (cf. Fig. 3.7, right bottom) with low values of the mean. The variability in the small mean model values is very large. Rad-TRAM has more skill in discrimination than COSMO-DE-EPS over all lead times. Again, only considering the variability of the mean with the standard deviations enables the identification of a cross-over point.
All three score reflect the fact that on 9 August 2007, with the mean skill hardly a cross-over point can be identified. With the consideration of the standard deviation of the mean it is possible. The earliest cross-over point can then be found with the CSRR after four hours lead time.

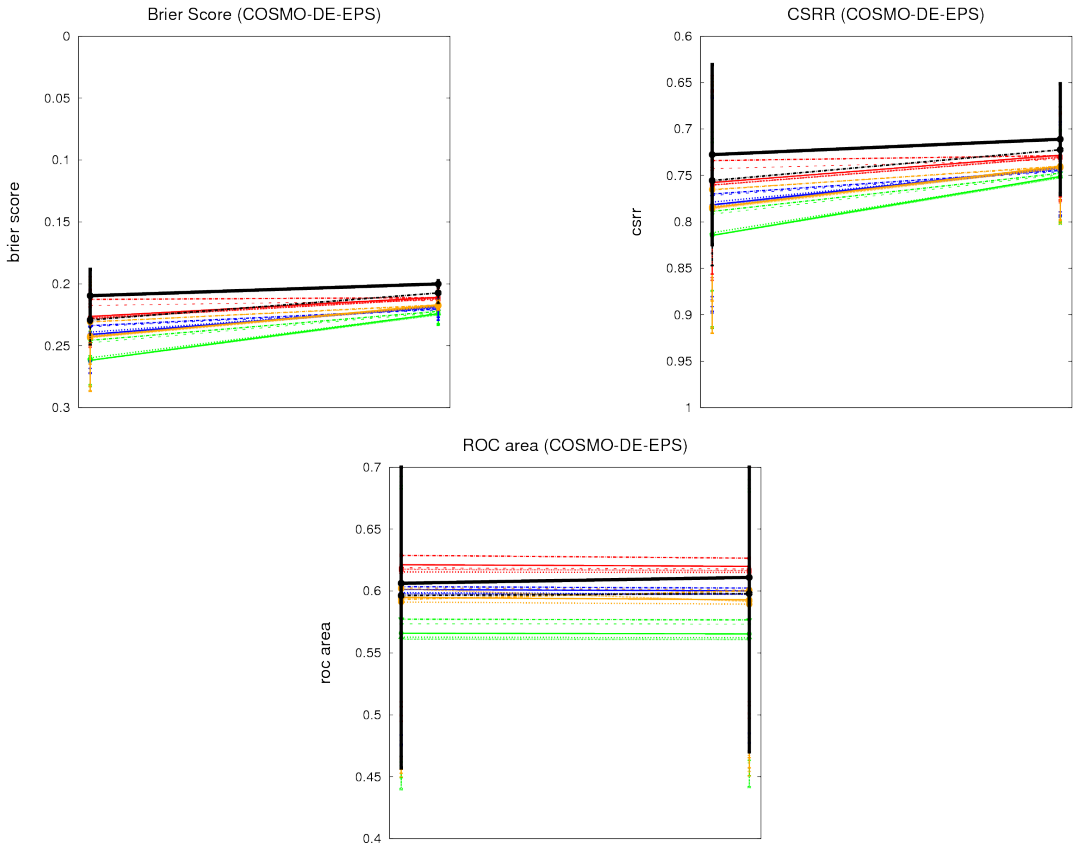

Figure 3.9: Effect of calibration of COSMO-DE-EPS probabilities on 9 August 2007 in Brier score, CSRR, and area under ROC curve with the calibrated forecasts on the left and the calibrated on the right. The forecasts are colour-coded as in Fig. 3.7.

The effect of calibration on the mean skill of COSMO-DE-EPS forecasts as evaluated in Fig. 3.8 is investigated with the change of mean values and their standard deviations from raw to calibrated probabilities with Brier score, CSRR, and the ROC area (Fig. 3.9). The Brier score and the CSRR show a very similar behaviour: the mean values of all solutions decrease and the spread of the different methods is reduced. Nevertheless, the spread between the methods and the different neighbourhood members is large before and after calibration

so that always a ranking can be established. As well with the Brier score and the CSRR, the effect on the neighbourhood members varies. The members with lateral boundary conditions from DWD global model (green) are improved more efficiently through calibration than the originally more skillful members based on ECMWF (red). The ranking is similar before and after calibration: the mean method has smaller errors than the neighbourhood members. As the fraction method is improved most efficiently, it is on rank two behind the mean method after calibration.
The area under the ROC curve shows hardly an effect of calibration (Fig. 3.9, bottom). Only the mean of the forecasts based on the mean of the neighbourhood members is slightly larger after calibration. The ranking is unchanged through calibration: the ECMWF based neighbourhood members have more skill than the mean method, the fraction method, and the other neighbourhood members.
To conclude, the effect of calibration on the mean skill of the COSMO-DE-EPS forecasts on 9 August 2007 is a reduction of spread between the different members and an improvement of the mean skill in Brier score and CSRR. In all three scores, the neighbourhood members are ranked following the lateral boundary conditions of the global models and the differences are visible even after calibration. The mean and the fraction method have more skill than the neighbourhood members consistently in all scores. The ROC area favours the members based on ECMWF before the mean and the fraction method. Note that the variability of the mean values is largest with the ROC area and smallest with the Brier score.

3.1.3 Discussion

The 9 August 2007 was characterised by a large amount of precipitation that covered large parts of the evaluation domain. Nevertheless, it was not an uniform, continuous field but divided into several intensive cells over different parts of the domain.
The quality concerning timing and location of the probabilistic forecasts based on Rad-TRAM was as expected high, especially for short lead times. With increasing lead time, the sharpness of the forecasts decreased. If the forecasted patterns existed long enough or a new was found in a similar location, the skill in location was high even for long lead times. If new precipitation came into the domain, Rad-TRAM was not able to represent this change as it is based on the extrapolation of already existing patterns.
The forecasts based on COSMO-DE-EPS performed not as well as Rad-TRAM in terms of timing and location. Generally, the first part of the day, with the intense precipitation field rotating north of the Alps around the low over northern Italy causing some convective cells over central Germany, was represented meaningful. But over the day the velocity with which the precipitation moved and its intensity was not well predicted by the 22 forecasts. This resulted in a large number of false alarms in the evening.
The time series of the Brier score, CSRR, and the ROC area of Rad-TRAM enhanced the impression from the snapshots in the eyeball investigation that the skill of forecasts with short lead times was very high. In each score, the skill decreased with lead time. The differences got small or even negligible for long lead times. The large difference of the short lead times' skill in Brier score in comparison to the uncertainty component leads to the conclusion, that these forecasts have skill in reliability and resolution. This skill decreased with lead time.
The skill of the forecasts based on COSMO-DE-EPS decreased over the course of the day. This can be seen with the CSRR and the ROC area. As the skill in discrimination as derived with the ROC area is low (values lower than 0.5) in the late evening, the forecasts even seem to be anticorrelated. The differences in skill of the different approaches applied

on COSMO-DE-EPS is in all scores smaller than the daily variability of the respective score. There seems to be no large skill in resolution and reliability as the shape of the Brier score varies only little from the uncertainty component.
The comparison of the development of skill of forecasts based on Rad-TRAM and on COSMO-DE-EPS with lead time meets the expectation that Rad-TRAM clearly has more skill than COSMO-DE-EPS for short lead times. On this particular day, Rad-TRAM was even superior for longer lead times and therefore, only considering the variability of each method enables the definition of a cross-over point. The reason for this behaviour is the relatively large amount of precipitation over the day. The chance of hitting a precipitation field in the observations with an extrapolation based forecast was very high. The phase where precipitation was interrupted and then followed by a field newly moving into the domain immediately showed up as a decrease in forecast skill. Of course, the late cross-over point also depends on the bad performance of COSMO-DE-EPS that, as discussed above, was not able to represent the actually observed situation adequately.
The effect of calibration on the mean performances of the different methods applied on COSMO-DE-EPS can be seen clearly with the Brier score and the CSRR. The sharpness of the model forecasts is decreased largely through calibration as no larger probabilities are forecasted yet. This has the positive effect that the possible magnitude of error is reduced. As the three calibration functions are very similar and for the neighbourhood members even equal, the differences between the methods are mainly the location of the probabilities. The amplitude of the probabilities is more similar after calibration. This explains the reduction in spread. The ROC area behaves different as it should not be affected by calibration. This condition is fulfilled. However, the mean shows a slight increase. As the ranking of the different methods is not affected, this can be neglected.

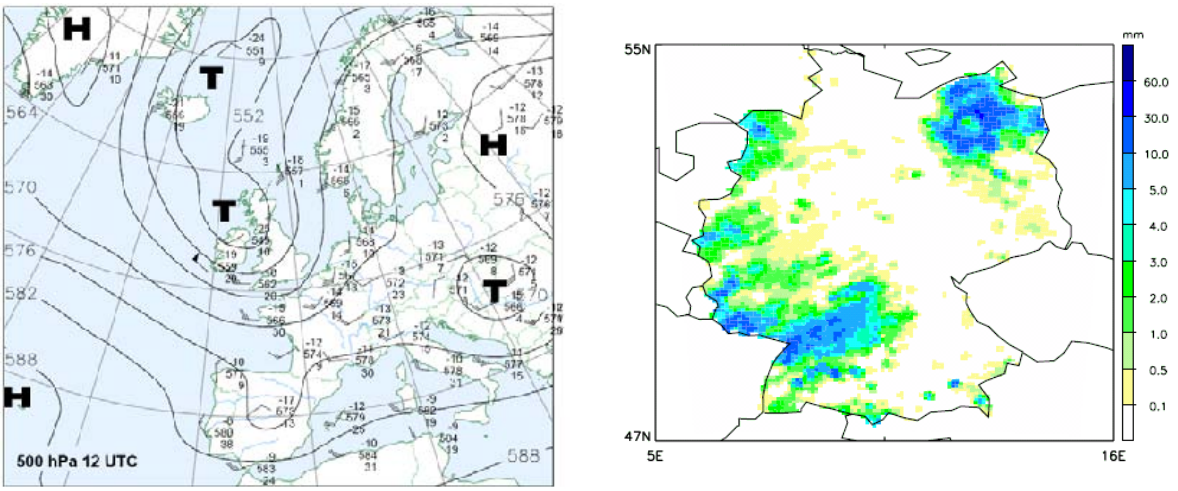

Figure 3.10: Synoptic overview: 500 hPa on 12 August 2007, 12 UTC and disaggregated daily precipitation sum over Germany (Zimmer and Wernli (2008)).

3.2 IOP 15: 12 August 2007 (Regime change)

3.2.1 Synoptic overview

In the first half of 12 August 2007, western Germany was under the influence of a weak ridge of high pressure causing large scale descent (Fig. 3.10, left). Nevertheless, some convective cells developed in the early morning over the south of Germany (Fig. 3.10, right). They were weak with radar reflectivities reaching not more than 28 dBZ (Fig. 3.11, left). After a short phase with almost no convective activity around noon (Fig. 3.11, middle), an upper-level trough approached from the West led by a cold front in the afternoon. This front caused ascent and some convective cells. The resulting total precipitation can be seen over western Germany (Fig. 3.10, right). The overall amount of precipitation was low as only some cells developed within the front. At 23:15 UTC, the front was located over western Germany and total precipitation was not very high as the frontal band was small (Fig. 3.11, right). However, some of the embedded cells caused heavy rain with large radar reflectivities.

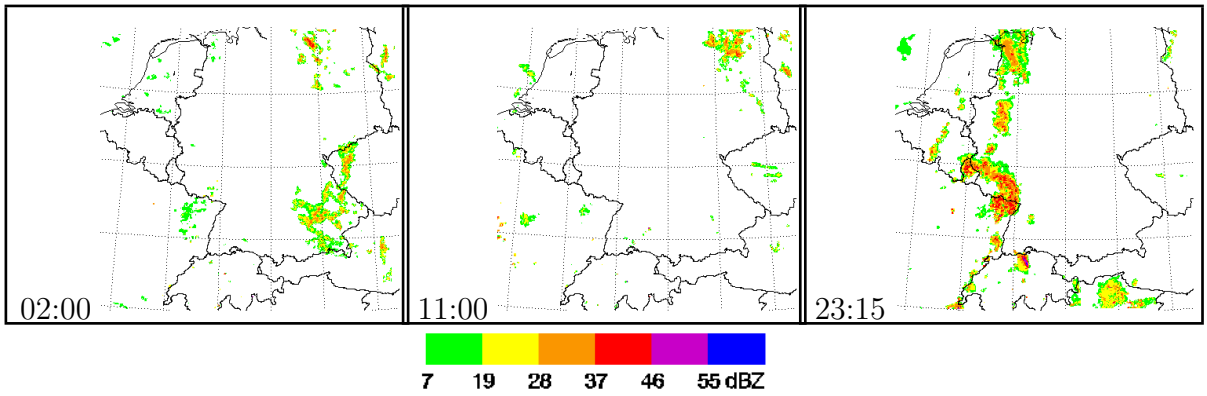

Figure 3.11: Observed radar reflectivity on 12 August 2007, 2:00 UTC, 11:00 UTC, and 23:15 UTC.

The uncertainty component of the Brier score reflects the synoptical development in terms of the relative frequency of 19 dBZ in the observations on 12 August 2007 (Fig. 3.12). The

total number of events (reaching threshold of 19 dBZ) was low in comparison to other days (e.g., Fig. 3.4). From 0:15 UTC to 4:00 UTC uncertainty was small due to a small number of events. Then, after a further decrease of uncertainty to almost zero until 14:00 UTC, the number of events, and therefore, uncertainty, increases with the front coming into the domain from the west. Uncertainty is highest between 19:00 and 23:45 UTC.

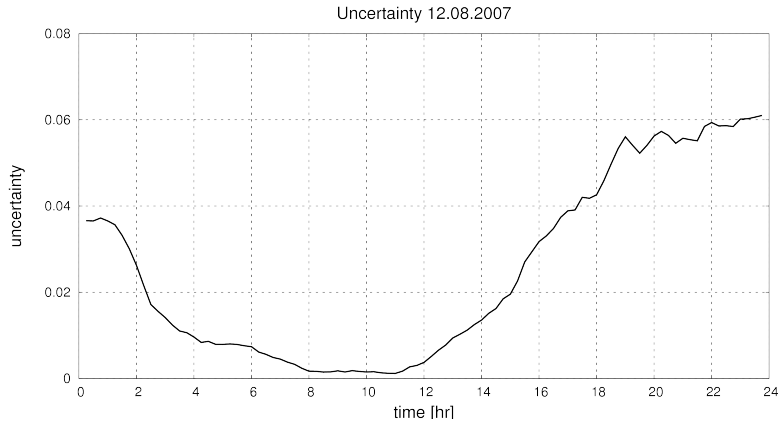

Figure 3.12: Uncertainty component of the Brier score on 12 August 2007.

3.2.2 Quality of probabilistic forecasts

Probabilistic forecasts provided by Rad-TRAM for 23:15 UTC are displayed in Fig. 3.13. The nowcasts are based on the observations 15 minutes (60 and 120 min respectively) ago. The forecast with the shortest lead time produces the sharpest probability field (Fig. 3.13a). With increasing lead time the probability field is smoothed out (Fig. 3.13b and c). This means, the probabilities are lower and the area over which the field extends increases. This reflects the increasing uncertainty in the forecasts and is a result of the increasing search area in which the probability is derived. The location of the probability forecasts of all lead times is reasonable compared to the observed reflectivity field (Fig. 3.11, right). But

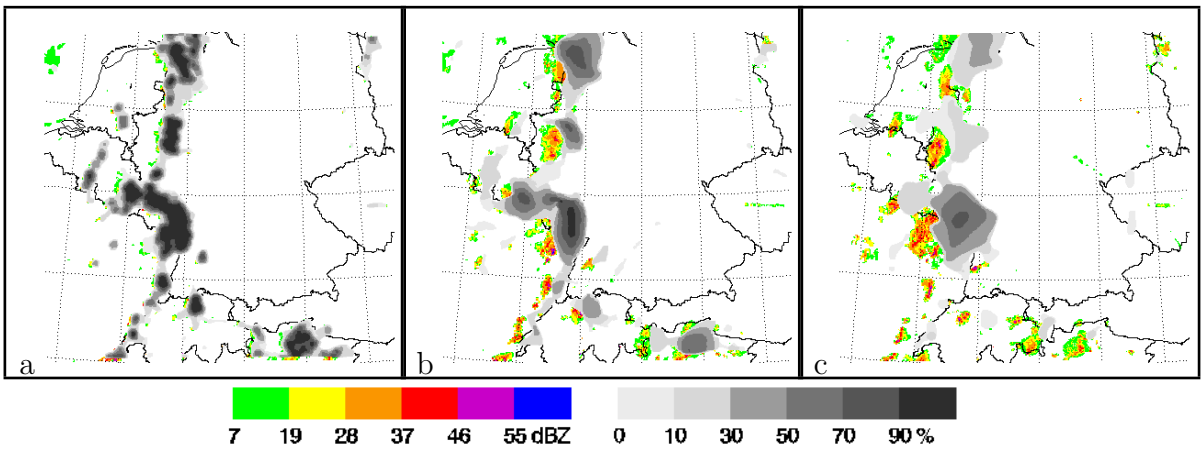

Figure 3.13: Probabilistic forecasts of Rad-TRAM for 12 August 2007, 23:15 UTC based on (a) 15 min forecast from 23:00 UTC, (b) 60 min forecast from 22:15 UTC and (c) 120 min forecast from 21:15 UTC grey-shaded and the reflectivity observations at the respective initial time colour-coded in the background.

investigating the position in detail shows some deviations. The two hour forecast already shows a deviation of the forecast position to the observed: the cells are extrapolated in a north easterly direction, but the front actually moved straight east.

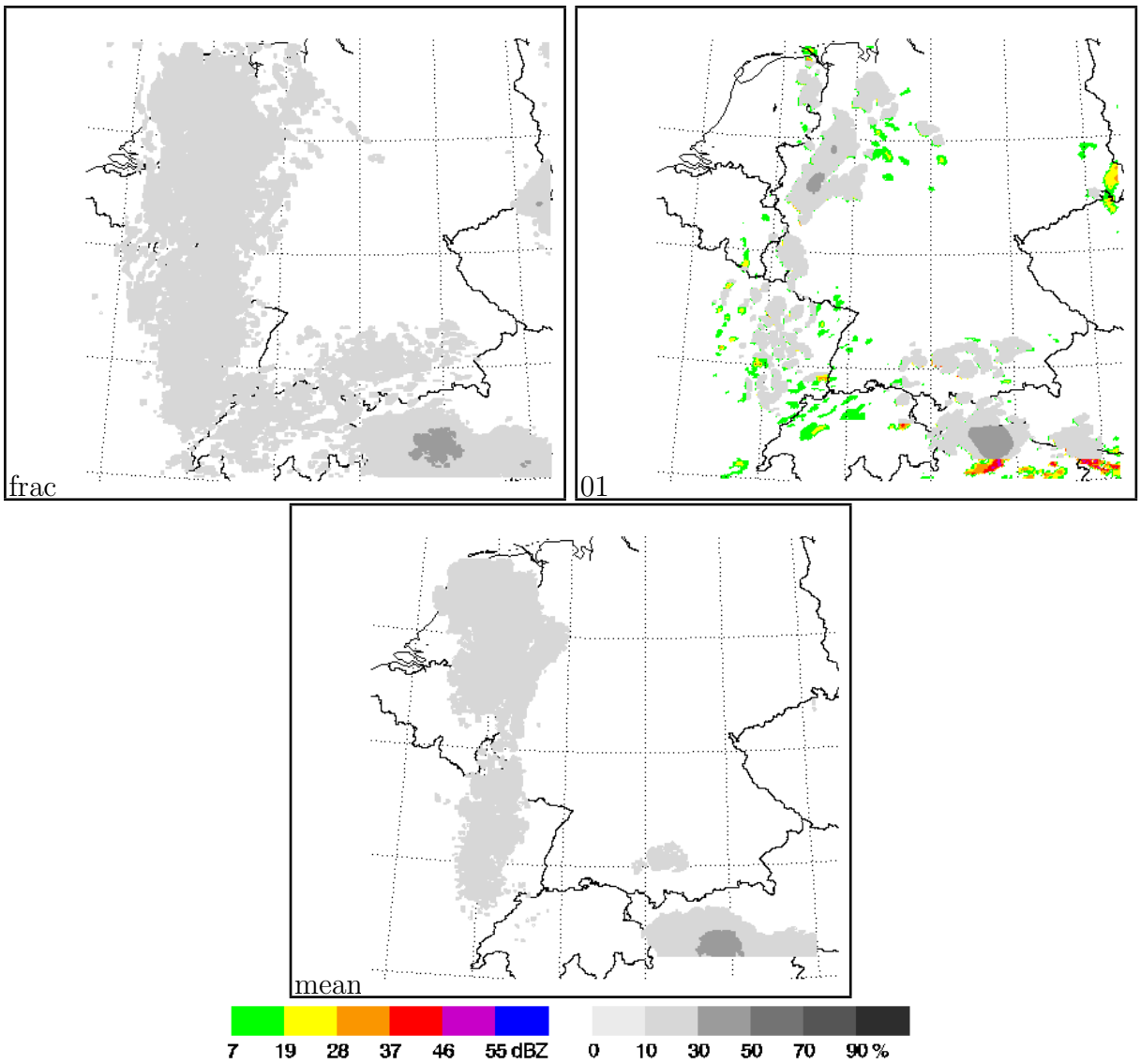

Figure 3.14: Probabilistic COSMO-DE-EPS forecasts for 12 August 2007, 23:15 UTC (grey-shaded) with the fraction method (frac), member 1 as representative for the ensemble, and the mean of the ensemble (mean). In the background of member 1 colour-coded the synthetic radar reflectivities in 850 hPa.

Figure 3.14 shows three of the 22 calibrated probabilistic forecasts based on COSMO-DE-EPS for 23:15 UTC: based on the fraction method (frac), member 1 based on the neighbourhood method representing the ensemble, and the mean of the neighbourhood probabilities (mean). Comparing these with the observations (Fig. 3.11) shows that they all predict a probability of precipitation larger zero in a meaningful location. Nevertheless, the maxima in probability are low ranging between 10 and 30 %. Location and intensity of the probability fields differ within the methods. The fraction method has a large, but spotted field. The area where the precipitation actually was observed is only partly covered anyway. Therefore, and due to the size of the probability field, there is a large number of false alarms. Member 1

has a small, spotted probability field. The size and the distribution of the precipitation in the field are reasonable compared to the actual meteorological situation. Although some of the observed precipitation patterns are predicted, there are still some false alarms and even misses. For example, the cell over western Germany is not captured (Fig. 3.11). The mean of the 20 neighbourhood probabilities is a large smooth field with low probabilities. As the probabilities for the single members are low, the mean of course is even lower. As already seen with the fraction method, the large area that is covered by the probabilities enables a good probability of hitting the precipitation leading to a high false alarm rate. The mean reflects the high variability in solutions of the ensemble.

The evaluation of the forecast quality on 12 August 2007 as time series is shown in Fig. 3.15. The forecasts based on Rad-TRAM show in Brier score a high daily variability (Fig. 3.15, left top). At the beginning and especially the end of the day, the the error is larger than at the time around noon. In the morning hours, the performance of the different lead times is very similar and only the first two lead times can be clearly separated from the others. After noon, when the cold front comes into the domain, the number of distinguishable lead times increases and they are ranked following lead time. At 19:00 UTC, even the fourth forecast hour can be distinguished. The comparison with the uncertainty component of the Brier score (Fig. 3.12) reveals a large agreement with the longer lead times. Therefore, the skill in reliability and resolution is only high for lead times up to four hours, especially in the second half of the day.
The CSRR has a large variability as well (Fig. 3.15, left middle). The period with almost no precipitation can be identified as a first minimum in CSRR. Earlier, the forecasts are ranked following lead time, but as within these are forecasts based on the first run, the shortcomings mentioned in the previous section have to be considered. In the second half of the day, the ranking following lead time is very clearly defined. For longer lead times that are based on observations during the phase with almost no precipitation (12:00 to 20:00 UTC), there are no differences in quality. During the frontal passage in the afternoon and evening, the skill for short lead times even improves (15min forecast: 0.6 at 14:00 UTC and 0.4 at 22:00 UTC). The number of distinguishable lead times increases from one at 13:00 UTC to five hours at 22:00 UTC.
The increase in skill and number of distinguishable forecasts can be seen with the ROC area as well (Fig. 3.15, left bottom). During the dry phase, there is a loss in skill. The equality of the forecasts based observations during this time is seen with the different lead times in the following hours. In the first half of the day, it is hard to identify a distinct ranking of the forecasts with different lead times except in the first forecast hour. The longer lead times follow no structure. In the afternoon, the skill of forecast concerning discrimination is well ranked following lead time with the shortest lead times showing most skill. Differences get smaller with increasing lead time.
The Brier score for COSMO-DE-EPS is at the beginning and particularly at the end of the day higher than around noon (Fig. 3.15, right top). The Brier score shows no large spread within the different solutions. The daily variability is larger than the spread within the solutions. Only the fraction method varies compared to the other methods. At the beginning, it is worse than the other neighbourhood members. At the end of the day, the fraction method has higher skill than the others, together with member 1. As the deviation of the Brier score from its uncertainty component (Fig. 3.12) is very small, the forecasts hardly have skill in resolution or reliability.
The CSRR (Fig. 3.15, right middle) shows in comparison to the Brier score a variability between the different approaches applied on COSMO-DE-EPS comparable to that during

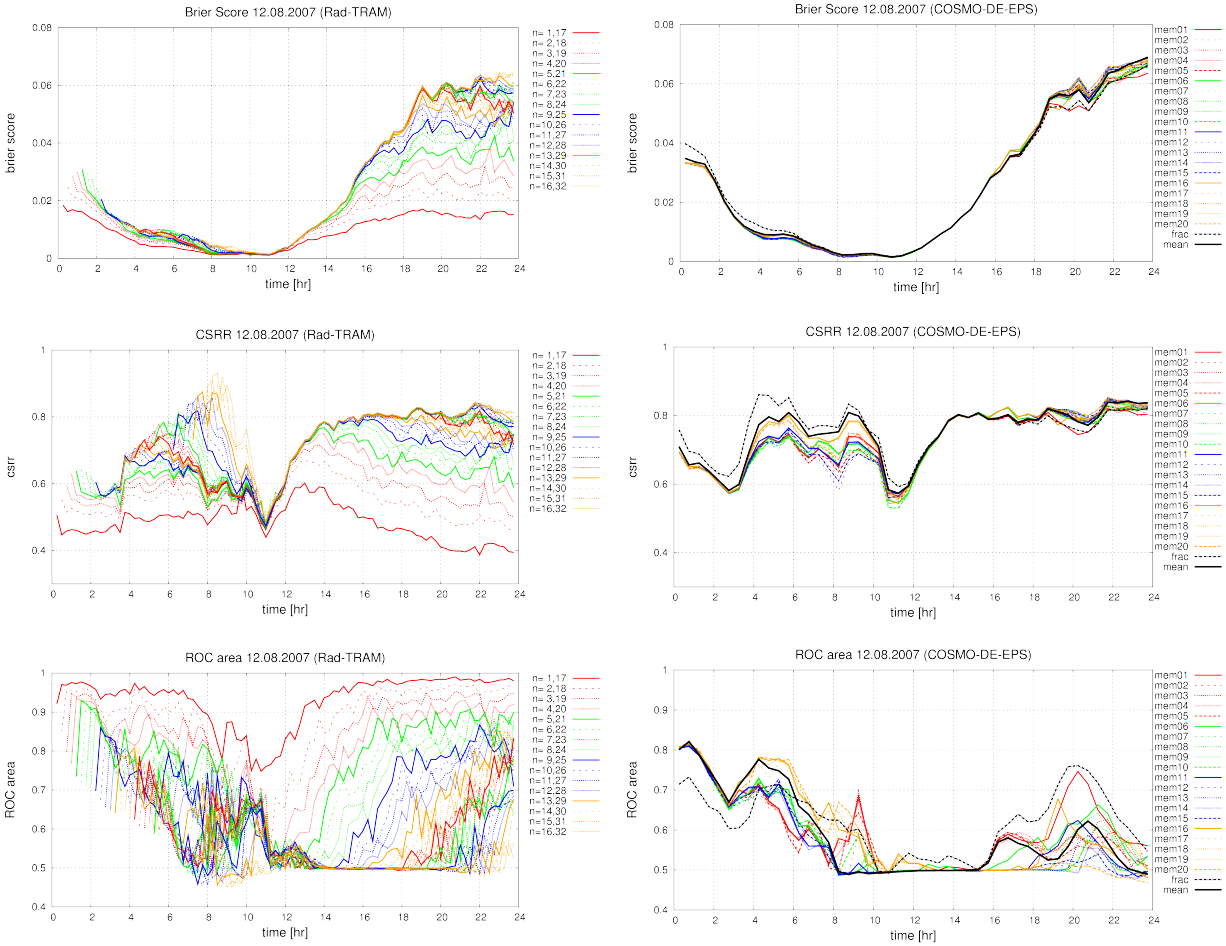

Figure 3.15: Development of Brier Score, CSRR, and area under ROC curve for Rad-TRAM (left) and COSMO-DE-EPS (right) forecasts from 12 August 2007. Colours and lines as explained in Fig. 3.7.

the course of the day. The increase of skill in the dry phase can be seen during 8:00 UTC and 12:00 UTC. This implies that the different forecasts are able to reproduce the period without precipitation. The spread between the approaches and members is larger than on 9 August, especially in the first half of the day. The neighbourhood members are ranked roughly following global models. The fraction method is as in the Brier score remarkable as it is until 10:00 UTC worse than the other methods and members. At the end of the day, the fraction method has more skill than the others together with member 1 in agreement with the Brier score.

The ROC area has a maximum value of 0.8 at the beginning of the day (Fig. 3.15, right bottom). During the day, values are seldom larger than 0.7 and therefore, the skill in discrimination is small. During the time when almost no events occurred, the low values of the area under the ROC curve imply that no forecast is able to discriminate between occurrence and nonoccurrence of the event. The ranking of the methods varies over the day and variability is large. The fraction method is during the first forecast hours significantly worse than the others. The neighbourhood members are, in agreement with the CSRR, ranked following global models. But different members have the highest skill. After the regime change, the fraction method clearly has more skill than the other solutions. Again, member 1 is good in discrimination, but as well the other members based on the first physical perturbation

(entrainment rate of shallow convection).

Comparing the skill of COSMO-DE-EPS with Rad-TRAM forecasts in the Brier score (Fig. 3.15, top) reveals a similar development with a minimum around 10:00 UTC and a maximum in the late evening hours. The uncertainty component determines the shape of the curves (cf. Fig. 3.12). As the uncertainty only depends on the observations, the development of the Brier score is mainly determined by the frequency of the event during the day. During the entire day, Rad-TRAM forecasts based on the latest observations have smaller errors than any of the forecasts based on COSMO-DE-EPS.
The CSRR reflects the specific meteorological situation with the dry period around noon before the frontal passage as well (Fig. 3.15, middle). Again, the latest Rad-TRAM forecasts are better than COSMO-DE-EPS forecasts. But the model forecasts are in the range of the Rad-TRAM forecasts based on four hours or older observations.
The ROC area varies over a large range for Rad-TRAM (Fig. 3.15, bottom). The skill of the COSMO-DE-EPS forecasts is clearly lower than Rad-TRAM's with lead times smaller than four hours. But for longer lead times, Rad-TRAM's skill decreases rapidly and therefore, COSMO-DE-EPS gains skill in comparison.

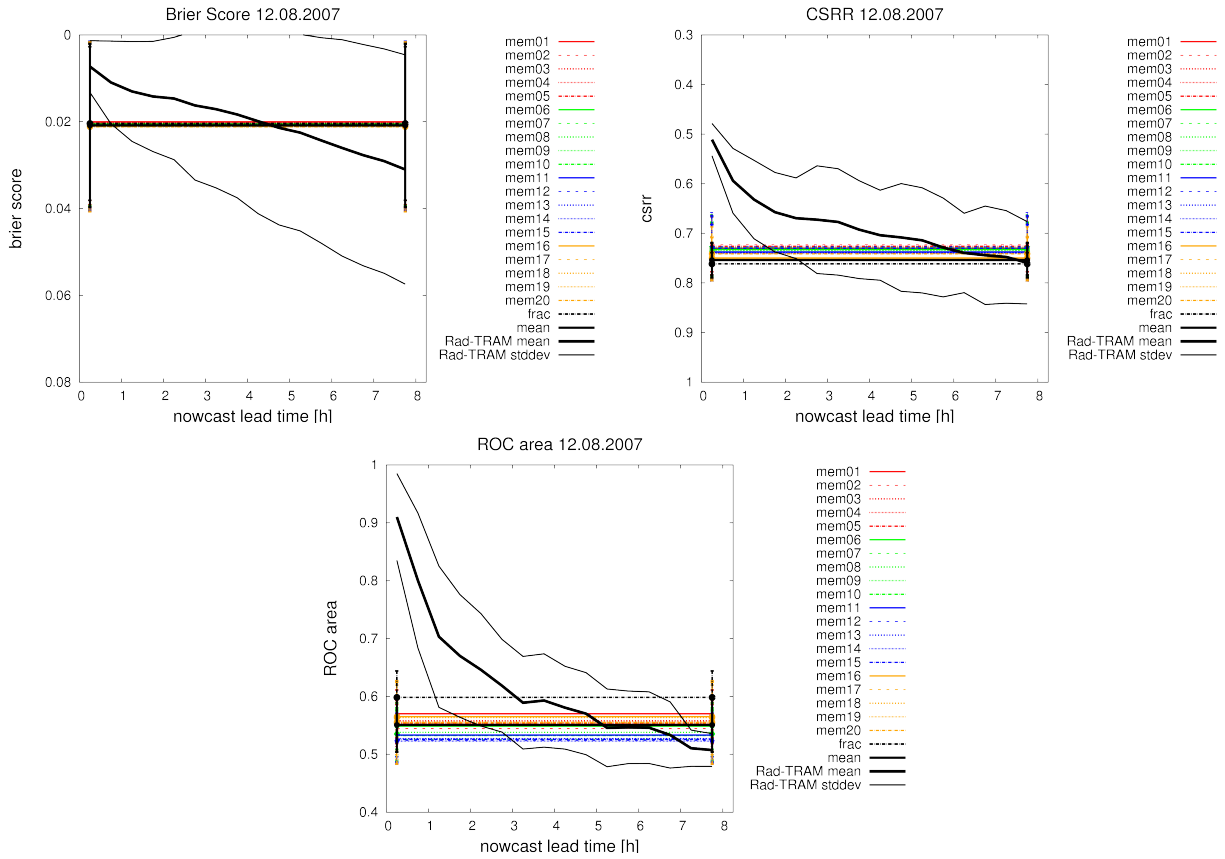

Figure 3.16: Development of Brier score, CSRR, and area under ROC curve with lead time for Rad-TRAM and and calibrated COSMO-DE-EPS forecasts on 12 August 2007. Colours and lines as explained in Fig. 3.8.

The development of forecast skill of Rad-TRAM and COSMO-DE-EPS with lead time is evaluated in Fig. 3.16. The scores show differences in the development of Rad-TRAM's mean

skill. In the Brier score (Fig. 3.16, top left), the difference between Rad-TRAM and COSMO-DE-EPS forecasts is small (cf. Fig. 3.8). But in the first four forecast hours, Rad-TRAM is better and in the second four hours worse than COSMO-DE-EPS forecasts. The variability of the mean value as seen with the standard deviation is large for both, Rad-TRAM and COSMO-DE-EPS. The differences between the methods applied on COSMO-DE-EPS to derive probabilistic forecasts are very small. The lead time since when the mean model forecasts outperform the nowcasts (cross-over point) is around 4.5 hours.
The behaviour as shown with CSRR is different (Fig. 3.16, top right). Here, Rad-TRAM forecasts have more skill for longer lead times. The rate of decrease in skill of the mean value is more rapid in the first three forecast hours than later. Rad-TRAM's mean skill falls below the COSMO-DE-EPS forecasts between 5.5 and 7.75 hours. CSRR shows some spread between the different model forecasts. The forecasts are ranked following the first physical perturbation (entrainment rate of shallow convection) and DWD and ECMWF lateral boundary conditions as best neighbourhood members. The fraction and the mean method have the lowest skill. In CSRR, the variability of Rad-TRAM's mean skill is larger than the variability within COSMO-DE-EPS based forecasts.
The ROC area shows a very rapid decrease of initially high skill for Rad-TRAM. The variability of the mean increases with lead time, but gets smaller in the last forecast hour. The cross-over point with the COSMO-DE-EPS forecasts is between 3 and 7 hours. This long range shows that the spread between the methods is very large. The fraction method performs significantly better than the neighbourhood members and their mean. Interestingly, the members with the lowest skill in CSRR now have the highest.

The effect of calibration on the mean and the standard deviation of the different approaches applied on COSMO-DE-EPS forecasts is shown in Fig. 3.17. The Brier score shows hardly a change through calibration (Fig. 3.17, top left). Only the fraction method is slightly improved. Before calibration, the mean method was ranked before the neighbourhood members and the fraction method. After calibration, there is no visible difference.
With CSRR a larger effect of calibration can be seen (Fig. 3.17, top right). Before calibration, there is a large spread between the methods and members. Two physical perturbations of the turbulence scheme (subscale cloud cover and asymptotic mixing length) of the NCEP and the ECMWF driven members have more skill than the others. The mean method is within the neighbourhood members and the fraction method has the lowest skill. The calibration improves all forecasts but the impact is different. Therefore, spread is reduced and ranking changes slightly. As the effect of calibration on the mean method is very small, its position in ranking changes. All neighbourhood members have smaller errors than the mean and the fraction method. The physical perturbations (subscale cloud cover and asymptotic mixing length) of the NCEP and the ECMWF are still the best neighbourhood members. The improvement of fraction method is the largest.
The area under the ROC curve is hardly affected and therefore, the ranking is the same before and after calibration (Fig. 3.17, bottom). The fraction method is better than the neighbourhood members and the mean method is within them. Only the value of the fraction method is improved. Generally, the values are very small (below 0.6), reflecting the low skill in discrimination of the model on this day.
Generally, the effect of calibration in this case study is small. Two of the three score hardly show a change. In CSRR, the effect can be summarised as reduction of spread within the different methods and members and improvement of the mean performance as in the first case study. The largest effect is seen in the fraction method.

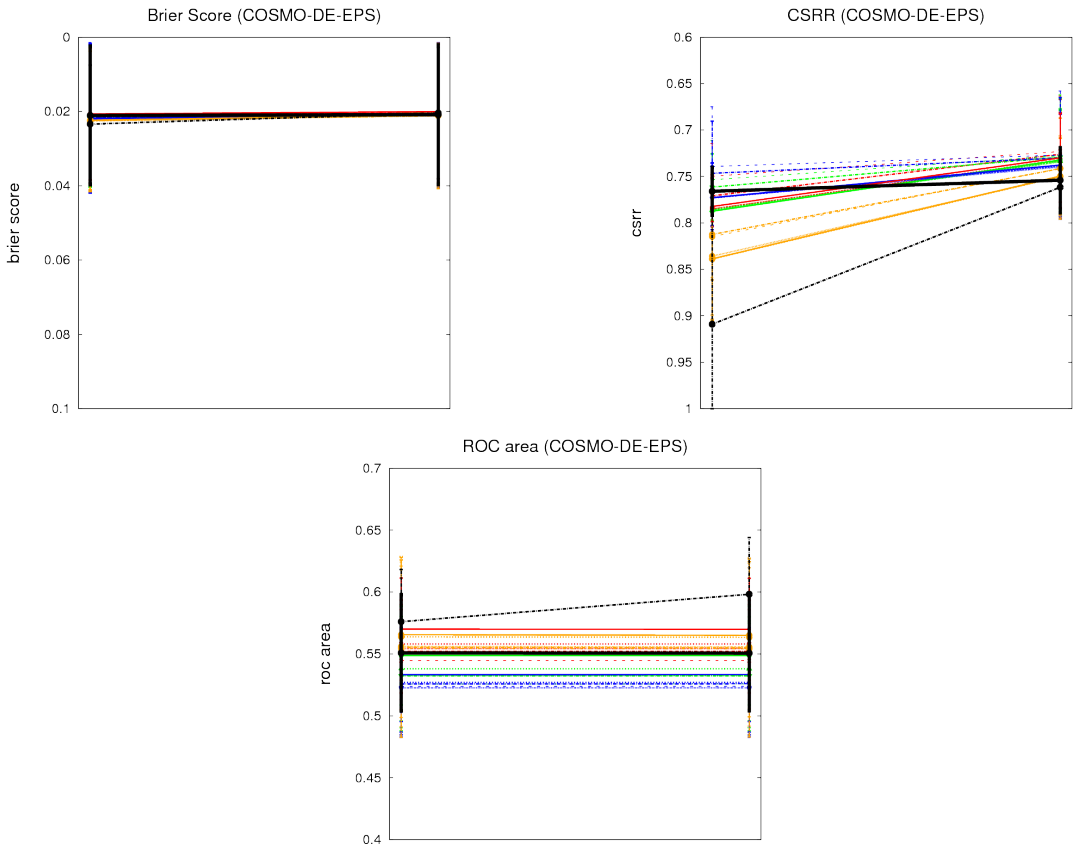

Figure 3.17: Effect of calibration of COSMO-DE-EPS probabilities on 12 August 2007 in Brier score, CSRR, and area under ROC curve (left: raw probabilities, right: calibrated). Colours and lines as explained in Fig. 3.9.

3.2.3 Discussion

On 12 August 2007, the dominating meteorological phenomenon was the regime change between a phase when hardly convection occurred to the passage of a weak cold front. The forecasts based on Rad-TRAM are able to capture the frontal situation very well, as this development can be very well described by advection. The change, when the cold front comes into the domain where no event occurred before, cannot be captured by the nowcaster. The forecasts based on COSMO-DE-EPS capture the dry phase around noon very good, but fail to predict the exact location and amount of precipitation caused ahead of the cold front. At the beginning of the day, the small cells are overestimated. Again, the forecasts based on Rad-TRAM have very intense sharpness for short lead times that decreases for longer lead times so that then, neither Rad-TRAM nor COSMO-DE-EPS forecasts are sharp.

The time series of the Brier score, CSRR, and the ROC area of both Rad-TRAM and COSMO-DE-EPS reflect the three phases with different synoptical situations of the day. Each phase shows a characteristic performance. Rad-TRAM has high skill in all scores for very short lead times. The skill decreases with increasing lead time and differences between the lead times get smaller. The number of distinguishable forecasts varies and is highest during the passage of the cold front. The skill of forecasts based on COSMO-DE-EPS is significantly lower in all three scores as Rad-TRAM's for short lead times. But it is in the magnitude of longer Rad-TRAM lead times. Surprisingly, the skill of the forecasts in dis-

crimination vanishes during the dry phase although the members correctly forecasted the nonoccurrence of the event. The ROC area varies over the entire range of possible values for Rad-TRAM. But neither in Rad-TRAM nor in COSMO-DE-EPS forecasts fall beneath the 0.5 threshold as in the case study of the 9 August 2007.

The comparison of the two forecast sources in the lead time dependent evaluation differs from the 9 August 2007 in some details. But it generally shows as well in all scores that short Rad-TRAM lead times have more skill than the forecasts derived from COSMO-DE-EPS. The Brier score is relatively small on this day due to the small number of the events. Rad-TRAM's mean skill decreases linearly in the Brier score and it is lower than the mean COSMO-DE-EPS skill already after 4 hours. The methods applied on the COSMO-DE-EPS hardly have spread. The variability of the mean is large and in the magnitude of the decrease in mean for Rad-TRAM for both forecast sources. The development of skill in the CSRR behaves different. Rad-TRAM's mean decreases very fast in the first two forecast hours but then much slower. Therefore, Rad-TRAM is outperformed by the model forecasts after 5.5 to 8 hours. The variability is in the magnitude of the 9 August, but the skill of the early Rad-TRAM forecasts is much worse with 0.5 in comparison to 0.4. Whereas the magnitude of the mean COSMO-DE-EPS skill is on both days between 0.7 and 0.8.
The ROC area of mean Rad-TRAM decreases more rapidly on 12 as on 9 August. Also the skill of the model is lower. Nevertheless, the cross-over points are already between 3 and 7 hours. The general skill of COSMO-DE-EPS based forecasts is lower, but as well the standard deviation.
The effect of calibration varies within the scores and is different in comparison to 9 August. Now, in the Brier score hardly any effect can be seen through calibration. But the values are a factor 10 smaller. The CSRR again is very sensitive to calibration with different intensity concerning the different methods. The effect on the fraction method and the neighbourhood members based on UKMO is largest and on the mean method smallest. The ranking is changed through calibration. The ROC area again is not sensitive to calibration in most of the methods. On 12 August, only the fraction method is changed. In contrast, on 9 August the fraction was insensitive but the mean method was changed.

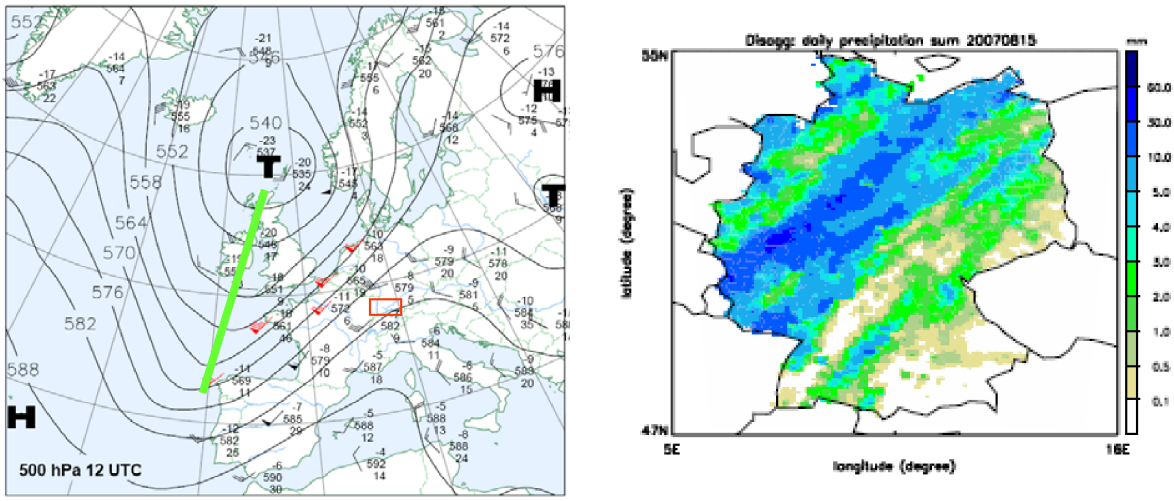

Figure 3.18: Synoptic overview: 500 hPa on 15 August 2007, 12 UTC and disaggregated daily precipitation sum over Germany (Zimmer and Wernli (2008)).

3.3 IOP 16: 15 August 2007 (Forced frontal convection)

3.3.1 Synoptic overview

On 15 August 2007 (IOP 16), a pronounced trough in 500 hPa over western Europe approached and slowly moved eastward (Fig. 3.18, left). A frontal system connected with this trough passed through the evaluation domain on this day. In the first third of the day, the warm front crossed the northern part of Germany and brought some precipitation (Fig. 3.19, 4:00 UTC). The warm front moved straight easterly. The warm air sector after the front and the large scale descent of air resulted in a dry period (Fig. 3.19, 10:00 UTC). Ahead of the following cold front, some isolated showers developed (Fig. 3.19, 13:00 and 19:00 UTC). The cold front passed Germany in the late evening (Fig. 3.19, 22:00 UTC) and caused convective development within the front and the frontal lines behind. The aggregated daily precipitation over Germany (Fig. 3.18, right) shows that the main precipitation took place over western and northern Germany. Nevertheless, some precipitation was observed over southern Germany. The cold front moved in north easterly direction as seen in the patterns in Fig. 3.18, right.

The uncertainty component of the Brier Score (Fig. 3.20) reflects the observed frequency of 19 dBZ in the evaluation domain. Clearly, the four above described phases can be separated. At the beginning of the day, there is the small scale precipitation connected with the warm front (0:15 to 6:00 UTC). The warm and dry air of the warm sector within the fronts and the resulting decrease of the frequency of the events is seen from 6:00 to 11:00 UTC. The precipitation ahead of the cold front leads to an increase of uncertainty (12:00 to 18:00 UTC). Uncertainty is maximised during the passage of the front as then almost half of the evaluation domain is covered with reflectivities of at least 19 dBZ (12:00 to 23:45 UTC).

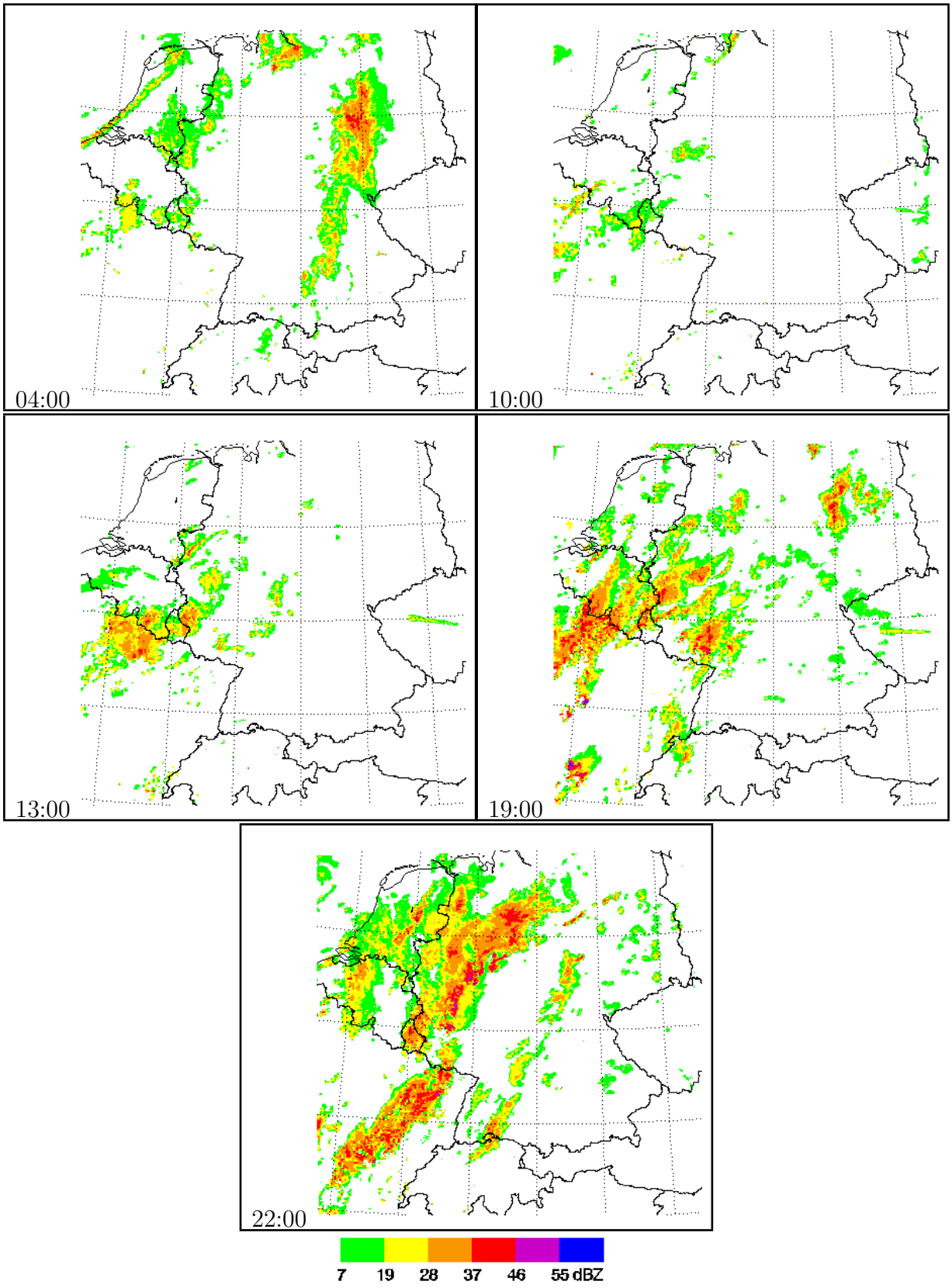

Figure 3.19: Observed radar reflectivity on 15 August 2007, 04:00 UTC, 10:00 UTC, 13:00 UTC, 19:00 UTC, and 22:00 UTC.

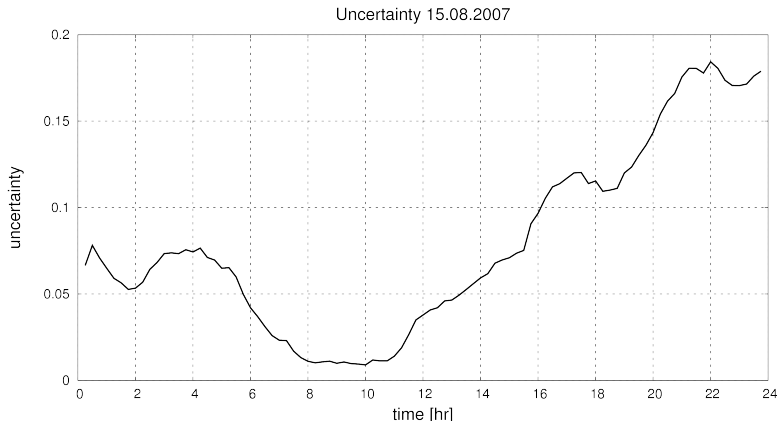

Figure 3.20: Uncertainty component of the Brier score on 15 August 2007.

3.3.2 Quality of probabilistic forecasts

Figure 3.21 shows forecasts based on Rad-TRAM with different lead times for 19:00 UTC. During this phase of the day, some intensive convective cells developed ahead of the cold front (cf. Fig. 3.19, 19:00 UTC). The forecast based on the latest observation (Fig. 3.21a) is very sharp and reflects all details of the small scale structure of the precipitation field. The comparison with the observation indicates no errors in location. The forecast based on the observations one hour ago (18:00 UTC, Fig. 3.21b) is still relatively sharp and detailed, but the probability field is smoother with lower maxima. Generally, the location of the probability field is in the area where the precipitation actually occurred. But there are already some deviations in small scale location. The forecast based on the observation two hours ago (17:00 UTC, Fig. 3.21c) still has two maxima in the probability of reaching 19 dBZ over western Germany, but generally the field is smoother and there is no small scale structure. The probability field is so widespread that there is a good chance of hitting the event but accompanied by a large false alarm rate. Some smaller cells like the one over north eastern Germany and over France/Switzerland are kept well. The low probability reflects the fact that the actual position in the small scale cannot be predicted with a high certainty. The last forecast is based on the observation three hours ago (16:00 UTC, Fig. 3.21d). Also in this forecast, there is still some skill in the location of the probability field. But the position of the highest probabilities does not fit to the position of the actually observed patterns (cf. Fig. 3.19, 19:00 UTC). The small patterns over southern Germany are not predicted at all. The later observed patterns are not yet in the domain at the time the forecast is created and therefore, an extrapolation forecast cannot predict them.

Six of the 22 calibrated forecasts derived from the COSMO-DE-EPS output for 13:15 UTC are compared with the observation: the fraction method, 4 members based on the neighbourhood method, and the mean method (Fig. 3.22). The neighbourhood members are all based on the first perturbation in model physics (entrainment rate of shallow convection) but are driven with different lateral boundary conditions. At this time, only in the western part of the domain some precipitation occurred ahead of the cold front (cf. Fig. 3.19). The forecasts differ in synthetic reflectivity fields and therefore, in the derived probability fields. The fraction method predicts low probabilities (mainly between 10 and 30 %) of reaching 19 dBZ over a relatively large area of the western domain. But large parts of the probability field are too far north in comparison to the observation. Also a probability of precipitation

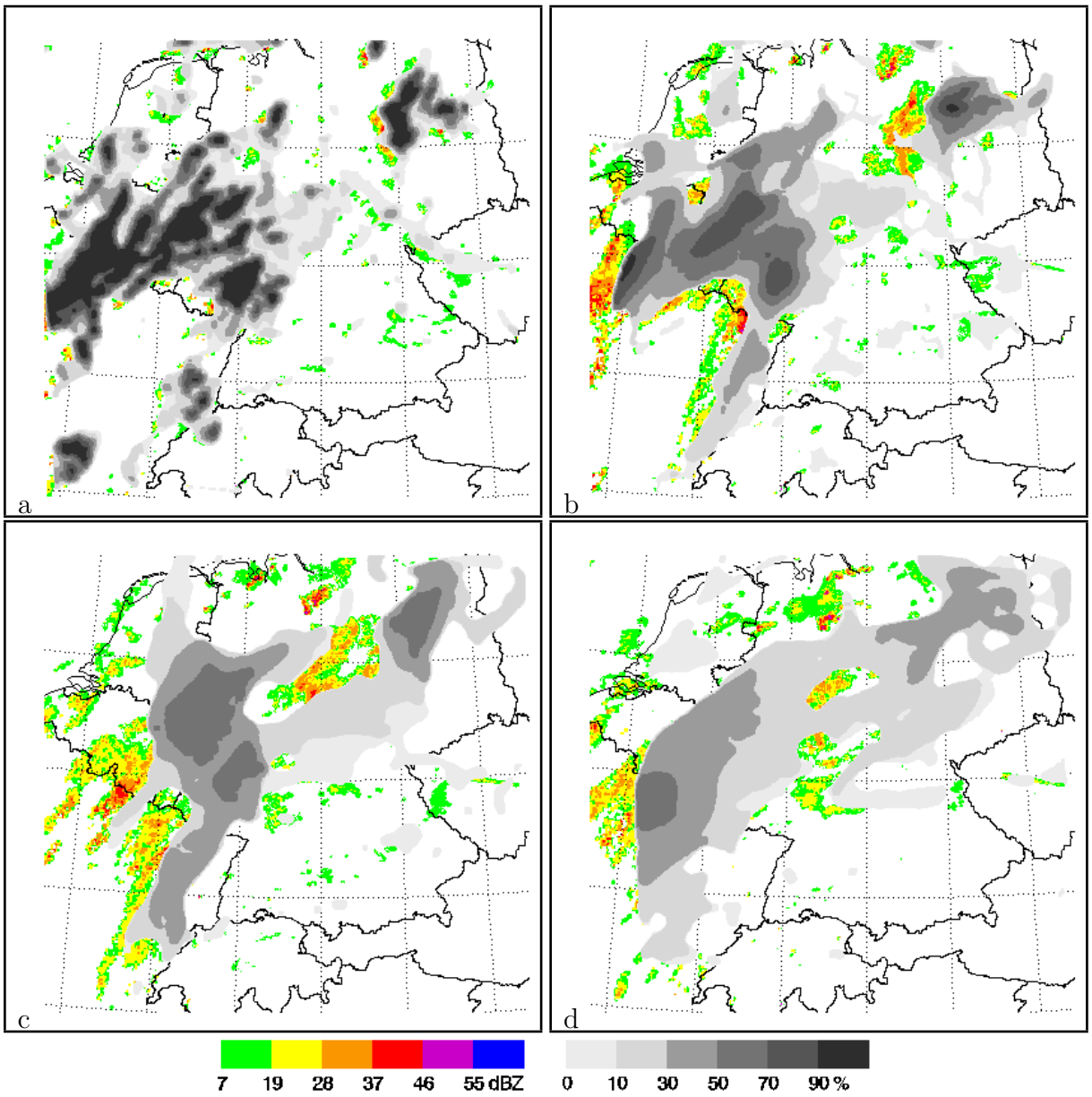

Figure 3.21: Probabilistic forecasts of Rad-TRAM for 15 August 2007, 19:00 UTC based on (a) 15 min forecast from 18:45 UTC, (b) 60 min forecast from 18:00 UTC, (c) 120 min forecast from 17:00 UTC, and (d) 180 min forecast from 16:00 UTC grey-shaded and the reflectivity observations at the respective initial times colour-coded in the background.

larger zero is predicted over eastern Germany where actually no event occurred. Members 1 and 6 (based on lateral boundary conditions from ECMWF and DWD) predict a small and spotted probability field with low amplitude (hardly probabilities larger than 30 %). This is a large underestimation and does not represent the observed prefrontal precipitation as the rain area was relatively large and uniform over northern France and Belgium. Members 11 and 16 (NCEP and UKMO respectively) both predict a significantly larger rain area. Member 16 still predicts probabilities larger than 50 % after calibration. As this member also predicts at least partly the right position it clearly fits better to the observation than member 11. Here, the area with probability of precipitation is located too far north. The mean method, that can also be understood as the conclusion of the neighbourhood members, predicts a uniform probability of precipitation field that covers large parts of Belgium and

western Germany. As there the precipitation field was observed, the forecast has skill. But as the observed precipitation field was smaller, a large number of false alarms is predicted. Again, the forecast based on the mean method looks similar to the fraction method but is less spotted.

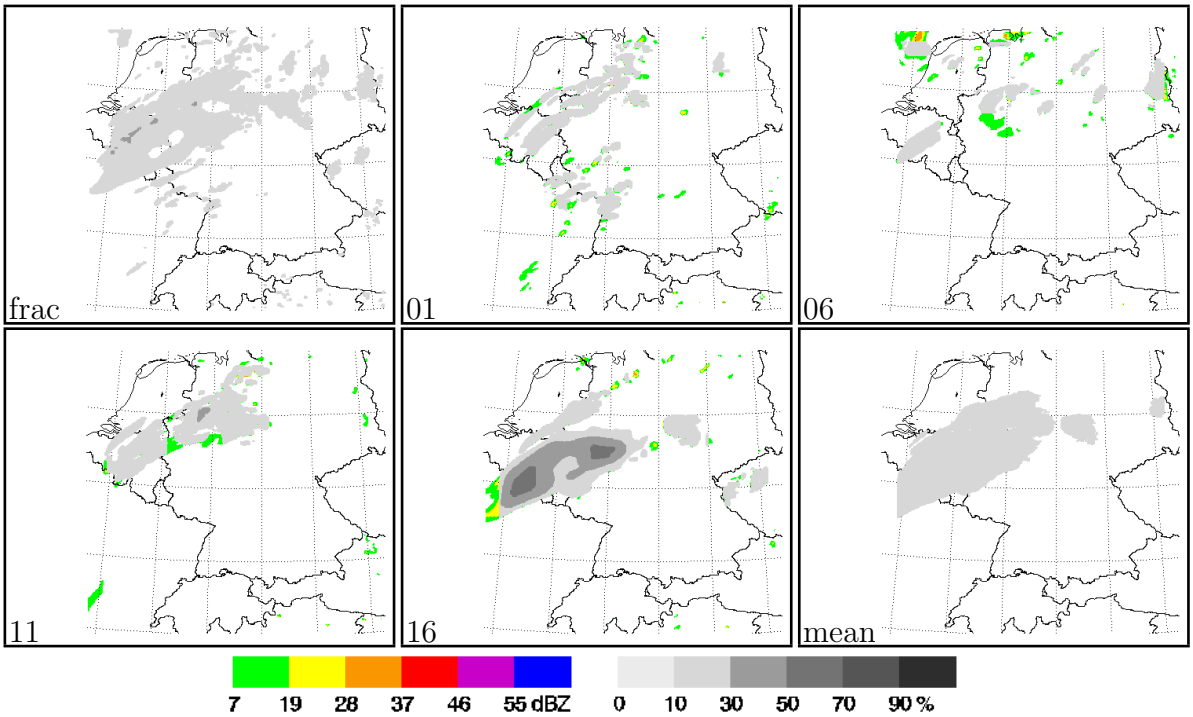

Figure 3.22: Calibrated probabilistic COSMO-DE-EPS forecasts for 15 August 2007, 13:15 UTC (grey-shaded) for the fraction method (frac), member 1, 6, 11, and 16 as examples for the neighbourhood members, and the mean of the neighbourhood members (mean). In the background of the neighbourhood members colour-coded the synthetic radar reflectivities in 850 hPa.

Figure 3.23 reveals the quality of the probabilistic forecasts based on Rad-TRAM (left) and calibrated COSMO-DE-EPS (right) as time series. The overall shape of the Brier score for Rad-TRAM (Fig. 3.23, left top) is dominated in the long lead times, as in the two investigated cases before, by the shape of the uncertainty component of the decomposed Brier score (cf. Fig. 3.20). But the short lead times reveal a different representation. This indicates, that there is resolution and reliability in the forecasts with short lead time. As in the two investigated cases before, there different forecasts are ranked following lead time. In the first phase of the day, when the warm front crosses the domain, the longer lead times are not completely available. Therefore, no ranking can be identified in this period. But the lead times that could already be verified are ranked following lead time. When the cold front with the prefrontal convection moved through the domain in the second half of the day, the number of distinguishable forecasts increases.
The CSRR (Fig. 3.23, left middle) for short lead times behaves relatively uniform over the entire day. The longer lead times show some variability. The values of the eighth forecast hours have a large peak around 8:00 UTC. This is due a combination of the problem with the very first forecasts and the very small precipitation field at this time. The number of distinguishable forecasts is very small in agreement to the Brier score: during the passage of the warm front, the first two forecast hours are well ordered in comparison to only one

during the dry period. In the second half of the day, the number of distinguishable forecasts further increases. The magnitudes of the CSRR for longer lead times are very similar.
The values of the ROC area range over the entire possible range of values on 15 August (Fig. 3.23, left bottom). Rad-TRAM forecasts up to one hour lead time have large skill in discrimination over the entire day. The decrease of skill with lead time is very rapidly, but varies in the four phases of the day. For longer lead times, the variability is very high without a clear ranking. The values of long lead times fall beneath 0.5 during the dry phase. This is the threshold where the forecasts have no skill in discrimination.
The three quality measures agree in their judgement of Rad-TRAM's skill. Even an exception like the decrease in skill for one time step (forecasts based on 15:45 UTC) is visible in all scores, as well for longer lead times. At this time step, there was an anomaly in the observations where a relatively large part of the precipitation field stayed stationary for one time step. It is not clear if this phenomena actually occurred or if there was an error in postprocessing the data by the weather service.

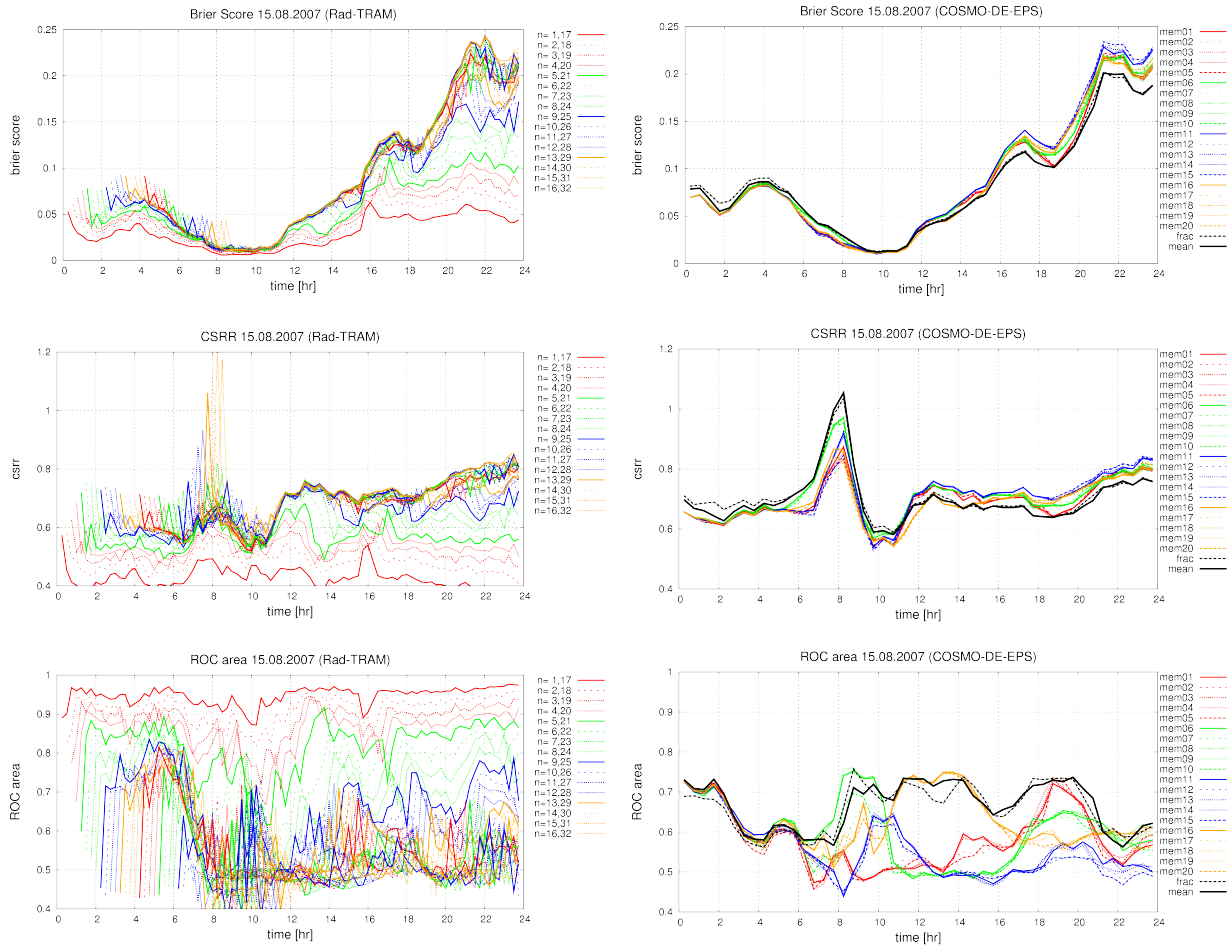

Figure 3.23: Development of Brier Score, CSRR, and area under the ROC curve for Rad-TRAM (left) and COSMO-DE-EPS (right) forecasts from 15 August 2007. Colours and lines as explained in Fig. 3.7.

The quality of the forecasts based on the COSMO-DE-EPS is displayed in the right column of Figure 3.23. The Brier score shows small differences between the different solutions (Fig. 3.23, right top). In the first part of the day, the variability is even smaller, but fraction

and mean method have the lowest skill. During the course of the day, the spread increases. During the passage of the cold front through the domain, the fraction and the mean method have more skill than the neighbourhood members. These are ranked following global models with small differences.
The CSRR shows only a slightly larger variability between the solutions (Fig. 3.23, right middle), but the variability of the score over the day is different. As seen with the Brier score, the daily variability is larger than the spread between the solutions. All members generally have the same shape of the curves. Especially the peak at 8:00 UTC is seen with all members. In this phase of the day, the number of events was very small (warm sector after warm front), but all members predicted precipitation as in the model the front moved significantly slower. The weighting with the small area of precipitation in the observations together with the overestimation in the forecasts caused the increase in CSRR. The ranking of the different solutions is in agreement with the Brier score: during the passage of the warm front, the fraction and the mean method have lowest skill and after the dry phase and during the passage of the cold front, the mean and the fraction method have highest skill. The neighbourhood members are as well ordered following global models.
The most interesting behaviour is seen with the ROC area (Fig. 3.23, right bottom). In the first six forecast hours, the difference between the methods is small. It can only be distinguished that the fraction method at the beginning has the lowest skill as seen on the other investigated days. But after six hours, there are large differences between the different solutions and especially the neighbourhood members have large spread. They are clearly ranked following the lateral boundary conditions form the different global models. The ranking of the neighbourhood members changes several times. Values of the ROC area are always beneath 0.8 and therefore, the general skill is small for all methods. Some members even fall beneath the no skill in discrimination value (0.5). During the entire period (after 6:00 UTC), the mean and the fraction method are very well in comparison to all neighbourhood members. But there is always at least one group of the global models that performs at least as good or better. This group changes several times in this period. For example, during 8:00 and 10:00 UTC, the members with the lateral boundary conditions from DWD have more skill than the other neighbourhood members. This is quite surprising, because in this period, in the observations the warm front already left the evaluation domain (Fig. 3.19) and the DWD members predicted the largest amount of precipitation. In contrast to the area under the ROC curve, the CSRR correctly ranks them worst. In the following (11:00 to 16:00 UTC), the DWD members rapidly loose skill and the members with the lateral boundary conditions from UKMO perform best, together with the fraction and the mean method. This is in agreement with the CSRR and can be confirmed by the comparison of forecasts and observations. During this phase (Fig. 3.19, 13:00 UTC), a small precipitation pattern ahead of the cold front moved into the domain from the west. The members based on UKMO (as representative member 16 in Fig. 3.22), predicted a large covered probability field for this area whereas the other members predicted smaller, spotted fields. Later, the members based on UKMO loose skill as they predicted the precipitation embedded in the cold front too small. The frontal precipitation in the cold front was predicted best by the members based on ECMWF (18:00 to 20:00 UTC). This is confirmed by the Brier score and the CSRR. The ECMWF members predicted the largest reflectivity fields and as the position of the field also is meaningful, all scores favour these forecasts. After 22:00 UTC all members loose skill and the differences get smaller.

Comparing the forecasts based on Rad-TRAM and COSMO-DE-EPS in Brier score (Fig. 3.23, top), it is seen, similar to the previous case studies, that the shape of the curves for long

Rad-TRAM lead times and COSMO-DE-EPS is strongly related to the uncertainty component of the Brier score (cf. Fig. 3.20). Only short Rad-TRAM forecast differ and therefore, show skill in reliability and resolution. As well the CSRR (Fig. 3.23, middle) shows similarities in the development of skill of the two forecasts sources. The large error through the overestimation of the rain area in the dry phase around 8:00 UTC is seen as well with COSMO-DE-EPS and long Rad-TRAM forecasts. Again, the Rad-TRAM forecasts based on the latest observation (0-3 hours) have significantly more skill than the forecast derived from COSMO-DE-EPS. The ROC area (Fig. 3.23, bottom) repeats the behaviour seen in the first case studies: Rad-TRAM's skill varies over the entire range of possible values and the COSMO-DE-EPS forecasts have lower skill with relatively large spread.

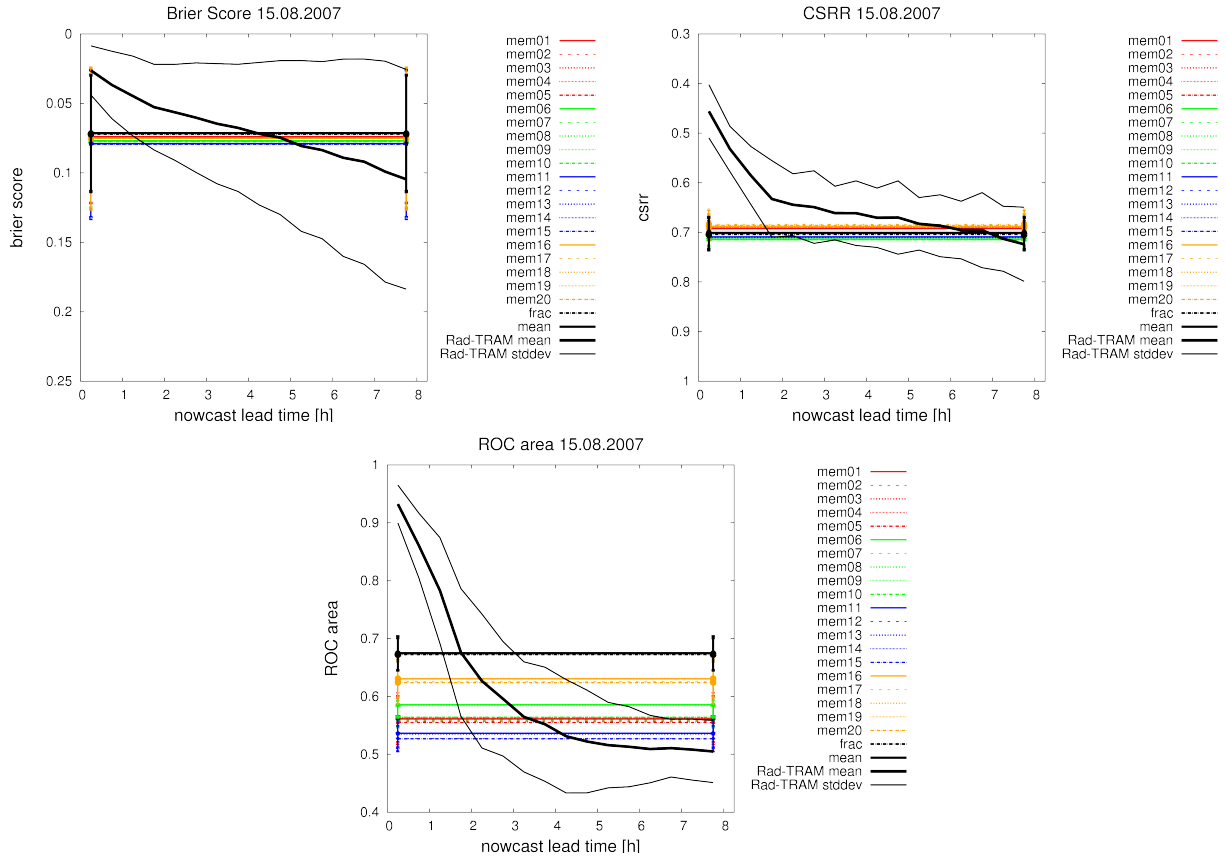

Figure 3.24: Development of Brier score, CSRR, and area under ROC curve with lead time for Rad-TRAM and calibrated COSMO-DE-EPS forecasts on 15 August 2007. Colours and lines as explained in Fig. 3.8.

In Fig. 3.24, the development of forecast quality with lead time is displayed for Rad-TRAM and the COSMO-DE-EPS forecasts. The Brier score shows a constant decrease of the mean skill of Rad-TRAM forecasts over the eight hours lead time (Fig. 3.24, top left). The variability is as large as the rate of decrease. The difference between the different methods applied on the COSMO-DE-EPS is small, but the fraction and the mean method have a higher skill than the neighbourhood members. These are ranked following global models. As well the forecasts based on COSMO-DE-EPS have large standard deviations. The cross-over points when the model forecasts have smaller errors than the mean Rad-TRAM forecasts is between 4 and 5.5 hours.
Regarding the CSRR (Fig. 3.24, top right) a different behaviour in the development of skill

is seen. The decrease of Rad-TRAM's mean forecast skill is very fast in the first two forecast hours and gets slower afterwards. The standard deviation is smaller than in Brier score. The standard deviations of the COSMO-DE-EPS forecasts is also clearly smaller. The spread between the methods is in the range of the Brier score but the ranking differs. The members based on the ECMWF and the UKMO have more skill than the mean method, the fraction method, and the other neighbourhood members. The cross-over point is found between 5.25 and 7.5 hours.
The area under the ROC curve (Fig. 3.24, bottom) reflects the large variability and spread of the COSMO-DE-EPS forecasts as seen in the time series (Fig. 3.23). The mean values vary over a large range. The mean and the fraction method clearly beat the neighbourhood members. These are ranked following global models with the UKMO members leading. The decrease of the mean values of Rad-TRAM is fast. Already after four hours of lead time, the mean almost has no skill in discrimination (near 0.5). Considering the standard deviation, even earlier cross-over points are possible. As the variation within the methods is large the range for the cross-over points is large as well. Depending on method, it ranges between 2 and 4.75 hours. This is clearly earlier than in the first case studies.

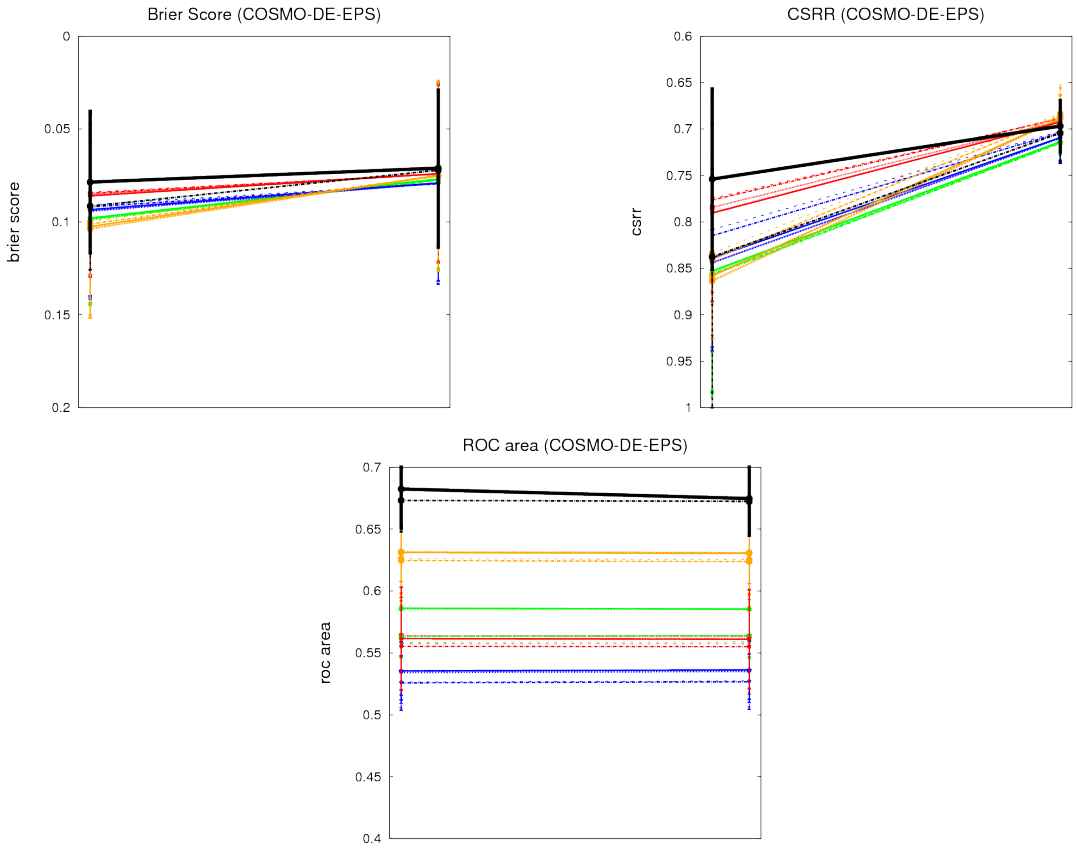

Figure 3.25: Effect of calibration of COSMO-DE-EPS probabilities on 15 August 2007 in Brier score, CSRR, and area under ROC curve (left: raw probabilities, right: calibrated). Colours and lines as explained in Fig. 3.9.

The effect of calibration on the 15 August 2007 can be seen in Fig. 3.25. The Brier score and the CSRR show the behaviour of increasing skill and reduced spread as seen on the 12 August 2007. Again, the methods differ in sensitivity and therefore, the ranking in Brier score and CSRR is changed through calibration. The mean method performs best (Brier score) or

within the best members (CSRR). The area under the ROC curve on 15 August 2007 shows a small decrease in skill for the mean method. The other forecasts stay unchanged through calibration in the ROC area.

3.3.3 Discussion

The 15 August was from a meteorological point of view a very interesting day as a complete frontal system with a warm front, a sector of warm air and a cold front with prefrontal convection passed the evaluation domain. This means, the forecasts had to reflect several regime changes. The nowcaster Rad-TRAM, as already seen on 9 and 12 August, fails to predict changes, but is very good within a specific situation. Therefore, short lead times show high skill in agreement in all scores. The forecasts based on COSMO-DE-EPS were special on this day as the differences between the members with the different driving global models were very large. Whereas those based on physical perturbations were negligible small. Interestingly, in each situation of the day, at least one group of the neighbourhood members gave meaningful results.
The evaluation of the development skill with lead time confirmed some findings of the other two cases. The Brier score of both Rad-TRAM and COSMO-DE-EPS is highly dominated by the uncertainty component. Deviations are only seen for short lead times of Rad-TRAM. The CSRR shows for the model forecasts that if no or only a small rain area was observed but a not negligible amount was predicted, a large error in CSRR is the result. The area under the ROC curve is the only score that reflects the large differences between the model forecasts. The overall skill in discrimination of the forecasts based on COSMO-DE-EPS is low, but the variability is very interesting. Short Rad-TRAM forecasts have very high skill whereas long forecasts perform much worse than seen on 9 August.
The evaluation of the development of forecast skill with lead time fits quite well to the performance seen on 12 August. The decrease in Rad-TRAM is also with a constant rate, although the absolute values are a factor of 10 larger. Also the standard deviation is similar (in the magnitude of the decrease of the mean). The cross-over points in Brier score and CSRR are found in the same time frame. The magnitude of COSMO-DE-EPS skill in CSRR is very similar as found on 9 and 12 August. The ROC area behaves different as on the other investigated cases. The decrease of mean Rad-TRAM skill is large and falls to very small values. Whereas the skill in discrimination of COSMO-DE-EPS forecasts is extraordinary good (almost 0.7 for the fraction and the mean method).
The effect of calibration as seen in the other cases is confirmed. The effect on Brier score is small but observable. The effect on CSRR is large in terms of improving values and decreasing spread. The ROC area is hardly affected as only the mean method is slightly reduced.

3.4 Summary

The systematic evaluation of the forecast skill of Rad-TRAM and COSMO-DE-EPS in different meteorological situations revealed some similarities. On all three days, the high skill of the nowcaster concerning sharpness, reliability, and resolution for short lead times was seen. The rate of decrease with lead times varied on the three days. The skill of the model forecasts was clearly lower but differences decreased with increasing lead time. For long lead

times, as well Rad-TRAM and COSMO-DE-EPS forecasts hardly had skill in sharpness. The comparison of the Brier scores with the uncertainty component indicated that the skill in resolution and reliability was low for both methods as well.
On all days, as expected, the forecasts could not reflect changes in the overall meteorological situation. Furthermore, in all cases the differences between the different lead times of Rad-TRAM were larger than the differences between the model solutions. Although the spread was small, a specific ranking could be established on the single days, especially on 15 August. This ranking changed between the different days. The neighbourhood members were either ordered following the lateral boundary conditions of the driving global models (15 August) or the perturbations in the model physics (12 August). As well the ranking of the mean and the fraction method in comparison to the neighbourhood members varied. The effect of calibration of all days can be summarised as a reduction of spread and an increase of skill in CSRR and the Brier score. The area under the ROC curve was not affected. The largest effect is seen on the fraction as their calibration results in the largest loss of sharpness.
Disagreements between the different case studies are found between the different skill scores on 12 August around noon, whereas on the other days they generally agreed. On 15 August, the forecasts revealed to be anticorrelated in the ROC area, whereas on 12 August, the forecasts never fall beneath the 0.5 threshold.
But the main difference between the three case studies is the time frame of the cross-over time. As expected, on days with a large amount of precipitation, the cross-over time is late and even not found within the investigated eight hour time frame under consideration of the standard deviations. Whereas, on days with small amounts of precipitation and small cells, the cross-over time is earlier. In frontal situations, the nowcasts are superior for longer lead times.

Chapter 4

Quality of probabilistic forecasts - Overview over general performance

Whereas in the last chapter three specific case studies were discussed, in this chapter, the entire period from 8 to 16 August 2007 will be investigated. Generally, the investigation is conducted similar as the single case studies with the evaluation of the development of skill with time and with lead time. But now, additionally reliability diagrams will be evaluated. This was not reasonable in case studies due to the relatively small amount of data on one day, but with the entire period it is possible. Furthermore, the calibrated and the uncalibrated probabilities from COSMO-DE-EPS are evaluated not only in mean but also in timeseries and reliability diagrams to see the differences and the effect of calibration more systematically.

In this evaluation, the main similarities of the single case studies should be confirmed. The quality of Rad-TRAM forecasts should be superior to COSMO-DE-EPS forecasts in the first forecast hours. It is expected that the differences between the three methods applied on the ensemble output will be small. The effect of calibration should be visible as decrease in spread and increase in mean skill for all three methods. The effect on the fraction method should be the largest due to the initially high sharpness. For the blending procedure that is the final goal of this thesis, the general cross-over time that will be evaluated in this chapter, is of high importance as it is the basis for the definition of the weighting functions.

4.1 Overview over relative frequency of event in observations

Figure 4.1 shows the uncertainty component of the decomposed Brier score from 8 to 16 August 2007. The x-axis displays the time in hours, starting at 0 (8 August 2007, 0:00 UTC) and ending at hour 216 (16 August 2007, 23:45 UTC). During this period, the uncertainty varied almost over all possible values (0.0 to 0.25). Only with the information from the uncertainty, it is not possible to distinguish if uncertainty is small due to an observed frequency near 1 or near 0. Nevertheless, there were significantly different days during the period as already seen in the last chapter. They were characterised by a very large (i.e. 09 (sec. 3.1) or 16 August 2007) or a very small uncertainty (12 (sec. 3.2) or 13 August 2007).

Also significant changes in uncertainty in short time intervals (i.e. 12 (sec. 3.2) or 15 August (sec. 3.3)) were further investigated.

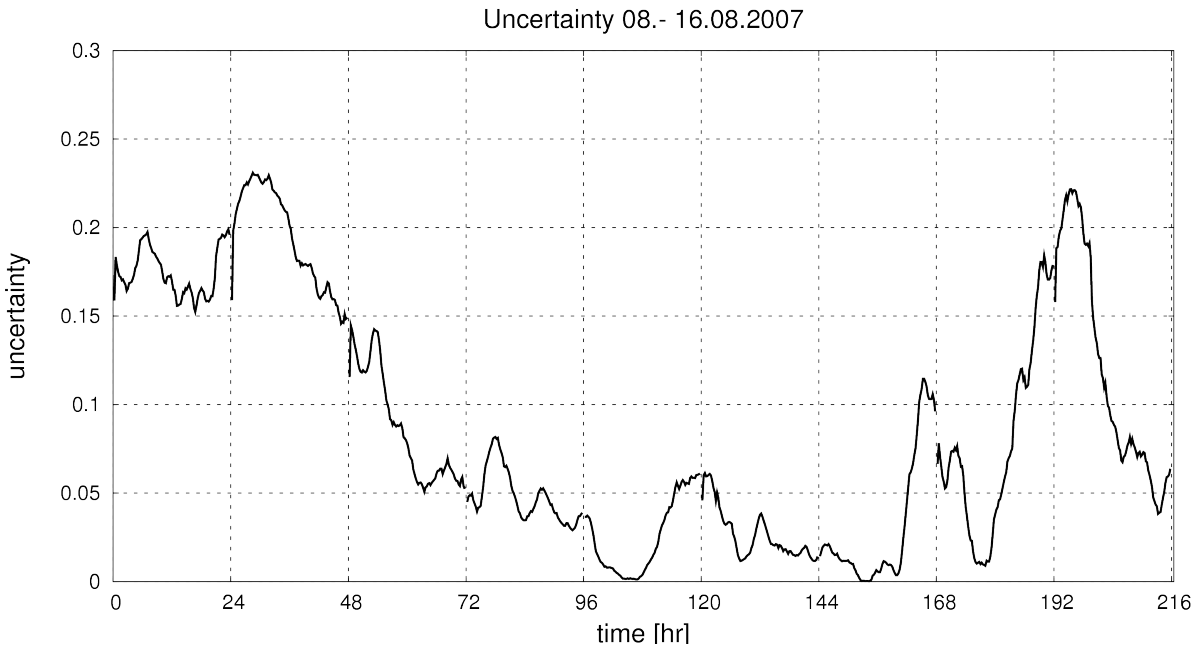

Figure 4.1: Uncertainty component of the Brier score from 8 to 16 August 2007.

4.2 Quality of probabilistic Rad-TRAM

Figure 4.2 shows the performance of the probabilistic radar tracker Rad-TRAM as time series of the Brier score (top), the CSRR (middle), and the area under the ROC curve (bottom) over the entire period from 8 to 16 August 2007.
The Brier score (Fig. 4.2, top) gives information about the magnitude of the forecast error. During the entire investigated period, the Brier score has a large variability. In comparison to the uncertainty component (Fig. 4.1), a similar shape of the curves can be seen. The deviation of short lead times from uncertainty is an indication for their skill in reliability and resolution. Hence, the performance of the forecasts in terms of the Brier score is highly dependent on the observed relative frequency of the event. A clear ranking following lead times can be identified on all days. The number of distinguishable lead times varies and seems to be dependent on the meteorological regime. If more than the first forecast hour is distinguishable the loss in skill is largest in this first hour.

The use of the CSRR (Fig. 4.2, middle) has the advantage that the exact values of various days can be compared more reliably. The score is independent of the observed frequency as it is weighted with the size of the rain domain (sec. 2.4). Therefore, the variability within the days is smaller. The quality of the forecasts and the rate of loss decreases with increasing lead time. Also in the CSRR, the number of distinguishable forecasts might be taken as indication of the predictability of the situation. The more lead times are distinguishable, the larger the amount of precipitation and the better the description of the development of the precipitation field by advection. Large outliers in the CSRR are a sign that there was almost no precipitation in the domain and should not necessarily be taken as a very large error of the forecast (i.e. 14 August 2007).

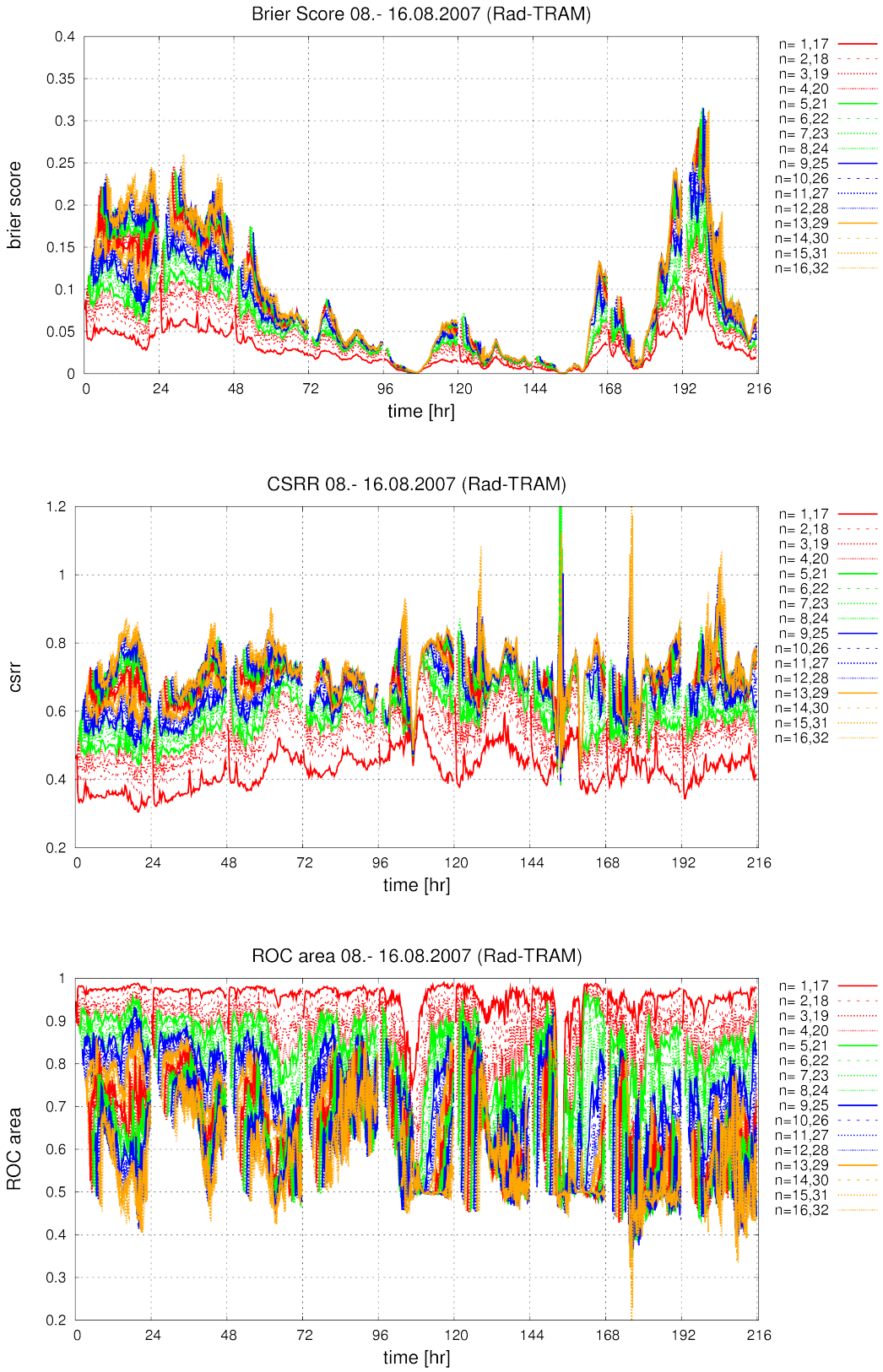

Figure 4.2: Development of Brier Score, CSRR, and area under ROC curve for Rad-TRAM based probabilistic forecasts from 8 to 16 August 2007. Colour-coded as explained in Fig. 3.7.

The area under the ROC curve (Fig. 4.2, bottom) varies for the different lead times and days over the entire possible range of values. The skill of the forecasts concerning discrimination is in the first forecast hour very high (between 0.9 and 1.0), except some outliers (i.e. 12 and 14 August). The ROC area has a clear ranking in quality of the forecasts following lead time. But longer lead times are not necessarily well ordered (8 August), but might show a larger variability (14 August). On eight of the nine investigated days, values of longer lead times are below the 0.5 threshold. Then, the forecasts do not have any skill in discrimination if the event occurs or not, but are even anticorrelated.
Comparing all three scores, there are some similarities. All scores show a ranking following lead time with the forecasts based on the latest observations having more skill than the older. All scores show in agreement that the loss of skill is largest in the first forecast hour. After the first hour, it depends on the meteorological situation if the differences between the lead times can be distinguished or if they are well ordered. The number of distinguishable lead times can be seen as an indication for the predictability. In this overview, the results the scores provide are consistent. That means, a forecast with high skill in the Brier score is also good in CSRR and the area under the ROC curve. Nevertheless, there are some exceptions to this finding. For example, on 12 August 2007 around noon, the Brier score and the CSRR show a rapid increase in forecast skill, whereas the area under the ROC curve shows a decrease (sec. 3.2).

Reliability diagrams display the full distribution of forecasts and observations in terms of the refinement calibration distribution (sec. 2.4). The reliability diagram consists of the so-called calibration function and the refinement distribution (small diagram in the top left of each plot). The refinement distribution is shown on a logarithmic scale (Fig. 4.3). The 15 minutes nowcasts have very high skill in reliability as the calibration function is very close to the diagonal (Fig. 4.3, top left). The aspect of resolution is well represented by the forecasts as well as the distance to the horizontal no resolution line is large. The histogram shows that the forecasts are relatively sharp, because the extreme bins are two mostly populated bins. As reaching 19 dBZ is a relatively rare event, the 0 % bin is mostly populated.
The forecast with 2.25 hours lead time also has very high skill (Fig. 4.3, top right). Reliability and resolution are still large. But sharpness has already decreased as can be seen by the lower population of the bins near 1.0. This is also visible in the calibration function that already is lower than the prefect reliability line for forecasted probabilities of 0.9 and 1.0. Nevertheless, forecasts of all categories are skillfull as they are far above the no skill line.
The reliability diagram of the 4.25 hours forecast (Fig. 4.3, middle left) shows that for this lead time, not all bins are populated. This means, the 1.0 forecast is never issued at this lead time. The other forecast categories have a high reliability and resolution except the 0.0 bin. It can be seen that it is above the no skill line, indicating that if 0.0 was forecasted there was an observed frequency of nearly 0.05.
In the reliability diagram of forecasts with a lead time of 6.25 hours (Fig. 4.3, middle right) a decrease in skill can be seen. Although the forecasts are still above the no skill line, their distance to the perfect reliability line is for bins larger 0.3 larger than to the no skill line. The high bins are rarely populated resulting in a further loss of sharpness.
The forecasts of a lead time of almost 7.75 hours (Fig. 4.3, bottom) show further decreased skill. The forecasts up to 0.7 are very close to the no skill line. This indicates that they hardly have skill concerning reliability and resolution. The high bins are if populated only sparsely populated but their calibration function is above the no skill line.
In summary, Rad-TRAM forecasts of different lead times show high skill in reliability dia-

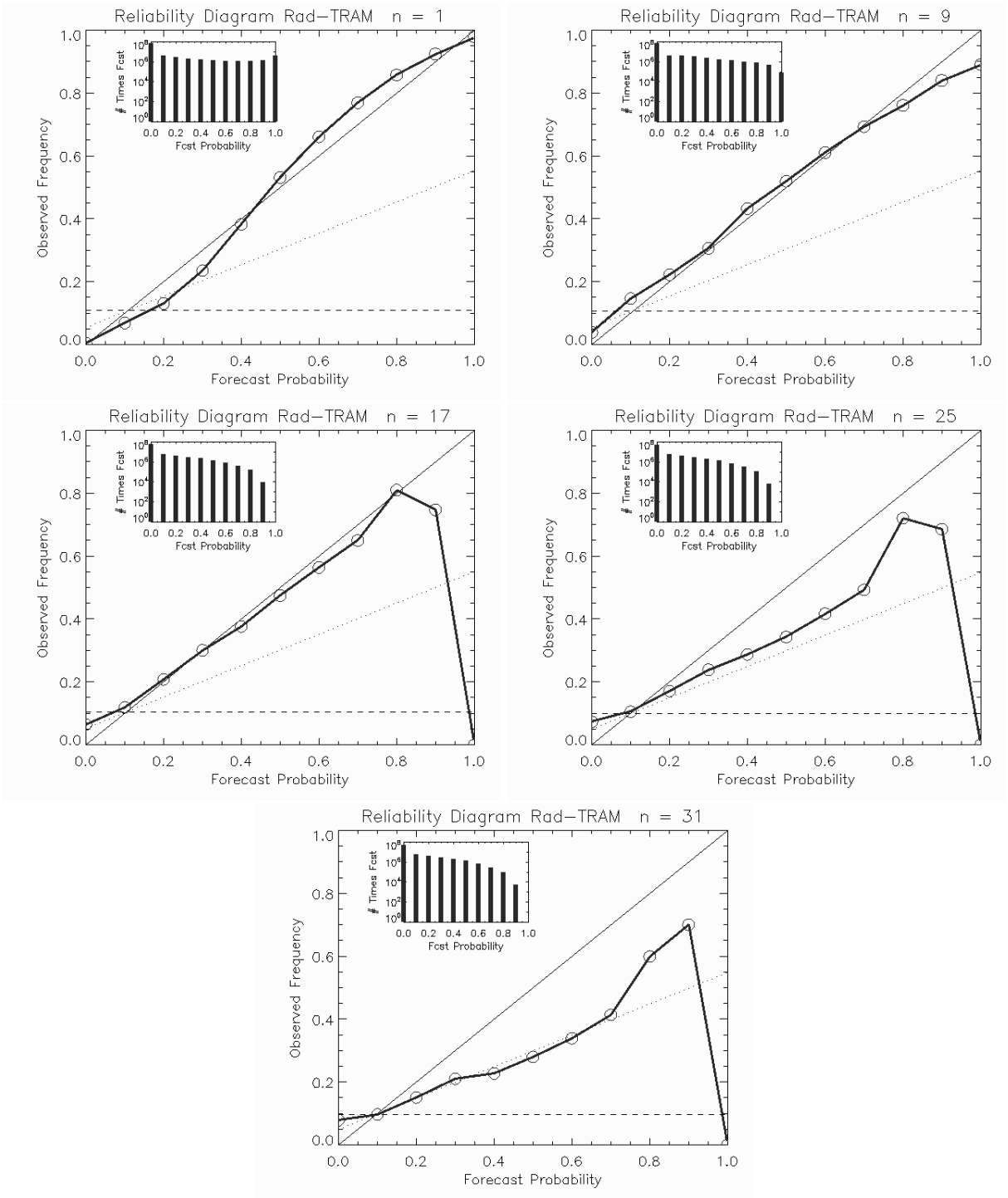

Figure 4.3: Reliability diagram for Rad-TRAM based probabilistic forecasts from 8 to 16 August 2007, for forecasts with lead times of 15 minutes (n=1), 135 minutes (n=9), 255 minutes (n=17), 375 minutes (n=25), and 465 minutes (n=31) with the perfect reliability (solid), the no resolution (dashed) and the no skill line (dotted).

grams. Even the long forecasts are skillful and the short lead times have very high skill in sharpness, reliability, and resolution.

4.3 Quality of COSMO-DE-EPS

In this section, as well the uncalibrated and the calibrated probabilistic forecasts derived from COSMO-DE-EPS output will be evaluated. Analogous to the forecasts based on observations, time series of the Brier score, the CSRR, and the area under the ROC curve and reliability diagrams are shown. Furthermore, the mean skill as evaluated for the comparison with Rad-TRAM will be compared to investigate the effect of calibration.

4.3.1 Quality of uncalibrated COSMO-DE-EPS probabilistic forecasts

Figure 4.4 shows the time series of the Brier score, the CSRR, and the ROC area of the uncalibrated probabilities based on COSMO-DE-EPS. The Brier score reveals a large variability within the different days (Fig. 4.4, top). As already seen in the single case studies with the calibrated probabilities, the shape of the curves is similar to the uncertainty (Fig. 4.1). Again, this is an indication that the relative frequency of occurrence of 19 dBZ in the observations strongly influences the forecast quality in Brier score. Compared to the daily variations, the spread within the different forecasts is small. Nevertheless, the spread varies dependent on the meteorological situation (large on 9 August, small on 11 August). It is hardly possible to identify a general ranking for the different methods in this representation. Note that the fraction method behaves differently than the other solutions. During the first forecast hours of each run, it performs significantly worse than the others. This is due to the fact, that spread within the different methods needs some integration time to develop. Depending on the definition of fraction method, the forecasts are very sharp if spread is small. The sharper the forecasts are the larger the possible error in comparison to the observations. The CSRR shows a different variability than the Brier score as the effects of the relative frequency of the event in the observations are eliminated (Fig. 4.4, middle). The development of skill on the various days is more comparable. But again, the variability depending on the meteorological situation is larger than the spread within the different methods. There are two outliers (14 and 15 August) that are very large (cf. Fig. 4.2). They result from the weighting with a very small rain area together with an overestimation in the forecasts. Again, the fraction method is the only method that can clearly be separated from the others in this time series.
The area under the ROC curve shows the largest variability within the methods (Fig. 4.4, bottom). Here, differences within the neighbourhood members can be identified (i.e. on 15 August, sec. 3.3). The forecasts vary over a large range of the score, but rarely exceed 0.8. This only happens in the very first forecast hours (except on 14 August). On 7 out of the 9 days, at least some of the members fall beneath the no discrimination threshold of 0.5. On 9 August (sec. 3.1), the values even fall beneath 0.4 indicating anticorrelation of the forecasts and the observations. Taking 0.7 as threshold for skill in discrimination (Buizza et al., 1999), it could be summarised that COSMO-DE-EPS forecast hardly have skill in discriminating if 19 dBZ occurs in the synthetic radar reflectivity field or not.

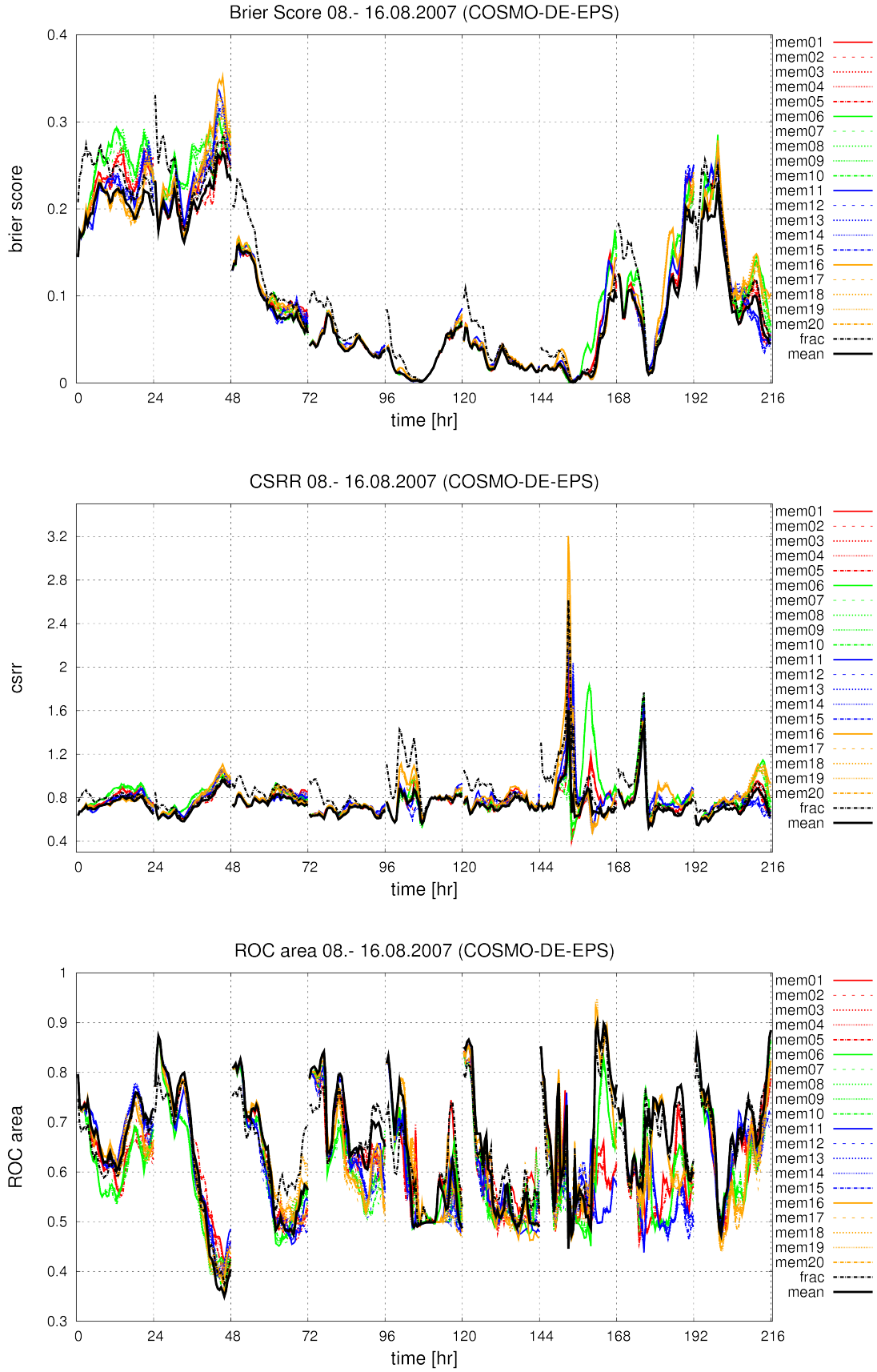

Figure 4.4: Development of Brier Score, CSRR, and area under ROC curve for forecasts based on uncalibrated probabilities derived from COSMO-DE-EPS from 8 to 16 August 2007. Colour-coded as explained in Fig. 3.7.

Reliability diagrams are shown for the fraction method, the first member as representative of the ensemble as the differences between the members are small, and the mean of the neighbourhood members (Fig. 4.5). The reliability diagram for the probabilities derived from COSMO-DE-EPS with the fraction method (Fig. 4.5, top left) shows a sharp refinement distribution. The lowest and the highest bin are most populated. The calibration function is relatively flat indicating low skill in resolution. The difference to the diagonal is large so that very low skill in reliability can be concluded as well. The low skill in reliability and resolution can be derived from the fact that the calibration function is below the no skill line as well.

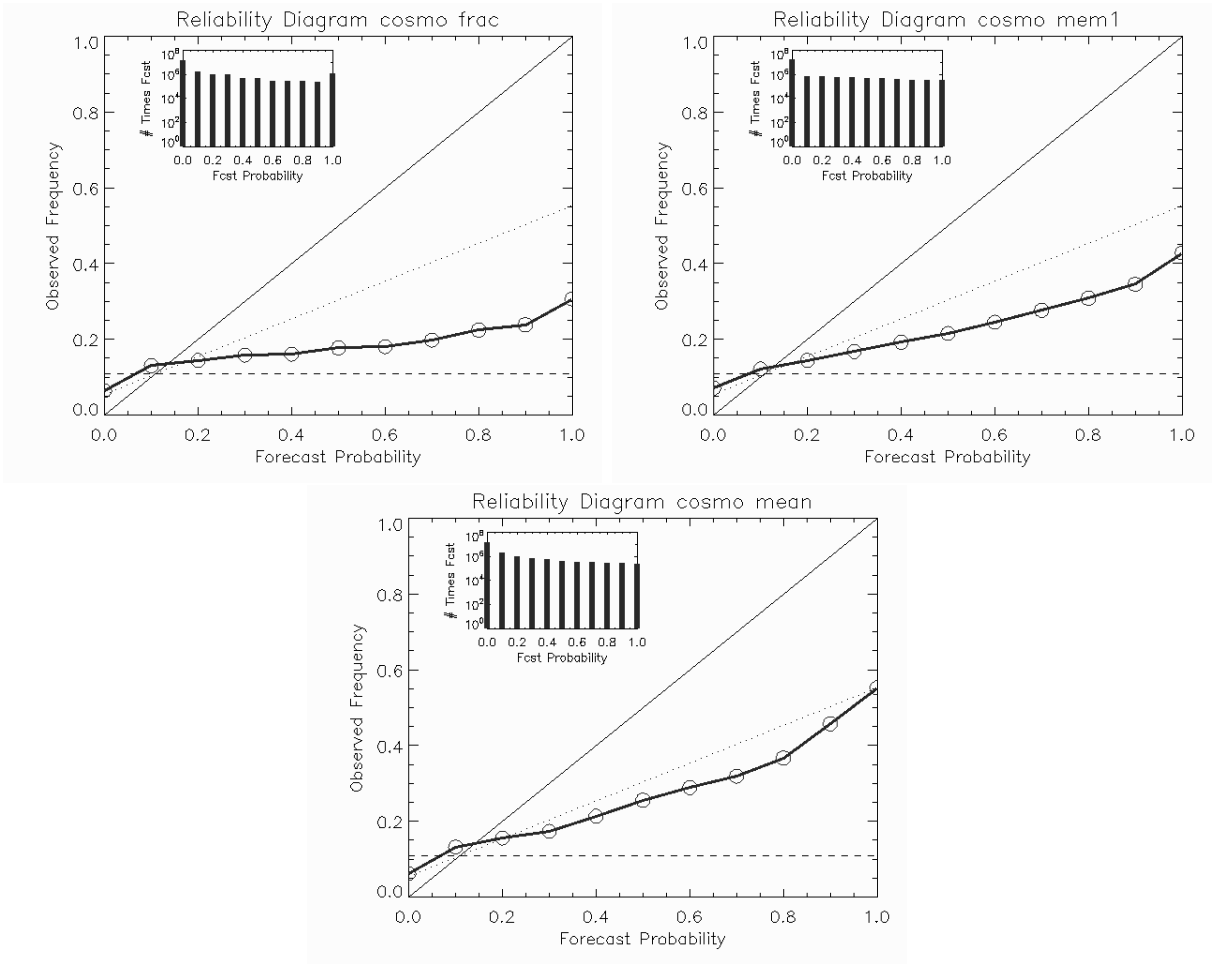

Figure 4.5: Reliability diagram for uncalibrated COSMO-DE-EPS forecasts based on the fraction method, member 1 as representative of the neighbourhood members, and the mean of the neighbourhood members from 8 to 16 August 2007. Lines as defined in Fig. 4.3.

The reliability diagram of forecasts based on member 1 reveals a different performance (Fig. 4.5, top right). The skill in sharpness is smaller as all bins, except the 0.0, are almost equally populated. The gradient of the calibration function is higher than for the fraction method, so there is more skill in resolution. Nevertheless, the difference to the diagonal is large and the no skill line is large and therefore, the overall skill is low.
The reliability diagram of the mean method (Fig. 4.5, bottom) looks most skillful in terms of the calibration function in comparison to the fraction method and member 1. The lowest and the highest bins are in the range of the no skill line. The bins within are lower indicating

underforecasting, but their difference to the no skill line is smaller than of the fraction method and member 1. But concerning the sharpness the mean method is inferior as the histogram shows a lower population on higher bins. In summary, this suggests that the quality of the uncalibrated COSMO-DE-EPS forecasts concerning reliability and resolution is low in comparison the Rad-TRAM even for long lead times. The forecasts based on COSMO-DE-EPS are often affected by location errors and false alarms as the observed frequency is always significantly lower than the forecasted probability.

4.3.2 Quality of calibrated COSMO-DE-EPS probabilistic forecasts

Figure 4.6 shows the performance of the calibrated probabilities (sec. 2.5) derived from COSMO-DE-EPS as time series of the Brier score, the CSRR, and the area under the ROC curve. The development of forecast skill with time in terms of the Brier score (Fig. 4.6, top) again reflects the shape of the uncertainty (Fig. 4.1) as already seen for the uncalibrated probabilities. Therefore, the general performance is similar to the uncalibrated forecasts (Fig. 4.4, top). But the difference between the solutions (spread) is reduced. As spread in Brier score was already small for the uncalibrated forecasts, in this representation hardly a difference can be identified (e.g. 13 August). The difference between the fraction method and the other methods in the first forecast hours every day is reduced as well.
The CSRR of the calibrated forecasts still has a larger daily variability than spread (Fig. 4.6, middle). The amplitude of the values and especially of the outliers (e.g. 14 August) is reduced. As well in CSRR, the spread is significantly reduced through calibration. The fraction method is inferior to the other solutions in the first forecast hours, but the deviation to them is smaller.
The area under the ROC curve is hardly changed through calibration as this score describes discrimination (Fig. 4.6, bottom). Still, the values vary largely over the course of the day with no very good values. On seven days, single members have no skill in discrimination at least during some hours. On one day (9 August), they are even anticorrelated with values below 0.4 at the end of the day.

The reliability diagrams of the calibrated COSMO-DE-EPS probabilities is shown in Fig. 4.7 for the fraction method, member 1 as representative of the neighbourhood members, and the mean method. The refinement distribution of the fraction method shows that after calibration, not all bins are populated (Fig. 4.7, top left). The highest forecasted probability is 50 % and the 30 % bin is not populated. The new extremes (0.0 and 0.5) are highly populated indicating skill in sharpness. The calibration function as well shows that not all bins are used after calibration. The comparison with the uncalibrated reliability diagram (Fig. 4.5, top left) reveals that the skill is increased as the difference to the no skill line is reduced. As the well in reliability as in resolution, skill is increased.
The reliability diagram of member 1 looks slightly different (Fig. 4.7, top right). Here, all bins up to 0.5 are populated. The 0.0 category contains most of the values, whereas the others vary indicating a low sharpness of the forecasts. The calibration function in the diagram shows skill in all used forecast categories except in 0.0 that was already above the no skill line in the uncalibrated forecasts. The difference to the diagonal is small so that the forecasts have a high skill in reliability. As the gradient is high, the skill in resolution is high as well. The comparison with the uncalibrated forecast (Fig. 4.5, top right) clearly shows

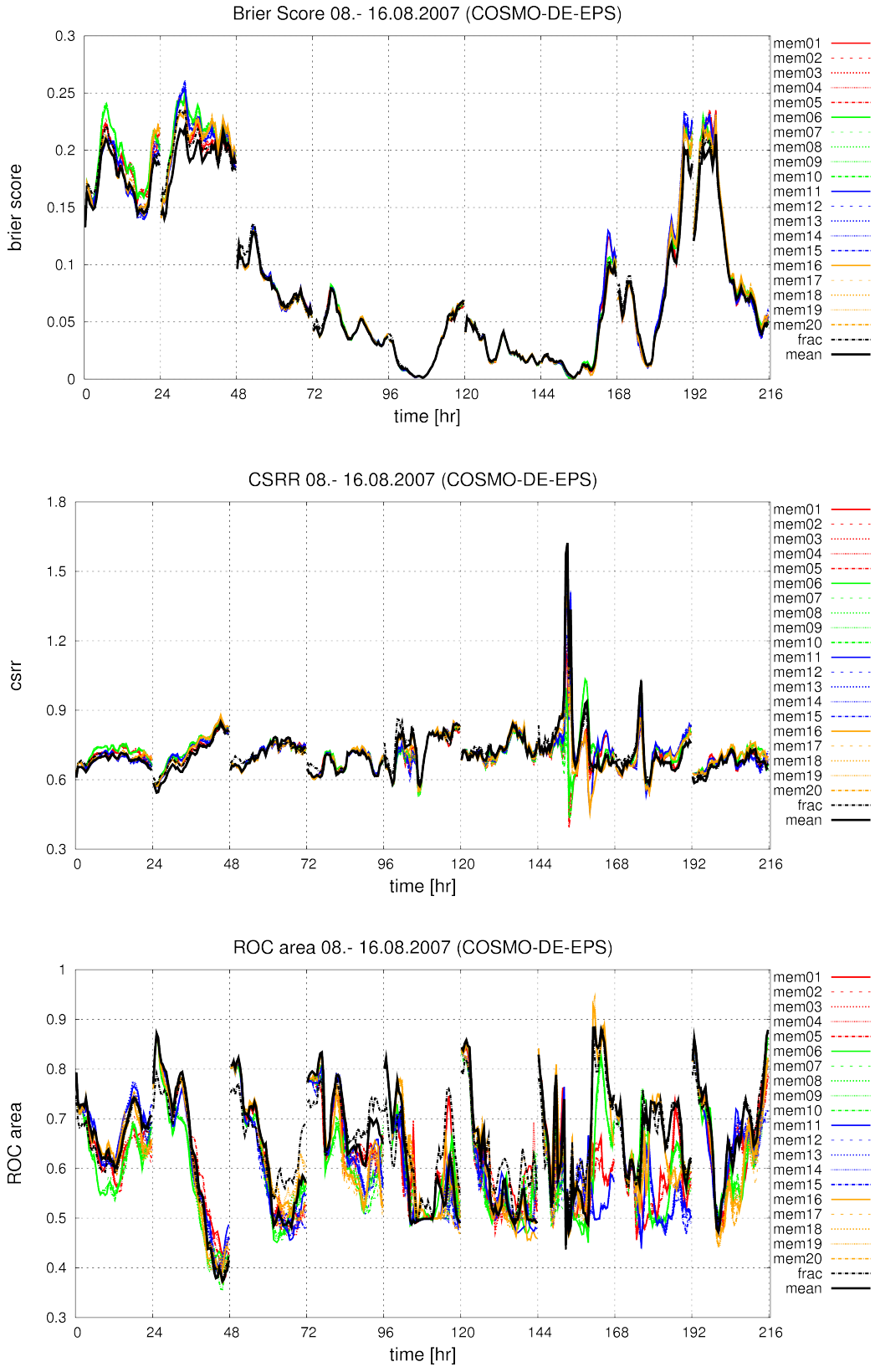

Figure 4.6: Development of Brier Score, CSRR, and area under ROC curve for forecasts based on calibrated probabilities derived from COSMO-DE-EPS on 8 to 16 August 2007. Colour-coded as explained in Fig. 3.7.

the positive effect of the calibration.
The reliability diagram of the mean method shows as already seen with the fraction method that not all forecast categories are used (Fig. 4.7, bottom). The 0.1 and 0.6 bin are not populated due to rounding (cf. Fig. 2.12). The calibration function of the other categories indicate skill as well in general and specifically in reliability and resolution. Also in the mean method, the calibration procedure clearly improved the overall skill of the forecasts although not all bins are populated.
Concluding, it is seen that the calibration is successful in the sense of improving the reliability component without significantly changing the resolution of the forecasts. But the cost is, that the bins are differently populated and spread between the methods is further reduced as the calibration functions are similar or even the same for all neighbourhood members (sec. 2.5).

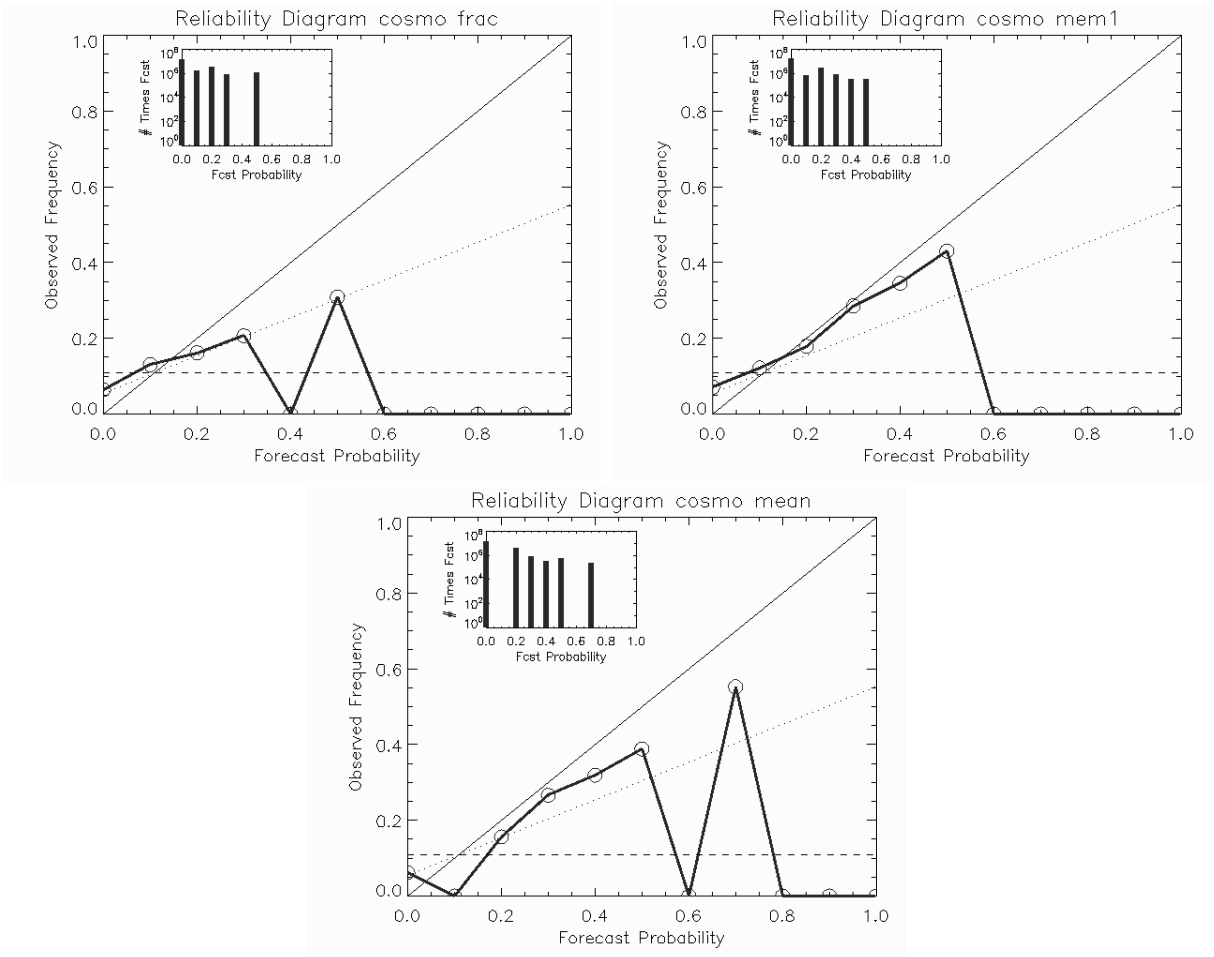

Figure 4.7: Reliability diagram for calibrated COSMO-DE-EPS forecasts based on the fraction method, member 1 as representative for the neighbourhood members, and the mean of the neighbourhood members from 8 to 16 August 2007. Lines as defined in Fig. 4.3.

4.3.3 Effect of Calibration on quality of COSMO-DE-EPS probabilities

Figure 4.8 shows the mean and the standard deviation over the entire period for uncalibrated and the calibrated probabilities of COSMO-DE-EPS as described in Fig. 3.1. The change in

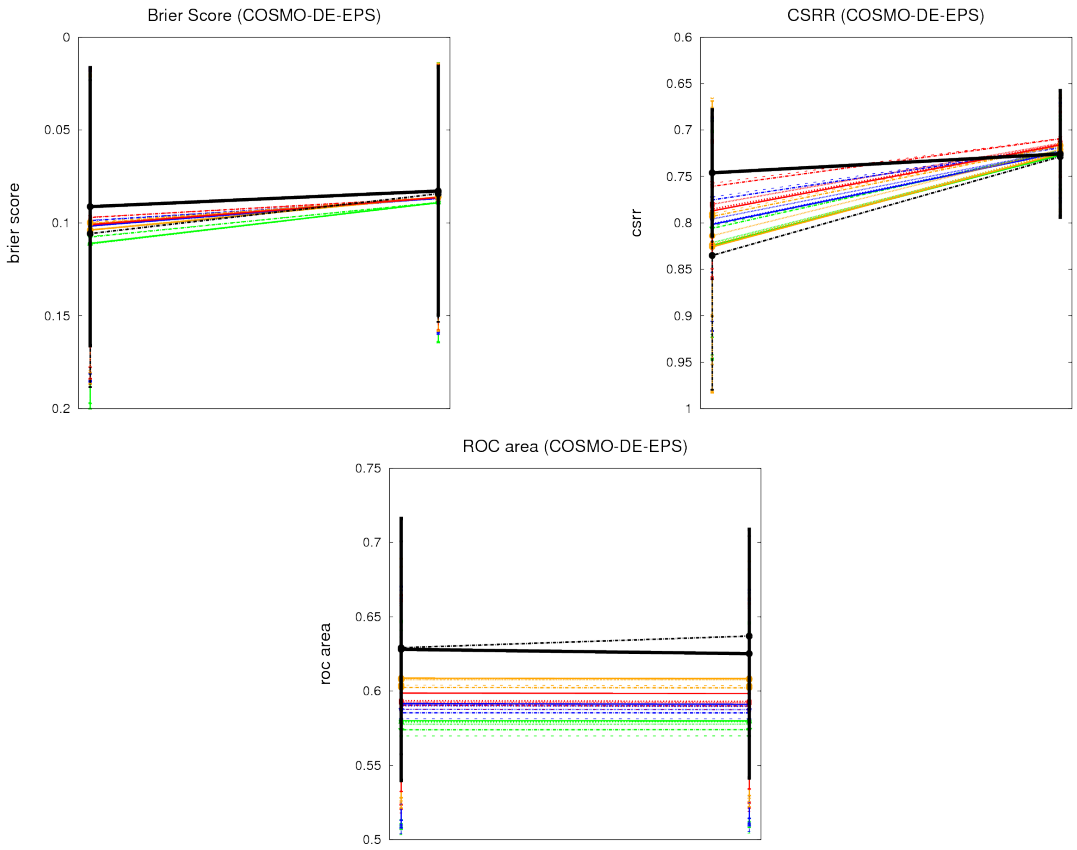

Figure 4.8: Effect of calibration of COSMO-DE-EPS probabilities in Brier score, CSRR, and area under ROC curve (left uncalibrated, right calibrated). Colours and lines as defined in Fig. 3.9.

the mean performance summarises the effect of calibration on the mean skill of the respective method over the entire investigated period.
The effect on the Brier score (Fig. 4.8, top left) is seen as an improvement (reduction) of the mean values (Fig. 4.8, top left). The spread within the different methods is reduced as well. The effect differs from method to method and therefore, the ranking is changed through calibration. Before calibration, the mean method clearly outperformed the neighbourhood members and the fraction method. The neighbourhood members were ranked with the physical perturbations 1 (entrainment rate of shallow convection) and 5 (asymptotic mixing length of turbulence scheme) of the ECMWF and NCEP driven neighbourhood members outperforming the others. After calibration, the differences are very small and the specific neighbourhood members cannot be distinguished. They are worse than the mean and the fraction method.
With the CSRR (Fig. 4.8, top right) the largest effect of calibration can be seen. As well the spread of the mean values as their variability before calibration is larger. Again, the effect of calibration is different on the respective methods and therefore, the ranking is changed. Before calibration, the mean of the neighbourhood members have highest skill. The neighbourhood members follow and are ranked following physical perturbation 1 and 5 and the lateral boundary conditions from ECMWF and NCEP. The fraction method clearly has the lowest skill. After calibration, the ECMWF and NCEP members have lower errors than the mean and the fraction method. As well the differences as the variability of the mean values is reduced.

The area under the ROC curves is hardly affected by calibration (Fig. 4.8, bottom). Only the fraction method is slightly sensitive to the calibration. The ranking and the variability is not changed. The fraction method has higher skill in discrimination than the mean method and the neighbourhood members. They are ranked following global models with the UKMO members leading.
All in all, the effect of calibration in Brier score and CSRR can be summarised as a reduction of spread and an increase in skill. The methods are effected with different intensities resulting in a changed ranking with the fraction method being most sensitive.

4.4 Comparison of performances of Rad-TRAM and COSMO-DE-EPS

The development of forecast skill of Rad-TRAM and calibrated COSMO-DE-EPS forecasts with nowcast lead time over the entire investigated period as derived following Fig. 3.1 is displayed in Fig. 4.9. As in the case studies, the representation is chosen such, that highly located values in the figure represent high skill. This means, the axis of the negatively oriented scores (Brier score and the CSRR) have been switched.

Rad-TRAM's mean skill in Brier score (Fig. 4.9, top left) decreases with lead time. In the first three hours of forecast time, the rate of decrease is faster than later. The standard deviation (thin black solid lines) as a measure for the variability of the mean value shows as already seen in the investigation of single days a very large variability of the mean value at each lead time. It is larger than the variability or decrease of the mean values with lead time. The forecasts based on the calibrated COSMO-DE-EPS are constant due to the evaluation set up (Fig. 3.1). Their skill is very similar to each other. Nevertheless, a ranking can be identified with the forecasts based on the mean and fraction method having smaller errors than the forecasts based on the neighbourhood method. The variability of all mean model forecasts is also very large and in the range of Rad-TRAM's variability. The time frame when the COSMO-DE-EPS forecasts start having smaller errors than Rad-TRAM (cross-over time) is between 4.75 hours (mean method) and 7 hours (member 6).

The performance with the CSRR (Fig. 4.9, top right) as well shows a rapid decrease of skill for Rad-TRAM, especially in the first three hours. The variability of the mean value is significantly smaller than with the Brier score. It is also smaller than the decrease of the mean value. The different methods for the derivation of probabilities from COSMO-DE-EPS have more spread than in Brier score. The members based on the ECMWF (1-5) have more skill than the other neighbourhood members. The mean and the fraction method have the smallest skill. Also in the forecasts based on the COSMO-DE-EPS, the standard deviations are smaller than in Brier score. The cross-over points are between 5.5 hours (member 5) and 7 hours (fraction method).
The area under the ROC curve (Fig. 4.9, bottom) shows that Rad-TRAM probabilities vary over the entire range of the possible values of the skill score. But on average, they are larger than 0.7 up to three hours lead time and always larger than 0.5. The rate of decrease of skill is higher in the first four forecast hours and then gets slower. In ROC area, the COSMO-DE-EPS forecasts have the largest spread between the 22 solutions in comparison to the other quality measures. The variability of the mean as seen in the standard deviation is relatively small. The fraction and the mean method have larger skill in discrimination

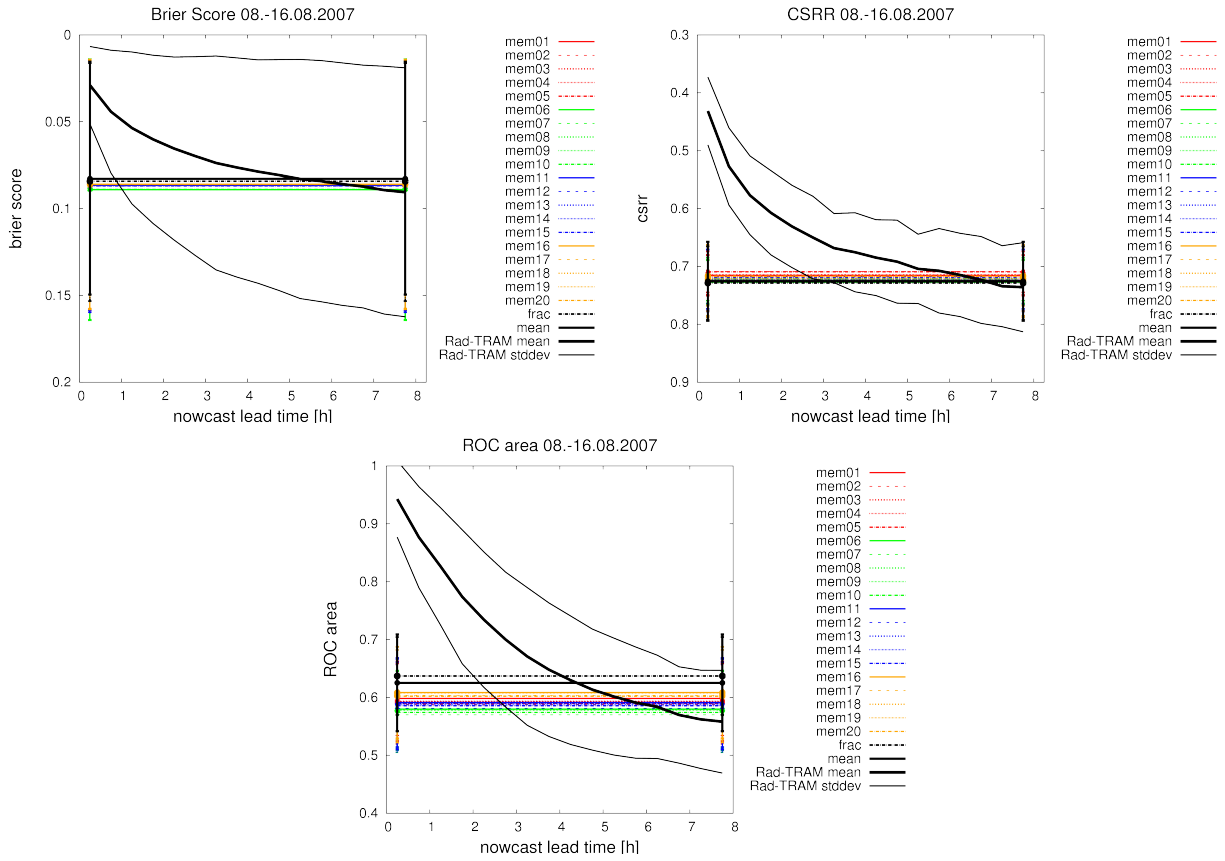

Figure 4.9: Development of Brier score, CSRR, and area under ROC curve with lead time for Rad-TRAM and calibrated COSMO-DE-EPS forecasts from 8 to 16 August 2007. Lines as defined in Fig. 4.3.

than all neighbourhood members. The neighbourhood members based on the lateral boundary conditions of DWD (6-10) have the lowest skill. As the spread is large, the time frame in which COSMO-DE-EPS becomes better than Rad-TRAM is large as well: 4 (fraction method) to 7 hours (member 7). Remarkably, as this was not necessarily seen in the case studies, all probabilities based on the model are with respect to their standard deviation larger than the no discrimination threshold of 0.5.
All three scores show in agreement a loss of forecast skill with lead time for Rad-TRAM. The mean model performance is significantly worse in the first forecast hours, but latest after 6 hours in all scores COSMO-DE-EPS has more skill than the nowcaster. Regarding the variability, an earlier or a later cross-over point is possible as well. The differences between the two forecasting methods are smaller for long lead times as they were for short lead times.

The mean performance of Rad-TRAM in terms of the CSRR (Fig. 4.9, top right) is the basis for the definition of the weighting functions and the calculation of the blended probabilities discussed in the next chapter.

4.5 Discussion

The investigation of the skill over the entire investigated period in this chapter was revealed as in the previous section with the case studies with timeseries and the development of skill

with lead time. Furthermore, reliability diagrams were evaluated to quantify the skill in reliability, resolution, and sharpness.
Some findings of the case studies were confirmed. Concerning Rad-TRAM, on all days the forecasts based on the latest observation have the smallest errors in Brier score and CSRR and the highest skill in discrimination. If longer lead times are distinguishable, the rate of decrease is larger in the first four hours than later. The evaluation of the entire COSMO-DE-EPS time series showed that even with the uncalibrated forecasts, the differences between the different solutions derived with the three methods were small. The calibration procedure even reduced these differences. The model forecasts do not reach the high skill of the nowcaster at short lead times.
The investigation of reliability diagrams showed very high skill of Rad-TRAM forecasts, even for lead times of 8 hours. The COSMO-DE-EPS forecasts showed skill at least after calibration. But as the sample size is limited, after calibration not all bins are populated. Here would be room for improvement if more data was available.
The lead time dependent evaluation showed that overall, forecasts cross around 6 hours. The variability reflected the findings of the case studies that this value depends on the meteorological situation. The effect of calibration in Brier score and CSRR is a reduction in spread and error. Therefore, with the uncalibrated forecasts the cross-over points would have been even later.

Chapter 5

Blending of probabilistic forecasts from Rad-TRAM and COSMO-DE-EPS

In the previous chapters, the quality of the probabilistic forecasts based on Rad-TRAM and COSMO-DE-EPS was evaluated systematically with different quality measures and in different evaluation set ups in case studies and over the entire period. Now, on basis of the knowledge of the overall skill and the development of skill with lead time, the two probabilistic forecasts will be combined. The goal is to create a probabilistic forecast that combines the skill of the forecast sources such that a seamless and an optimal forecast at lead times from 0 to 8 hours is created. The blended forecasts should combine the probabilities such that they represent a seamless transition from one forecast field to the other under consideration of the strength and weaknesses of the methods at the specific lead times. This means, the skill of the blended forecasts should be at least as high as the respective best forecast at the different lead times.
In the first section of this chapter, existing approaches will be reviewed in a short literature overview. Then, the method applied in this study will be introduced. Finally, the results of the evaluation of the quality of the blended probabilistic forecasts will be presented and discussed for the three case studies and for the entire period. The chapter closes with a discussion of the results.

5.1 Literature overview

Various systems and approaches exist to combine forecasts based on observations and NWP models. The most common approach is an additive combination of the two forecasts. The basis is the knowledge about the development of their forecast skill with lead time as this development determines the weighting functions for the combination. It has to be distinguished if probabilistic or deterministic forecasts are evaluated with conventional or probabilistic quality measures. The combination can be conducted in different ways as well. It has to be distinguished which variable is combined (precipitation rate, radar reflectivity, or probabilities of exceeding a precipitation threshold) in which way (linear or exponential weights, stochastic noise). In the following, the most important approaches are described briefly.

As first, the UK MetOffice evaluated the quality development of precipitation forecasts with lead time systematically. The investigation resulted in the development of Nimrod (Golding, 1998). They used deterministic quality measures (RMS, RMSF, FAR, POD, and CSI) to evaluate the performance of a radar data based extrapolation technique and NWP fields (Unified Model (UM)) for hourly accumulated rainfall. The blended forecast is produced by calculating dynamic weights with correlation coefficients derived from the forecast quality of the previous hour (Golding, 2000). The advection forecast correlation is assumed to fall exponentially whereas the model correlation is assumed to increase steadily to a maximum of 0.7. The combined forecast has more skill than each single forecast separately at every lead time.

STEPS (Bowler et al., 2006) has been developed as a joint project of the MetOffice and the Australian Bureau of Meteorology and follows a different approach. Three so called cascades are blended in terms of accumulated rainfall to produce an ensemble of possible future scenarios: nowcasts based on SPROG (Seed, 2003), downscaled NWP forecasts from UM and stochastic noise dependent on scale. STEPS provides deterministic (best guess advection) and probabilistic products (probability of precipitation). As well probabilistic quality measures (Brier skill score and area under the ROC curve) as conventional scores (bias) showed that STEPS forecasts have skill even for different precipitation thresholds.

At NCAR, two different approaches are under development. First, in the project NIWOT (Wilson and Xu, 2006) several configurations are tested in order to optimally combine high reflectivities based on radar observations and model forecasts deterministically for forecasts of thunderstorms.
In a second group, probabilistic forecasts are combined in order to consider the inherent uncertainty of both methods (Pinto et al., 2006) for aviation applications (COSPA[1]) (Wolfson et al., 2008). Here, the Local Lagrangian method (Germann and Zawadzki, 2004) is applied on the radar composite in NCWF-2 (Megenhardt et al., 2004) and enlarged to 6 hours (NCWF-6). As forecasts based on NWP, the RUC Convective Probability forecast (RCPF) is used that provides probabilities by applying a spatial filter similar to the neighbourhood method (Weygandt and Benjamin, 2004). Phase errors in the RCPF are corrected. The quality of the forecasts is calculated with the conventional scores like CSI and bias. The probabilities are merged similar to Golding (2000) and result in a more skillful forecast.

In Hong Kong, an advanced integrated system called RAPIDS (Rainstorm Analysis and prediction Integrated Data-processing System) was developed (Wong et al., 2009). RAPIDS consists of a nowcasting (SWIRLS) and a NWP component (NHM) and works with radar reflectivities. The system is able to produce probabilistic forecasts as the model WRF is run as time-lagged ensemble (at least during the 08FDP) with three nests (27, 9, and 3 km). The RAPIDS blending algorithm consists of three parts: i) phase error correction in model QPF; ii) correction of model intensity based on radar-based quantitative precipitation estimate; and iii) merging of model QPF with radar nowcast with a hyperbolic tangent weighting function (Li et al., 2005). A probability of precipitation can be estimated with the time-lagged ensemble as additional information about increasing uncertainty in forecast hour 3 to 6. The value of the blended forecast is shown with the conventional skill score POD .

[1]Consolidated Storm Prediction for Aviation

At the McGill University, different models (GEM/HIMAP, ETA, GEM (operational), WRF) were compared with two different MAPLE (Turner et al., 2004) versions (original MAPLE (McGill Algorithm for Precipitation Nowcasting by Lagrangian Extrapolation) and MAPLE-NOFF where small nonpredictable scales are removed) (Lin et al., 2005). In this study, the forecast quality of hourly accumulated rainfall was evaluated with conventional scores like POD, CSI, FAR, and conditional mean absolute error (CMAE). In a following study, weighting functions based on the CSI were applied on the two data sources (Kilambi and Zawadzki, 2005). It is shown that the blended forecasts have more skill than the single components in the 4-12 h period and are slightly worse than the nowcasts in the first four forecast hours.

5.2 Method for blending the probabilistic forecasts

In this study, the probabilistic forecasts of exceeding the reflectivity threshold $\mathcal{L} = 19\ dBZ$ based on Rad-TRAM and COSMO-DE-EPS are combined. The basis for the additive combination of the probabilities is the knowledge of the development of their forecast quality with lead time. In the last two chapters, this development was evaluated with the Brier score, the CSRR, and the area under the ROC curve in three case studies (Fig. 3.8, Fig. 3.16, Fig. 3.24) and for the entire period (Fig. 4.9).

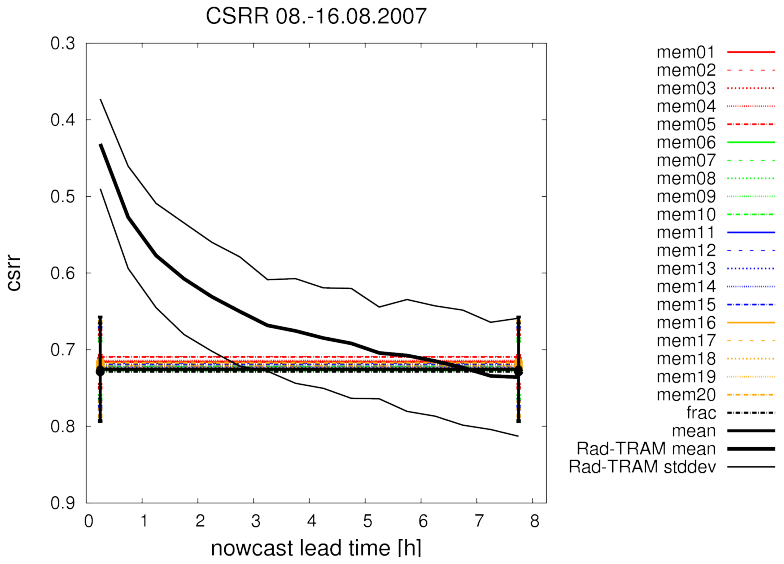

Figure 5.1: Development of CSRR with lead time for Rad-TRAM and calibrated COSMO-DE-EPS forecasts from 8 to 16 August 2007.

The development of forecast skill as evaluated with the CSRR is chosen to be the basis for the derivation of the weighting functions for the additive combination (Fig. 5.1). As discussed in detail in Chapter 4.3, the skill of Rad-TRAM forecasts decreases steadily over all lead times in all scores. The CSRR is found to be more reliable than the Brier score as it does not depend on the observed frequency of the event and has smaller standard deviations. The mean performance of COSMO-DE-EPS forecasts is significantly worse, but as Rad-TRAM's skill as well decreases to low values, after about six hours a first cross-over point can be identified. The differences between the skill of the different approaches applied on the ensemble output are small.

The weighting functions are derived similar to Kilambi and Zawadzki (2005) as their approach was simple and straight forward. They defined their weighting functions according to the performance of the respective method i at the time of the forecast t with the critical success index (CSI)

$$weight = \frac{1}{1 - CSI_{i,t}^{2.5}} - 1. \tag{5.1}$$

In their study, i denoted four different forecast types. Two were based on extrapolation (MAPLE and OMAPLE) and two on NWP models (GEM and WRF).

In this study, the weighting functions are defined based on the mean performance of Rad-TRAM with CSRR. The weight for Rad-TRAM w_r is defined depending on lead time τ as

$$w_r(\tau) = 2.11 - \frac{1}{1 - CSRR(\tau)^{2.8}} \tag{5.2}$$

and normalised to one at the first lead time. Due to the model set up for COSMO-DE-EPS forecasts, a real lead time dependent evaluation of forecast skill was not possible. Therefore, the weighting function of the model forecasts has to be defined differently. Since the combined quantity is probability of precipitation, the weights of both methods should sum to one. The weight for all COSMO-DE-EPS based forecasts w_c is calculated on basis of Rad-TRAM's weighting function

$$w_c(\tau) = 1 - w_r(\tau). \tag{5.3}$$

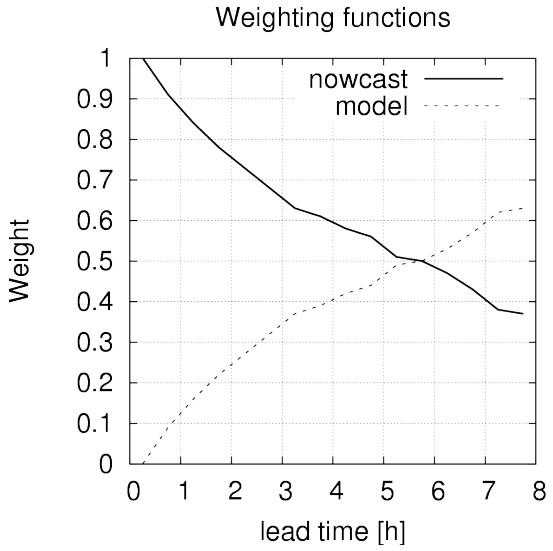

Figure 5.2: Weighting functions for the combination of probabilities based on radar extrapolation with Rad-TRAM w_r and probabilities derived from COSMO-DE-EPS w_c.

The skill of the three methods applied on the COSMO-DE-EPS output to derive probabilistic forecasts does not vary significantly or systematically. Hence, one common weighting

function is defined for all model forecasts. The resulting weighting functions are displayed in Fig. 5.2.

The weight for the extrapolation based probabilities w_r decreases steadily from one. As long lead times of Rad-TRAM might show some skill as well and the differences between Rad-TRAM's and COSMO-DE-EPS's forecast skill then are small, w_r does not fall to zero but reaches a minimum at 0.38. The cross-over point is after 5.75 hours in agreement to the findings in Fig. 4.9. This means, after this time more weight is given to the model derived probabilities. Note, the maximum weight for COSMO-DE-EPS is 0.62.

The weighting functions are multiplied to the respective probabilistic forecasts from Rad-TRAM, P_{LL}, and COSMO-DE-EPS, P_{EPS}, to combine the two probabilities at each time step in the respective eight hour interval (Fig. 3.1) to a combined probability P_{blend} according to

$$P_{blend,i} = w_r(\tau) * P_{LL}(\tau) + w_c(\tau) * P_{EPS,i} \tag{5.4}$$

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.

Figures 5.3 and 5.4 reveal examples of the combination at two different lead times. These are chosen such, that at one lead time the maximum weight is at the nowcaster and on the other it is at the model forecast. As well the components for the combination as the resulting combined probabilities are shown. For the sake of clarity, only the fraction method of the 22 model forecasts is displayed. Of course, the blending procedure results in 22 different forecasts based on the combination of Rad-TRAM with the 22 different forecasts derived from COSMO-DE-EPS output.

Figure 5.3 shows probabilistic forecasts for 12 August 2007, 23:15 UTC with lead time of $\tau = 1.25$ h. As explained in sec. 3.2, during this phase of the day, precipitation ahead of a cold front was in the evaluation domain (Fig. 3.11). At the lead time $\tau = 1.25$ h, the Rad-TRAM forecast (Fig. 5.3, top left) is multiplied by a larger weight w_r than the COSMO-DE-EPS forecast (fraction method, Fig. 5.3, top right). Therefore, the combined probability field (Fig. 5.3, bottom) reflects the high probabilities from the Rad-TRAM forecast. Nevertheless, the influence of the forecast with the fraction method is visible in additional low probabilities.

Figure 5.4 displays forecasts valid at the same time, but with a lead time $\tau = 7.25$ h. The model forecast is the same as in Fig. 5.3 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 (Fig. 5.4, bottom) is dominated by the fraction method forecast. The probabilities of both components are low, and therefore, the combined probability is low as well.
Comparing Fig. 5.3, bottom and 5.4, bottom, the decreasing influence of the nowcaster can be seen clearly. Only with the information from the blended probability field, it is not possible to deduce which forecast source leads to which pattern in the blended probability field. This illustrates that the blended forecasts deliver a seamless combination of the Rad-TRAM

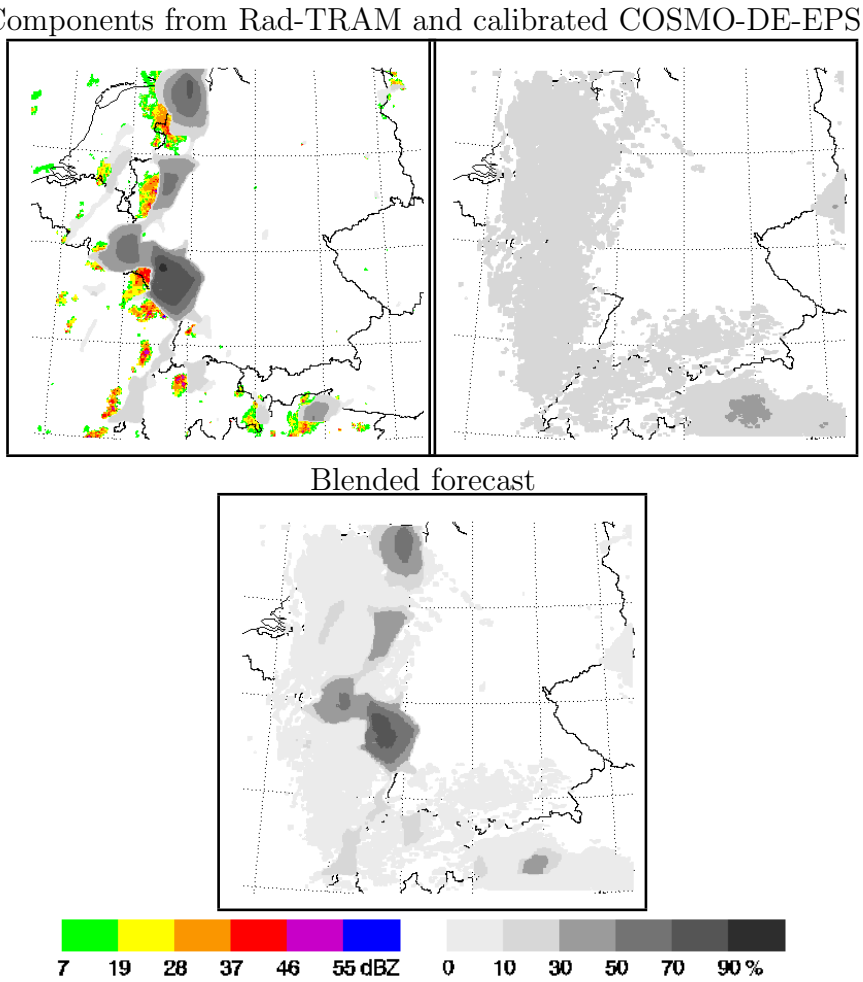

Figure 5.3: Blended probabilities for 12 August, 23:15 UTC (bottom) with components from Rad-TRAM and the calibrated COSMO-DE-EPS fraction method (top) for $\tau = 1.25$ h. Observations used to initialise the Rad-TRAM forecasts are shown in colour in the background.

and COSMO-DE-EPS based probabilistic forecasts.

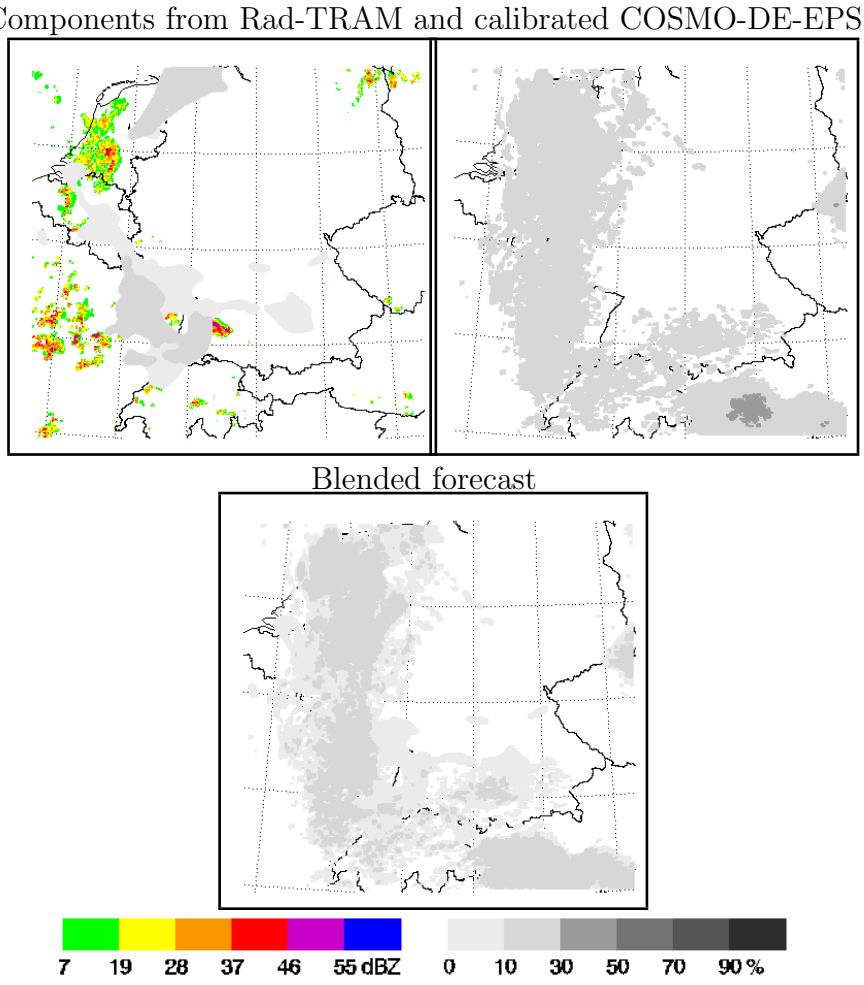

Figure 5.4: Blended probabilities for 12 August, 23:15 UTC (bottom) with components from Rad-TRAM and the calibrated COSMO-DE-EPS fraction method (top) for $\tau = 7.25$ h. Observations used to initialise the Rad-TRAM forecasts are shown in colour in the background.

5.3 Quality of blended probabilities

The quality evaluation with respect to the various scores is conducted for the blended probabilities $P_{blend,i}$ in the same way as explained in Fig. 3.1. The skill of the blended forecasts should be at least as high as the respective best single forecast at each lead time. Solely the calibrated probabilities of COSMO-DE-EPS are evaluated. First, the quality of the blended probabilities is evaluated for the three case studies of Chapter 3. Then, the overall skill over the entire period is investigated. The blended probabilities are calculated with the Rad-TRAM forecasts (P_{LL}) and the calibrated model probabilities ($P_{EPS,i}$; see Eq. 5.4). The quality of the resulting probabilities $P_{blend,i}$ is evaluated lead time dependent with the Brier score, the CSRR, and the area under the ROC curve. To investigate the differences in quality from the blended probabilities and the single components, their quality is displayed as well.

5.3.1 Skill of blended forecasts in the case studies

9 August 2007

On the 9 August 2007, large scale ascent caused heavy precipitation in large parts of the evaluation domain that lasted over the entire day (cf. sec. 3.1). On this day, the mean skill of the extrapolation based technique was higher than the model forecasts over all lead times. Only with CSRR a cross-over point was identified after eight hours (Fig. 5.5, right column). In terms of the Brier score, the skill of the blended probabilities decreases with increasing lead time (Fig. 5.5, top left). On this day, already after two hours differences between the different model solutions can be seen with the Brier score. The standard deviations are small as well. The comparison with the single components of the combination shows that the skill of the combined probabilities is in the range of Rad-TRAM's skill and higher than the one of the model solutions (Fig. 5.5, top right). This can already be seen after two hours but gets more clearly with increasing lead time. The differences between the methods applied on the COSMO-DE-EPS are slightly smaller after combination but still the same ranking can be identified with the mean and the fraction method having more skill than the neighbourhood members.
The development of the forecast skill with lead time of the blended probabilities in the CSRR shows a very high skill for short lead times (Fig. 5.5, middle left). This high skill decreases to values in the range of the model skill. The differences between the methods applied on the ensemble are smaller than in Brier score but still visible with a consistent ranking of the mean and fraction method before the neighbourhood members. The standard deviations are smaller than the decrease with lead time although they slightly increase with lead time. The performance of Rad-TRAM alone is very similar to the blended probability (Fig. 5.5, middle left). Evaluating the exact values shows that the combined values are slightly worse. This means, including the information from the model forecasts worsens the quality of the blended forecast in the last forecast hours.
Likewise, the skill of the blended forecasts as evaluated with the area under the ROC curve decreases with lead time (Fig. 5.5, bottom left). The decrease is constant with lead time. Differences between the different methods applied on the COSMO-DE-EPS can already be identified after two forecast hours. They are smaller as in the separate evaluation of the forecasts. The comparison with the single performances (Fig. 5.5, bottom right) shows that

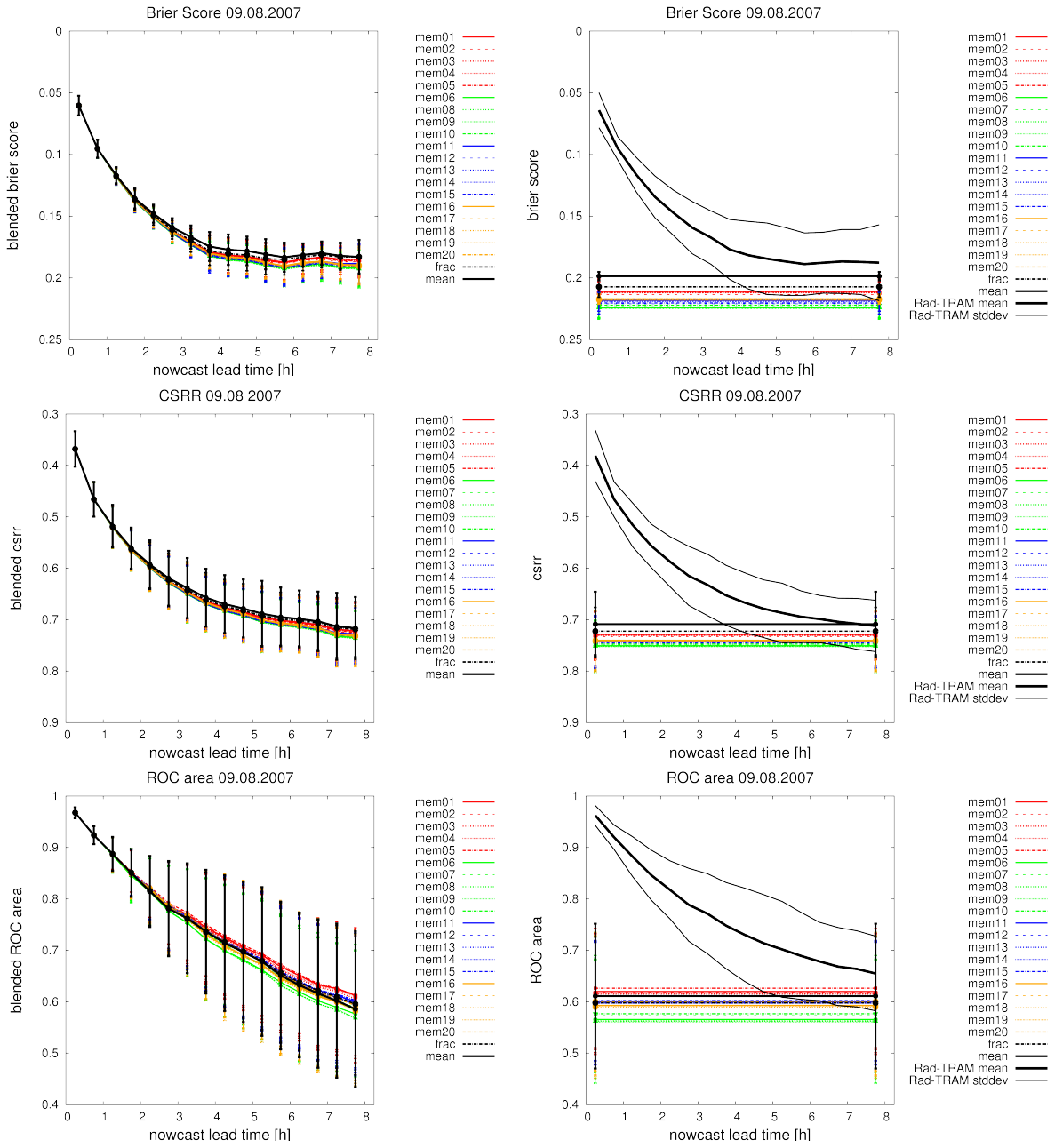

Figure 5.5: Development of Brier Score, CSRR, and area under the ROC curve with lead time for blended probabilities (left column) and the components separately (right column) on 9 August 2007. The different lines in the skill of the blended probabilities represent the combination of the different model solutions with the Rad-TRAM forecast.

without the information from the model forecasts, Rad-TRAM has more skill in discrimination. The standard deviations of long lead times increase through the combination.
To conclude, the development of skill of the blended probabilities on 9 August 2007, for short lead times, in all scores the high skill of Rad-TRAM forecasts and their rapid decrease is reproduced. But for long lead times, the combination with the clearly worse COSMO-DE-EPS forecasts worsens the skill of the blended probabilities. Nevertheless, the differences between the blended and the Rad-TRAM probabilities are small.

12 August 2007

The 12 August 2007 was dominated by the passage of a weak cold front in the early evening (cf. sec. 3.2). This regime change was characteristic for this day. Before the arrival of the frontal precipitation, only a small amount of rain was detected in the early morning. On 12 August, the development of skill in Rad-TRAM and COSMO-DE-EPS varied in the three scores. Therefore, the cross-over periods were different as well. They ranged from three (first cross-over in ROC area) to eight hours (last cross-over in CSRR).

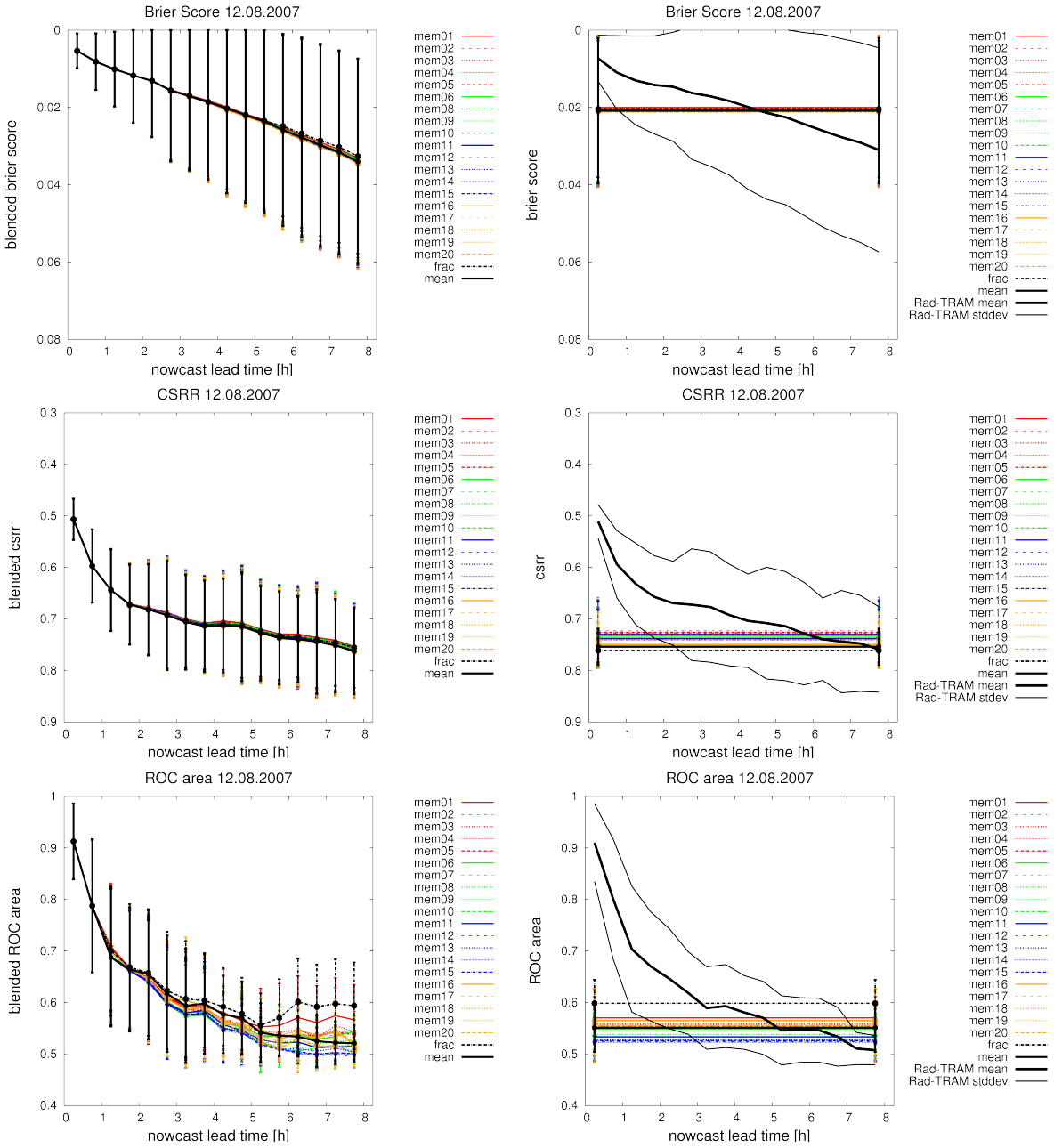

Figure 5.6: Development of Brier Score, CSRR, and area under the ROC curve with lead time for blended probabilities (left column) and the components separately (right column) on 12 August 2007. Lines as explained in Fig. 5.5.

The Brier score of the blended probabilities (Fig. 5.6, top left) is very small as the observed

frequency of 19 dBZ was small (cf. sec. 3.2). As on the other days, the skill of the blended forecasts decreases with lead time. The variability of the mean is high and in the order of the rate of decrease. The shape of the curve based on Rad-TRAM's mean skill (Fig. 5.6, top right) is hardly changed in the blended forecast. Regarding the ranking of the combined probabilities, hardly a difference can be identified between the methods as seen for the COSMO-DE-EPS forecasts separately.
The CSRR of the combined probabilities (Fig. 5.6, middle left) starts from a relatively large value on 12 August 2007 (around 0.5 at 15 minutes lead time) in comparison to other days indicating higher errors even for short lead times. Therefore, the decrease is not as rapid as on other days. As the differences between Rad-TRAM and the forecasts based on COSMO-DE-EPS were small at long lead times (Fig. 5.6, middle right), the resulting blended probability also does not vary significantly from the single values. Although before combination there was some spread between the solutions derived from COSMO-DE-EPS, after the blending procedure the forecasts are very similar. The standard deviation is as large as for Rad-TRAM forecasts and larger than on other days in CSRR.
The area under the ROC curve of the combined probabilities shows an interesting behaviour (Fig. 5.6, bottom left). After the typical decrease in skill in the first forecast hours, after five hours lead time the skill of the fraction method and member 1 increases again. In all methods an improvement in comparison to the single Rad-TRAM and partly to the COSMO-DE-EPS forecasts can be seen for long lead times. Already after two hours of lead time, differences between the methods can be identified. After six hours they show the same ranking as seen in the single forecast (Fig. 5.6, bottom right). On this day, the fraction method clearly bet the neighbourhood members and the mean method that was ordered within them.
Before the blending procedure, differences between the two forecast sources were small in Brier score and CSRR. Therefore, the effect of combination is relatively small but definitively no decrease in skill is seen. The area under the ROC curve is significantly improved especially with two methods (fraction method and member 1) through the blending procedure.

15 August 2007

The 15 August 2007 was of special interest as on this day a complete frontal system including a warm front, a dry warm air sector and a cold front with prefrontal convection passed the domain (cf. sec. 3.3). The evaluation of the two forecast sources separately showed that Rad-TRAM clearly has more skill than the forecasts based on COSMO-DE-EPS in the first forecast hours (Fig. 5.7, right). Later, differences were small in Brier score and CSRR. The area under the ROC curve differed from this behaviour as COSMO-DE-EPS forecasts had more skill than Rad-TRAM already after two forecast hours. Furthermore, there was a large spread between the methods ranging from relatively high values (mean method at 0.7) to values indicating no skill in discrimination (members based on NCEP around 0.5).

The combined probabilities evaluated with the Brier score (Fig. 5.7, top left) reflect the behaviour of Rad-TRAM. Mean skill decreases with increasing standard deviations. During the lead times when the mean values of Rad-TRAM and COSMO-DE-EPS are similar (three to five hours) (Fig. 5.7, top right), the combined values are slightly better. But for long lead times (larger six hours) the combined values are worse than the model based forecasts and follow the decrease of skill of Rad-TRAM. Already after two hours of lead time, differences between the solutions derived from COSMO-DE-EPS can be identified.
The CSRR reflects that the blended forecasts capture the high skill of Rad-TRAM in the first

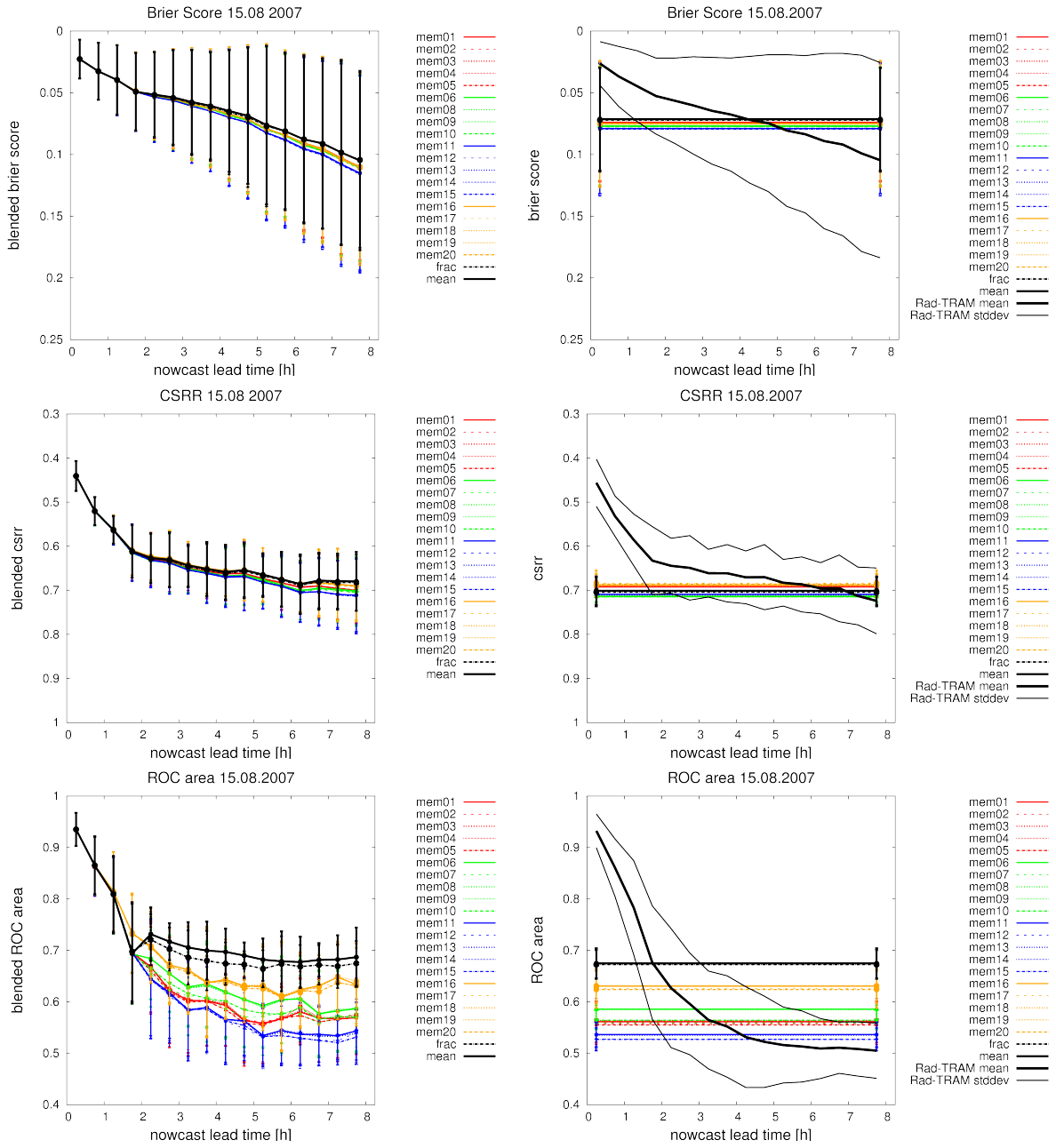

Figure 5.7: Development of Brier Score, CSRR, and area under ROC curve with lead time for blended probabilities (left column) and the components separately (right column) on 15 August 2007. Lines as explained in Fig. 5.5.

forecast hours (Fig. 5.7, middle left). In the last two hours, the combined probabilities have slightly more skill than each single component (Fig. 5.7, middle right). Interestingly, the ranking changes through the combination as the combination based on the mean method has the smallest error and in the model forecasts alone neighbourhood members were superior. But the differences in change are small. The reason for the changed ranking could be that for the combination not the mean value of the forecasts based on the COSMO-DE-EPS is taken but the actual probability at the time the combination is carried out.
The skill of the blended forecasts evaluated with the ROC area (Fig. 5.7, bottom left) is changed through the blending procedure in comparison to the single performances. Rad-TRAM showed a rapid decrease to values near the no skill line (0.5) (Fig. 5.7, bottom right).

Whereas COSMO-DE-EPS based forecasts have large differences but group with the fraction and the mean method followed by neighbourhood members sorted by the same driving global models. The development of skill in terms of discrimination of the combined probabilities varies significantly from Rad-TRAM's skill as the forecasts longer than one hour lead time are already improved. The amount of improvement varies with the respective method. Again, a ranking following global models can be clearly identified with the mean and the fraction method showing more skill than the neighbourhood members.
On the 15 August, it is interesting to see the different effect of the blending procedure with the three scores. In the Brier score, the higher skill of the model forecasts for the long lead times cannot be captured through the blending. Whereas the ROC area can see this improvement for all model solutions already after two hours lead time. The fraction and the mean method based combination even improves the forecast in comparison to the previous time step.

5.3.2 Skill of blended forecasts over entire period

Now, the entire period from 8 to 16 August 2007 is investigated. The Brier score of the blended probabilities decreases steadily with lead time (Fig. 5.8, top left). The rate of decrease is more rapidly in the first three forecast hours than in the following five. The variability of the mean values as seen with the standard deviations in the combined probabilities is still very large. The difference between the solutions based on the different approaches applied on COSMO-DE-EPS are very small and can only be identified after six hours. It is not possible to identify a specific ranking. The comparison with the skill of the single components (Fig. 5.8, top right) reveals that the high skill of Rad-TRAM in the first forecast hours is reproduced. For long lead times (7 and 8 hours), the combined skill is in the range of the skill of COSMO-DE-EPS. The differences between the model solutions are small as well but slightly larger than in the blended probabilities. The variability is in a similar range and larger than the decrease of skill with lead time.
With the CSRR, the skill of the blended probabilities (Fig. 5.8, middle left) decreases steadily over the eight forecast hours as well. The variability of the mean in CSRR is significantly smaller than in the Brier score. It is now smaller than the rate of decrease with increasing lead time. The differences between the solutions based on the different model forecasts are not distinguishable. The comparison with the single components of the combination (Fig. 5.8, middle right) shows that as well with the CSRR, the high skill of the respective components at the respective lead times is reproduced in the blended forecasts.
The area under the ROC curve of the blended forecasts shows a steady decrease of skill in discrimination with increasing lead time (Fig. 5.8, bottom left). Note that in this score a significant increase of skill in comparison to Rad-TRAM's skill alone (Fig. 5.8, bottom right) can be identified for lead times longer than three hours. In the blended forecast based on the best method (mean), the ROC area falls beneath 0.7 after 4 hours. Rad-TRAM's forecast crossed that value already after about 3 hours. The methods produce distinguishable differences between the solutions already after 2 hours. The differences increase with increasing lead time. The lowest values of combined probabilities are around 0.6 (after eight hours lead time). Whereas the single Rad-TRAM forecasts were significantly worse. At lead times where the respective model forecasts have skill in the same order of magnitude as Rad-TRAM, the blending procedure leads to improved forecast quality (i.e., fraction method at four hours). The ranking of the methods is similar to the single model forecast with the fraction and the mean method having more skill than the neighbourhood members that are

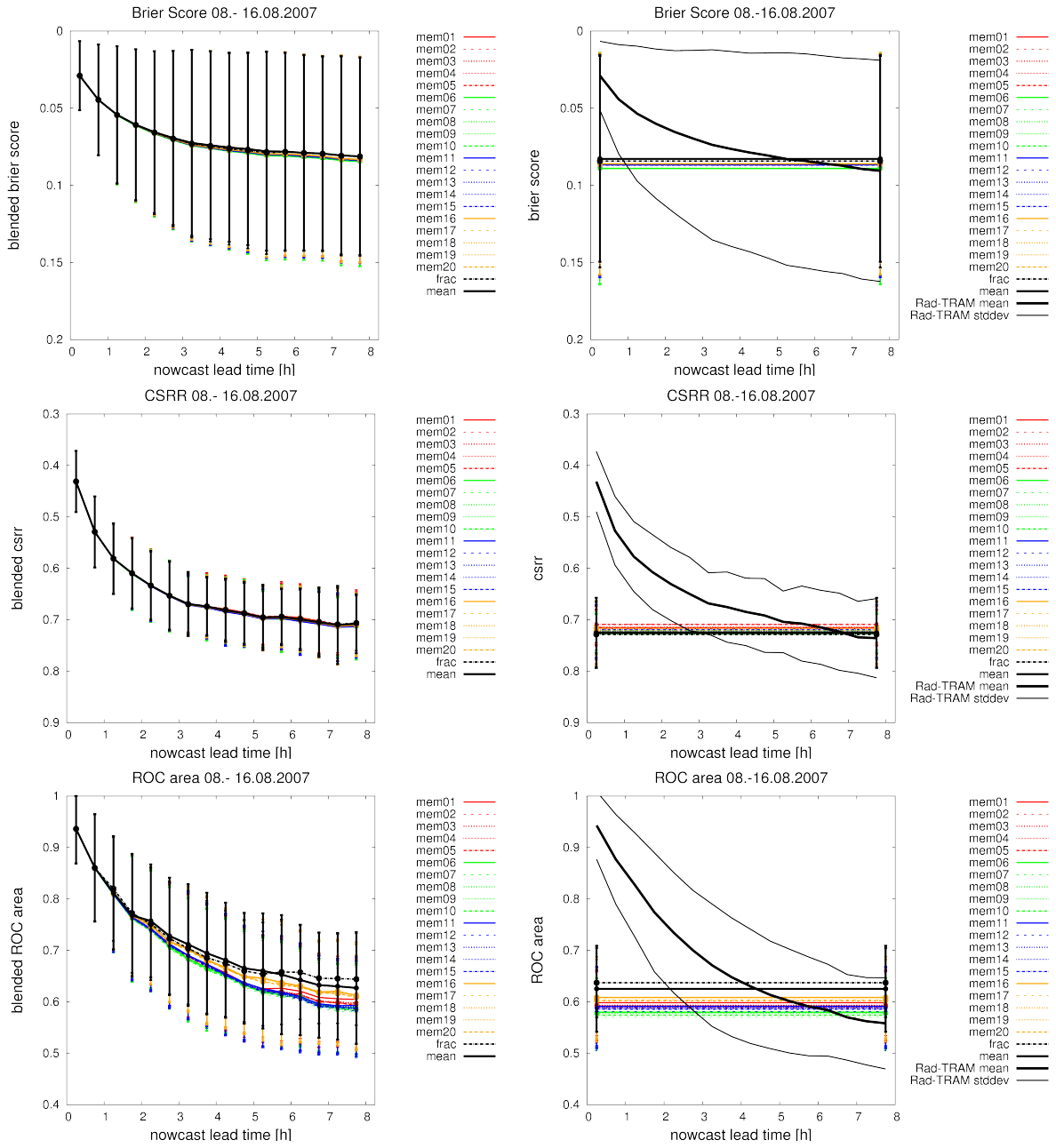

Figure 5.8: Development of Brier Score, CSRR, and area under the ROC curve with lead time for blended probabilities (left column) and the components separately (right column) from 8 to 16 August 2007. Lines as explained in Fig. 5.5.

ranked following global models.
Generally, in all scores the skill of the single methods as evaluated in Fig. 4.9 is reproduced. At each lead time, the blended forecasts have at least as high skill as the single forecasts. For lead times in which the respective components have similar skill, the blended forecast improves the quality. A clear ranking between the different methods can only be identified for the area under the ROC curve as only here large differences occurred.

5.4 Discussion

The relatively simple and straight forward combination of the probabilities based on Rad-TRAM and COSMO-DE-EPS with a method similar to Kilambi and Zawadzki (2005) was applied on the probabilistic forecasts derived in the previous chapter. The resulting probability fields look meaningful and are a seamless combination of the probabilities of the nowcaster and the model.
Generally, the high skill of the nowcaster for short lead times and the superior skill of the model for long lead times can be reproduced with all three investigated scores although the weighting functions are derived only on basis of the CSRR. In the investigation of the entire period it is shown that the blending procedure results in an improvement of the skill of the single components in the range of the cross-over time. In situations where the nowcaster performed better than the model for longer lead times, the combined forecasts might perform worse. This is due to the fact that only one weighting function is applied for all meteorological situations. A meteorological regime dependent combination could consider the different behaviour in different meteorological regimes. Then, the limitations of the applied method could be reduced.
The quality of the combination is limited by the fact that the ensemble is only run once a day and therefore, no real model lead time is available. Furthermore, the evaluation is only performed within one day. That means the latest eight hour interval is evaluated between 16 and 24 UTC. During this time frame, often intense convective events occurred. These are not evaluated for short Rad-TRAM lead times. Often, the performance in intensive situation has good skill for the nowcaster. Hence, as these values are not considered in the calculation of the mean, they tend to underestimate the overall performance. Another limitation of the nowcaster is seen on days with convective cells covering only small areas. The quality of the displacement vectors is not as reliable as on days where large parts of the domain are covered with precipitation. The enlargements of this lower quality displacement vectors up to eight hours leads to not negligible errors. Furthermore, it has to be mentioned that the probabilities are very low at the end of the combination period. The impact from the model are probabilities at different locations as the nowcaster had. As probabilities are low, the combined fields look reasonable and not artificially constructed.

Chapter 6

Discussion

As demonstrated in the previous chapter, the combination of the probabilistic nowcasting method Rad-TRAM with the COSMO-DE-EPS facilitates a seamless prediction of convective precipitation for lead times from 0 to 8 hours. For this purpose, Rad-TRAM has been extended to consider the intrinsic uncertainty in extrapolation forecasts. The output of COSMO-DE-EPS is postprocessed with three different methods to derive probabilistic forecasts. The quality of the combined probabilistic forecasts is evaluated by means of different skill scores in different evaluation setups. The skillful blending of both methods maintains the overall predictive skill for the entire forecast range. At times when the skill of both forecast sources is similar (near cross-over time), the blended forecast even improves the skill. The probabilistic approach of this study can be applied for forecasts of events which require the combination of extrapolation and NWP methods.

For this study, probabilities of exceeding a precipitation threshold of 19 dBZ are predicted. Generally, the methods are applicable to other thresholds as well. But in this study, the choice of the threshold is restricted by the availability of ensemble forecasts with the same configurations and by the availability of the radar observations in six reflectivity classes. Sensitivity studies applying a higher threshold (37 dBZ) result in fewer events. Therefore, the statistical evaluation of the probabilistic forecasts fails due to the limited amount of data from the ensemble. For other meteorological situations (mesoscale convective systems) a higher threshold might be appropriate. A lower threshold of 7 dBZ is not chosen as the 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 (e.g. Bowler et al., 2006). Instead of accumulated rain fall (Golding, 2000), here instantaneous radar reflectivities are employed. Hence, the evaluation of the probabilities with a threshold of 19 dBZ in instantaneous radar reflectivity are stricter compared to lower thresholds in hourly accumulated rain fall. Additionally, this study is conducted on a high resolution grid where the possibility of a double penalty error is higher than on coarser resolution (Ebert, 2008).

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. The fraction is highly dependent on the size of the search area. In this study, the side length of the quadratic search area is increased linearly up to 4 hours forecast time (in agreement with Germann and Zawadzki, 2004). After this

lead time, the search area is kept constant. This means, for lead times from 4 to 8 hours 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 RadTRAM 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 slower compared to pure convective situations. This would result in higher probabilities for longer lead times as the temporal variability of the precipitation field is smaller. For example, appropriate diagnostic means for meteorological regimes as the convective timescale (Done et al., 2006) could be utilised.
Additionally, the calculation of the displacement vector field impacts the forecast quality. Based on Zinner et al. (2008), the pyramidal image matcher is applied. This means, reliable displacement vectors can only be calculated in a neighbourhood of a point where precipitation actually occurred. Therefore, the approach applied in this work is not semi-Lagrangian as in Germann and Zawadzki (2002), but the probability is linearly extrapolated with the vector defined at the point of interest centred in the search area (in Germann and Zawadzki (2002) called 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 limited domain of this study, this effect is likely to be small and thus has little influence on the results.

For the model forecasts, the search area only appears in the neighbourhood method (Theis et al., 2005). As discussed in Chapter 2.3, the fraction of precipitation pixels is computed for a square region of fixed side length of 75 km.
A distinct ranking of the different methods applied on the COSMO-DE-EPS cannot be established as differences in the overall forecast quality between them are very small. However, in the three case studies and the overall time series of the ROC area they differ. This indicates potential for establishing an order with respect to the applied method. Based on such a ranking and assuming a nearly steady meteorological regime, best members can be identified with persistence (cf. Fig. 3.15). In these situations, a best member selection might be appropriate if the data base was larger than the one used in this study.
The ensemble used in this study leaves room for improvement. In the experimental COSMO-DE-EPS, there is no observational data assimilated at the beginning of the forecast. In the setup available for this study, the initial conditions were not perturbed. It is expected that these two factors influence forecast quality and the development of spread between the different members. Especially the assimilation of observations should improve the skill of the model forecasts (i.e. shorten cross-over time) significantly. Hence, the results of this study could as well be understood as conservative or worst case solution.

The application of different probabilistic quality measures in different evaluation setups revealed that the quality measures generally showed a similar behaviour. But in details like the cross-over time they differed. Therefore, it is always reasonable to cross check forecasts with different measures pointing out different aspects of quality. Especially when dealing with probabilities or conditional probabilities, it is important to consider the entire joint distribution of the forecasts and the observations.

An important feature of this study is the calibration of the NWP derived probabilities. Calibrating a relatively rare event in a inhomogeneous precipitation field is an active field of

research (Hamill et al., 2008). The calibration is conducted in a simple and straightforward way using the reliability diagram statistics method (Zhu et al., 1996) for the three methods separately. 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 amount of data was limited. Furthermore, more advanced approaches in calibrating the data (e.g. Hamill et al. (2008) or Raftery et al. (2005)) have to demonstrate that the outcome of the probabilistic forecast will be improved significantly.
The calibration with the reliability diagram statistics method reduced the reliability component of the Brier score (Tab. 2.4) and the sharpness. With the single calibration function, the spread of the neighbourhood members is reduced. Furthermore, the various methods differ marginally in their skills after the calibration as their calibration functions are similar as well. The main difference between the forecasts based on COSMO-DE-EPS with calibration to COSMO-DE-EPS without calibration is not the magnitude but the location of the probability fields.

The weighting functions are the basis for the additive combination of the probabilistic forecasts. Here, the study is restricted to a single weighting function w_r that is determined by the development of Rad-TRAM's forecast skill with lead time in CSRR. This score is chosen as its general decrease of skill with lead time was similar to the other scores but the standard deviations were smaller in CSRR (Fig. 4.9). It would be desirable to have a similar lead time dependent weighting function for the COSMO-DE-EPS output. Due to the setup of the ensemble runs, this quantity was not available. However, it was shown that COSMO-DE-EPS based forecasts in a first approximation do not depend on lead time (Chapter 3.2) for lead times larger than three hours (spin-up time). Therefore, and as the combined quantities are probabilities, the lead time dependence for all model forecast was calculated as $1 - w_r$. 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. A study with the methods derived in this thesis applied on a time-lagged ensemble constructed with COSMO-DE is currently conducted by a diploma student.
In the derivation of the weighting functions, the consideration of meteorological regimes could as well improve the results. The three case studies showed that in different meteorological situations, the development of skill varies. For example, in the frontal regime, for longer lead times a larger weight could be given to the nowcaster. Whereas, in purely convective situations, the model forecasts should be considered earlier. This might improve the quality of the blended probabilistic forecasts.
The application of the weighting functions on the two probabilistic forecasts results in blended probabilistic forecasts. The evaluation of their forecast skill with all quality measures used in this study shows consistently at all lead times at least the same skill as for the respective best single forecasts. In all scores, the skill is even improved for lead times around the cross-over time.

As a first attempt to construct a blending of probabilistic nowcasts and high-resolution ensemble forecasts, the methods in this thesis have been chosen to be as simple as possible. An important factor that has been ignored thus far but was already briefly mentioned in this discussion is the dependence of forecast skill on weather regime. If this is different for the nowcasts and ensemble forecasts, it may be possible to optimise 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 timescale 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. (2010), Zimmer et al. (2010)), and could be used to construct more optimal calibration and weighting functions for the short and long timescale regimes.

Several aspects of the study were discussed in this chapter. To summarise, the major results are pointed out. The strength of this study is that the major goal was reached. The framework to provide seamless probabilistic predictions that at least match the best result of the individual components was developed. Furthermore, the probabilistic forecast of exceeding a specific threshold might give guidance about specific hazard levels relevant to a user together with the confidence of the forecast. A further strength of the method is that once the forecast and the nowcast are available, the computational effort in the blending procedure is comparable small.
There are also several weaknesses as discussed above. Two points mainly influence the reaching of the overall goal. The poor quality of the ensemble that seems to be mainly caused by the not included data assimilation weakens the model in comparison to the nowcaster and as well the quality of the blended forecast. Furthermore, the limited availability of the experimental COSMO-DE-EPS bounded this study in several aspects (choice of threshold, consideration of meteorological regimes in calibration and weighting functions).
Although this work developed the fundamental framework to improve short-range forecast quality, there are still some opportunities that seem to be the most promising possibilities for further improvement. The most promising opportunities are the improvement of the quality of the ensemble forecasts and the systematic and analytic consideration of meteorological regimes. If these are fulfilled, interesting studies to sensitivity of the forecast quality on the precipitation threshold are reasonable.

In the long run, one might expect that blending of nowcasts with numerical forecasts could be replaced by direct assimilation of radar and other observation data into the numerical model. 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 models' 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 prepare 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.

Chapter 7

Conclusions and Outlook

The aim of this thesis was to develop a framework in which probabilistic forecasts of convective precipitation are provided such that over several forecast hours their skill is maximised. Therefore, two different forecasting methods, nowcasting and NWP, had to be combined. Considering their sources of error and consequently, uncertainty in the forecasts of both methods, probabilistic forecasts were derived.

The radar tracker Rad-TRAM was extended with an optional module that provides probabilistic forecasts of exceeding a certain threshold in radar reflectivity based on the Local Lagrangian method for eight hours in 15 min time steps.
For the model forecasts, the experimental ensemble COSMO-DE-EPS of the Deutscher Wetterdienst based on the COSMO-DE model was used. Three techniques were introduced to derive probabilistic information from the COSMO-DE-EPS output. After the calibration of the probabilistic model forecasts with the reliability diagram statistics method, the probability fields of the different solutions mainly differed in location but not in amplitude of the probability.
The skill of the probabilistic forecasts based on the two methods was evaluated with several probabilistic quality measures in order to consider different aspects of quality. As well the development of skill with time in time series as with lead time was evaluated for both forecasts. Regarding the model forecasts, the effect of the calibration on the forecast skill was investigated. It was found that the calibration results in a reduction of spread and an increase of mean skill. After calibration, no large difference in forecast skill between the solutions was found. The results of this investigation were robust in terms of the applied probabilistic quality measures. The investigation of the development of skill with lead time revealed in all scores consistently that over the entire period, Rad-TRAM forecasts are superior to COSMO-DE-EPS forecasts up to lead times of 5 to 7 hours.
The development of skill with lead time was the basis for the derivation of the weighting functions for the additive combination of the probabilistic forecasts in the blending procedure. Here, the weighting functions were defined on basis of the development of skill of Rad-TRAM forecasts solely. This was necessary due to the set up of the COSMO-DE-EPS that is only run once a day. Only one weighting function was applied on all 22 forecasts derived from COSMO-DE-EPS output. This function was defined based on the Rad-TRAM weighting function, as $1 - w_r$. The resulting blended probability forecasts revealed to be meaningful and seamless forecasts. The objective evaluation of the quality of the blended forecasts proved this subjective impression. The skill of the respective best forecast at the different lead times was reproduced in the investigation of the entire period. In the time

frame of the cross-over period when the single components had skill in a similar magnitude, the skill of the blended forecasts was even higher. Nevertheless, the investigation of specific case studies revealed that there is a dependence of skill of the blended forecasts in comparison to the single components on the respective dominating meteorological regime.

In recent years, several groups have addressed the challenge of accurately forecasting convective precipitation over lead times of several hours. It is agreed that with current knowledge, it is necessary to combine forecasts based on observations with forecasts based on numerical weather prediction to optimise skill. The combination in probability space, as performed in this study, seems to be the most promising approach. Deficiencies as known from combinations of radar reflectivities or rain amounts that involve the creation of non physical phenomena or the destruction of mesoscale organisations of convection cannot occur in the probabilistic approach.

Although the methods derived in this work have proved their skill and can already be used as forecast support in field campaigns (like Wetter & Fliegen), there are still several aspects that could be addressed in future work.
One major shortcoming of this study was the set up of the COSMO-DE-EPS. To evaluate the development of the forecast skill with lead time it would be necessary to have several model runs each day available. With such a time-lagged ensemble, a concrete evaluation of the development of skill would be possible. The application of data assimilation methods could improve the general performance of the COSMO-DE-EPS as well. Another shortcoming in the set up is that the initial conditions are so far not perturbed in the COSMO-DE-EPS. As this is another major source of uncertainty, this should be considered in continuative studies. The perturbation of model physics with a stochastic approach is another possibility to enlarge the spread of the ensemble.
The methods introduced in this work should be applied on several precipitation thresholds. For example a higher threshold representing heavy precipitation or hail as often found in thunderstorms could be of interest for several decision makers, e.g. at airports. As higher reflectivities are even rarer events than the threshold of 19 dBZ applied in this study, a larger amount of data is necessary to receive reliable statistics.
The most promising direction for future work seems to be the consideration of the dependence of forecast skill on meteorological regimes. First investigations conducted in this study indicate that the development of skill of the nowcasting method and the model differs with the meteorological situations. If further investigations with a larger amount of data confirm these findings, a regime dependent definition of the calibration and the weighting functions could increase the robustness of this study's results. The concept of convective timescale introduced by Done et al. (2006) has the potential to be an objective method to identify the relevant regimes. It measures the degree to which cumulus convection is controlled by larger-scale dynamical processes. Hence, it could be used to construct optimised calibration and weighting functions for the short and long timescale regimes.

Appendix A

List of abbreviations and symbols

08FDP	Beijing 2008 Forecast Demonstration Project
a	contingency table: forecast yes and observed yes
ASCII	American Standard Code for Information Interchange
α	displacement vector
b	contingency table: forecast yes and observed no
B	buoyancy
BS	Brier score
c	contingency table: forecast no and observed yes
C	constant of radar equation
CAPE	convective available potential energy
Cb-TRAM	Cumulonimbus Tracking and Monitoring
CIN	convective inhibition
CMAE	conditional mean absolute error
COPS	Convective and Orographically-induced Precipitation Study
COSMO	Consortium for Small-scale Modeling; the numerical model developed in this consortium
COSMO-DE	COSMO model with 2.8 km horizontal resolution
COSMO-DE-EPS	experimental ensemble based on COSMO-DE model
COSMO-SREPS	Short-Range Ensemble Prediction System
COSPA	Consolidated Storm Prediction for Aviation
COTREC	Continuity of TREC
CSI	Critical Success Index
CSRR	Conditional Squared root of ranked probability score
d	contingency table: forecast no and observed no
dBZ	decibels of Z
D	particle diameter
DLR	Deutsches Zentrum für Luft- und Raumfahrt e.V. (German Aerospace Center)
DWD	Deutscher Wetterdienst (German weather service)
ECMWF	European Centre for Medium Range forecasts
EPS	Ensemble Prediction System
F	False Alarm Rate/Ratio
g	acceleration due to gravity
g	antenna gain

GEM/HIMAP	Global Environmental Multiscale-High resolution Model Application Project
GHz	Giga Hertz
h	pulse length
hPa	hecto Pascal
H	Hit rate (also named Probability of Detection)
IOP	intensive observation periods
k	side length of search area
k	numbering of n forecast - event pairs
K	dielectric constant
K	diffusion coefficient
LCL	lifted condensation level
LFC	level of free convection
LNB	level of neutral buoyancy
$\mathcal{L}$	threshold in (observed or simulated) radar reflectivity
$\mathcal{L}_{max}$	maximum threshold
$\mathcal{L}_{min}$	minimum threshold
λ	wavelength of transmitted electromagnetic wave
m	refraction index
MAPLE	McGill Algorithm for Precipitation Nowcasting by Lagrangian Extrapolation
MAPLE-NOFF	near optimal forecast filtering
MASCOTTE	Maximum Spatial Correlation Tracking Technique
n	number of iterations
n	total number of forecast-event pairs
N	number of forecast event pairs
N_i	number of times each forecast y_i is used
NCAR	National Center for Atmospheric Research
NCEP	National Center for Ensemble Prediction
NCWF-2	National Convective Weather Forecast version 2
NCWF-6	National Convective Weather Forecast version 6
NHM	Non-hydrostatic Model
NWP	Numerical Weather Prediction
$\bar{o}$	overall frequency of the event
$\bar{o_i}$	relative frequencies of occurrence in each subsample
o_j	occurrence of event j in observations
ω_k	search area around the point of interest with the scale parameter k
Ω	space domain
$\tilde{\Omega}_{t_0+\tau}$	size of the rain domain
p	pressure
$p(o_j\|y_i)$	conditional probability that the event o_j occurred given the forecast y_i
$p(y_i)$	marginal distribution of the forecasts (also named refinement)
P	probabilistic forecast
$\hat{P}$	binary observation
$P_{blend,i}$	blended probabilistic forecast with i different model approaches
P_r	reflected power
P_t	transmitted power
P_{LL}	probabilistic forecast with Local Lagrangian method
$P_{EPS,i}$	probabilistic forecast with EPS system with i different approaches
POD	probability of detection
Ψ	Precipitation field

QPF	Quantitative Precipitation Forecast
r	distance from target to radar
r	position within search area
R_d	individual gas constant of dry air
Rad-TRAM	Radar Tracking and Monitoring
RAPIDS	Rainstorm Analysis and prediction Integrated Data-processing System
RDT	Rapid Developing Thunderstorms
RMS	Root mean squared error
RMSF	RMS Factor
ROC	Relative Operating Characteristics or Receiver Operating Characteristics
RCPF	RUC Convective Probability forecast
RUC	Rapid Update Cycle
SPROG	Spectral Prognosis
STEPS	Short-Term Ensemble Prediction System
SWIRLS	Short-range Warning of Intense Rainstorms in Localized Systems
σ	backscatter cross-section
σ_i	backscatter cross-section of the single scatterer i
t	time
TITAN	Thunderstorm Identification, Tracking, Analysis, and Nowcasting
TKE	Turbulent kinetic energy
TREC	Tracking Radar Echoes by Correlation
T_v	virtual temperature of environment
T_vp	virtual temperature of the parcel
τ	forecast lead time
Θ_0	opening angle
UKMO	United Kingdom Meteorological Office
UM	Unified Model
UTC	universal time coordinated
w_c	weight for COSMO-DE-EPS based forecasts
w_r	weight for Rad-TRAM based forecasts
WRF	Weather Research and Forecast model
x	location
y	meridional Cartesian coordinate
y_i	possible forecasts in i categories
z	radar reflectivity factor
Z	logarithmic radar reflectivity factor
z_i	height at level i

Bibliography

Andersson, T. and Ivarsson, K.-I.: 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. Wea. Rev.* **131**, 1509–1523.

Baldauf, M., Stephan, K., Klink, S., Schraff, C., Seifert, A., Förstner, J., Reinhardt, T. and Lenz, C.-J.: 2006, The new very short range forecast model LM-K for the convection-resolving scale, *Second THORPEX International Science Symposium.* **Volume extended abstracts, Part B**, 148–149, Available at http://www.pa.op.dlr.de/stiss/proceedings.html.

Barillec, R. and Cornford, D.: 2009, Data assimilation for precipitation nowcasting using Bayesian inference, *Adv. Water Res.* **32**, 1050–1065.

Battan, L.: 1973, *Radar observation of the atmosphere*, University of Chicago press Chicago.

Blackadar, A.: 1962, The vertical distribution of wind and turbulent exchange in a neutral atmosphere, *J. Geophys. Res.* **67**, 3095–3102.

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

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

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

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

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

Carvalho, L. and Jones, C.: 2001, A satellite method to identify structural properties of mesoscale convective systems based on the maximum spatial correlation tracking technique (MASCOTTE), *J. Appl. Met.* **40**, 1683–1701.

Craig, G. C., Keil, C. and Leuenberger, D.: 2010, Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection, *submitted to Quart. J. Roy. Meteor. Soc.* .

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.

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

Doms, G., Förstner, J., Heise, E., Herzog, H.-J., Raschendorfer, M., Reinhardt, T., Ritter, B., Schrodin, R., Schulz, J.-P. and Vogel, G.: 2007, A description of the nonhydrostatic regional model LM. Part II: Physical Parameterization, *Deutscher Wetterdienst (German Weather Service); available online at http://cosmo-model.org* .

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

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

Doviak, R. and Zrnic, D.: 1984, *Doppler Radar and Weather Observations*, Academic Press, Inc., Orlando, FL.

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

Emanuel, K. A.: 1994, *Atmospheric convection*, Oxford University Press.

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

Gebhardt, C., Theis, S., Paulat, M. and Bouallegue, Z. B.: 2010, Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries, *submitted to Atmos. Res.* .

Germann, U. and Zawadzki, I.: 2002, Scale-dependence of the Predictability of Precipitation from Continental Radar Images. Part I: Description of the methodology, *Mon. Wea. Rev.* **130**, 2859–2873.

Germann, U. and Zawadzki, I.: 2004, Scale Dependence of the Predictability of Precipitation from Continental Radar Images. Part II: Probability Forecasts, *J. Appl. Met.* **43**, 74–89.

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

Golding, B.: 2000, Quantitative precipitation forecasting in the UK, *J. Hydrol.* **239**, 286–305.

Hamill, T. and Colucci, S.: 1998, Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts, *Mon. Wea. Rev.* **126**, 711–728.

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

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

Höller, H.: 1994, Mesoscale Organization and Hailfall Characteristics of Deep Convection in Southern Germany, *Beitr. Phys. Atmos.* **67**, 219–234.

Holton, J.: 2004, *An introduction to dynamic meteorology*, Academic press.

Kilambi, A. and Zawadzki, I.: 2005, An evaluation of ensembles based upon MAPLE precipitation nowcasts and NWP precipitation forecasts, *32nd Conference on Radar Meteorology, Albuquerque, NM.*

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

Lewis, J.: 2005, Roots of ensemble forecasting, *Mon. Wea. Rev.* **133**, 1865–1885.

Li, L., Schmid, W. and Joss, J.: 1995, Nowcasting of Motion and Growth of Precipitation with Radar over a Complex Orography, *J. Appl. Met.* **34**, 1286–1300.

Li, P.-W., Wong, W.-K. and Lai, E.: 2005, RAPIDS- a new rainstorm nowcasting system in Hong Kong, *World Weather Research Program Symposium on Nowcasting and Very Short Range Forecasting, Toulouse, France.*

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

Lovejoy, S. and Mandelbrot, B.: 2010, Fractal properties of rain, and a fractal model, *Tellus A* **37**, 209–232.

Marsigli, C., Montani, A. and Paccagnella, T.: 2008, The COSMO-SREPS ensemble for the short-range: system analysis and verification on the MAP D-PHASE DOP, *Joint MAP D-PHASE Scientific Meeting-COST 71 mid-term seminar, Bologna, Italy*, pp. 9–14.

Megenhardt, D., Mueller, C., Trier, S., Ahijevych, D. and Rehak, N.: 2004, NCWF-2 probabilistic nowcasts, *11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis, Massachusetts*, p. 23.

Morel, C. and Senesi, S.: 2002, A climatology of mesoscale convective systems over Europe using satellite infrared imagery I: Methodology, *Quart. J. Roy. Meteor. Soc.* **128**, 1953–1971.

Mueller, C., Saxen, T., Roberts, R., Wilson, J., Betancourt, T., Dettling, S., Oien, S. and Yee, J.: 2003, NCAR Auto-Nowcast System, *Wea. Forecasting* **18**, 545–561.

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

Murphy, A.: 1993, What is a Good Forecast? An Essay on the Nature of Goodness in weather Forecasting, *Wea. Forecasting* **8**, 281–293.

Murphy, A. and Winkler, R.: 1987, A general framework for forecast verification, *Mon. Wea. Rev.* **115**, 1330–1338.

Pierce, C., Ebert, E., Seed, A., Sleigh, M., Collier, C., Fox, N., Donaldson, N., Wilson, J., Roberts, R. and Mueller, C.: 2004, The nowcasting of precipitation during Sydney 2000: an appraisal of the QPF algorithms, *Wea. Forecasting* **19**, 7–21.

Pierce, C., Hardaker, P., Collier, C. and Hagget, C.: 2000, GANDOLF: A system for generating automated nowcasts of convective precipitation, *Meteorol. Appl.* **7**, 341–360.

Pinto, J., Mueller, C., Weygandt, S., Ahijevych, D., Rehak, N. and Megenhardt, D.: 2006, Fusion observation- and model-based probability forecasts for the short term prediction of convection, *12th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis, Massachusetts*, p. 5.

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

Rinehart, R.: 1997, *Radar for meteorologists*, Dept. of Atmospheric Sciences, Center for Aerospace Sciences, University of North Dakota.

Rinehart, R. and Garvey, E.: 1978, Three-dimensional storm motion detection by conventional weather radar, *Nature* **273**, 287–289.

Roebber, P., Schultz, D., Colle, B. and Stensrud, D.: 2004, Towards improved prediction: high-resolution and ensemble modeling systems in operations, *Wea. Forecasting* **19**, 936–949.

Rotach, M., Ambrosetti, P., Ament, F., Appenzeller, C., Arpagaus, M., Bauer, H., Behrendt, A., Bouttier, F., Buzzi, A., Corazza, M. et al.: 2009, MAP D-PHASE: real-time demonstration of weather forecast quality in the Alpine region, *Bull. Am. Meteorol. Soc.* **90**, 1321–1336.

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

Schmid, W., Mecklenburg, S. and Joss, J.: 2002, Short-term risk forecasts of heavy rainfall, *Water Sci. Technol.* **45**, 121ff.

Schreiber, K.-J.: 2000, Der Radarverbund - Informationen zum Wetterradar-Verbundsystem, *Deutscher Wetterdienst* .

Schwartz, C., Kain, J., Weiss, S., Xue, M., Bright, D., Kong, F., Thomas, K., Levit, J., Coniglio, M. and Wandishin, M.: 2010, Towards Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Scale Ensemble Membership, *Wea. Forecasting* **25**, 263–280.

Seed, A.: 2003, A dynamic and spatial scaling approach to advection forecasting, *J. Appl. Meteorol.* **42**, 381–388.

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

Smith, R.: 1997, *The physics and parameterization of moist atmospheric convection*, Springer.

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

Stanski, H., Wilson, L. and Burrows, W.: 1989, Survey of common Verification Methods in Meteorology, *Technical report*, World Weather Watch Technical Report No. 8 WMO/TD No. 358. WMO, Geneva.

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

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

Sugihara, K. and Inagaki, H.: 1995, Why is the 3D Delaunay triangulation difficult to construct?, *Information Processing Letters* **54**, 275–280.

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

Tiedtke, M.: 1989, A comprehensive mass flux scheme for cumulus parameterization in large-scale models, *Mon. Wea. Rev.* **117**, 1779–1800.

Turner, B., Zawadzki, I. and Germann, U.: 2004, Predictability of precipitation from continental radar images. Part III: Operational nowcasting implementation (MAPLE), *J. Appl. Meteorol.* **43**, 231–248.

Weigl, E., Klink, S., Kohler, O., Reich, T., Rosenow, W., Lang, P., Podlasly, C., Winterrath, T., Adrian, G., Majewski, D. and Lang, J.: 2005, Abschlussbericht Projekt RADVOR-OP: Radargestützte, zeitnahe Niederschlagsvorhersage für den operationellen Einsatz (Niederschlag-Nowcasting-System), *Technical report*, Deutscher Wetterdienst, Abteilung Hydrometeorologie.

Weygandt, S. and Benjamin, S.: 2004, RUC model-based convective probability forecasts, *11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis, Massachusetts*, p. 11.

Wilks, D.: 2006, *Statistical methods in the Atmospherics Sciences*, Academic Press, San Diego, London.

Wilson, J., Crook, N., Mueller, C., Sun, J. and Dixon, M.: 1998, Nowcasting Thunderstorms: A Status Report, *Bull. Am. Meteorol. Soc.* **79**, 2079–2099.

Wilson, J., Ebert, E., Sleigh, M., Pierce, C., Saxen, T., Roberts, R., Mueller, C. and Seed, A.: 2004, Sydney 2000 forecast demonstration project: convective storm nowcasting, *Wea. Forecasting* **19**, 131–150.

Wilson, J. and Xu, M.: 2006, Experiments in blending radar echo extrapolation and NWP for Nowcasting convective storms, *4th Conference on Radar in Meteorology and Hydrology, Barcelona, Spain*, pp. 519–522.

Wolfson, M., Dupree, W., Rasmussen, R., Steiner, M., Benjamin, S. and Weygandt, S.: 2008, Consolidated storm prediction for aviation (CoSPA), *Integrated Communications, Navigation and Surveillance Conference, 2008. ICNS 2008*, pp. 1–19.

Wong, W., Yeung, L., Wang, Y. and Chen, M.: 2009, Towards the Blending of NWP with Nowcast: Operation Experience in B08FDP, *World Weather Research Program Symposium on Nowcasting, Whistler, B.C. Canada*.

Wulfmeyer, V., Behrendt, A., Bauer, H., Kottmeier, C., Corsmeier, U., Blyth, A., Craig, G., Schumann, U., Hagen, M., Crewell, S. et al.: 2008, The convective and orographically induced precipitation study: A research and development project of the World Weather Research Program for improving quantitative precipitation forecasting in low-mountain regions, *Bull. Am. Meteorol. Soc* **89**, 1477–1486.

Zhu, Y., Iyengar, G., Toth, Z., Traclon, S. and Marchok, T.: 1996, Objective Evaluation of the NCEP global ensemble forecasting system, *15th Conf. on Weather Analysis and Forecasting, Norfolk, VA, Amer. Meteor. Soc.* pp. J79–J82.

Zimmer, M., Craig, G. C., Wernli, H. and Keil, C.: 2010, Classification of precipitation events with a convective response timescale, *submitted to Geophys. Res. Lett.* .

Zimmer, M. and Wernli, H.: 2008, COPS ATLAS - The meteorological situation from June 1 till August 31, 2007, *Technical report*, Johannes Guttenberg Universität Mainz.

Zinner, T., Mannstein, H. and 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.

Acknowledgements

I am deeply grateful to Prof. Dr. George C. Craig who supervised this thesis. He always had time for interesting, motivating, and encouraging discussions. His positive and creative way of thinking considering several perspectives of problems was and is very inspiring. His suggestions and comments substantially improved this thesis.

I would also like to thank Prof. Dr. Ulrich Schumann for reviewing this thesis as co-examiner. He improved the work with helpful suggestions and comments.

I would like to thank Christian Keil for his continuous support over the last years. He was always open for discussions of any kind and provided me the COSMO-DE-EPS data.

I thank Arnold Tafferner for the impetus to this work. He gave me the possibility to develop my own ideas within this work freely.

Susanne Theis and Christoph Gebhardt at DWD made the COSMO-DE-EPS data accessible and were always open for discussions. Especially Susanne Theis' interest and comments helped to look at the work from another angle.

Andreas Dörnbrack's many comments are gratefully acknowledged, especially concerning writing. His positive way of thinking was and is very constructive.

Thanks also to Winfried Beer who always found quick and robust solutions for any kind of technical problems. Hermann Mannstein, Caroline Forster, and Patrick Tracksdorf were always open to find solutions to IDL problems.

Verena Fiedler, Christian Keil, and Kersten Schmidt kindly helped me to finish the thesis with helpful comments on the text.

I also like to thank all the members of the department for the pleasant working atmosphere. Also members from other departments made the time at work pleasant, especially Simone, Christian, Christoph, Verena, Ingo, Johannes, Rudi, Christine, Helge, and Dominik. A special thanks also to my room mates in the last years: Moni, Vera, and Klaus.

I especially would like to thank my friends for continuously remembering me that there is a life beside work and their understanding.

Most deeply I thank those who are closest to me: my grandparents, my mother, my brothers Markus and Martin (with Anja, Finja, Tim, and Maximilian) and Jochen. They believed in me and provided me the support I needed - everyone in his own way.

Curriculum Vitae

Kirstin Kober

Born on 27 September 1981 in Munich, Germany

University Education

2007–2010	Ph.D. student at DLR, Institut für Physik der Atmosphäre, Oberpfaffenhofen
2006–2007	Project scientist at DLR, Institut für Physik der Atmosphäre, Oberpfaffenhofen
December 2006	Diploma in Meteorology, Diploma Thesis: "Verfolgung von Gewitterzellen mittels Fernerkundungsdaten von Satellit und Radar"
2001–2006	Study of Meteorology at the Ludwig-Maximilians Universität München

Primary and Secondary School

1992–2001	Ignaz-Kögler-Gymnasium, Landsberg/Lech
1988–1992	Primary School, Landsberg/Lech