A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System

CHRISTIAN KEIL AND GEORGE C. CRAIG

Institut für Physik der Atmosphäre, DLR Oberpfaffenhofen, Wessling, Germany

(Manuscript received 19 September 2006, in final form 18 December 2006)

ABSTRACT

Errors in regional forecasts often take the form of phase errors, where a forecasted weather system is displaced in space or time. For such errors, a direct measure of the displacement is likely to be more valuable than traditional measures. A novel forecast quality measure is proposed that is based on a comparison of observed and forecast satellite imagery from the *Meteosat-7* geostationary satellite. The measure combines the magnitude of a displacement vector calculated with a pyramid matching algorithm and the local squared difference of observed and morphed forecast brightness temperature fields. Following the description of the method and its application for a simplified case, the measure is applied to regional ensemble forecasts for an episode of prefrontal summertime convection in Bavaria. It is shown that this new method provides a plausible measure of forecast error, which is consistent with a subjective ranking of ensemble members for a sample forecast. The measure is then applied to hourly images over a 36-h forecast period and compared with the bias and equitable threat score. The two conventional measures fail to provide any systematic distinction between different ensemble members, while the new measure identifies ensemble members of differing skill levels with a strong degree of temporal consistency. Using the displacement-based error measure, individual ensemble members are found to compare better with observations than either a short-term deterministic forecast or the ensemble mean throughout the convective period.

1. Introduction

In recent years numerical weather prediction models have become more complex and have been applied on finer scales. These high-resolution models have the potential to forecast phenomena that are highly localized and episodic, as for instance warm season precipitation events. Unfortunately, traditional approaches for the validation of spatial forecasts, including convection forecasts and quantitative precipitation forecasts, are inadequate to meet current needs. Generally, these traditional measures are computed on the basis of contingency tables constructed by comparing point values. Based on the counts, a variety of skill scores such as bias score or equitable threat score can be computed (Wilks 1995). However, a common problem of high-resolution forecast fields occurs in conditions where a weather system is properly developed in the model but improperly positioned. A forecaster or analyst would ascribe some skill to such a forecast, whereas conventional scores might not ("double penalty" problem). For such misplacement errors, a direct measure of the displacement is likely to be more valuable than traditional measures.

There may be only limited information in observations of quantities such as pressure to identify which forecasts accurately capture a convective storm. Rather than using conventional data and error measures to evaluate forecast quality, it may be better to use remote sensing information. For instance, composites of ground-based radar instruments deliver radar reflectivity maps indicating precipitation at high spatial and temporal resolutions. Geostationary satellite imagery display brightness temperature (BT) fields indicating, for example, atmospheric cloud or water vapor structures at high precision. A correct forecast of precipitation and clouds can be seen as a measure of the overall forecast quality (Mesinger 1996).

Attempts to exploit the information contained in re-

Corresponding author address: Christian Keil, Institut für Physik der Atmosphäre, DLR Oberpfaffenhofen, D-82234 Wessling, Germany.
E-mail: christian.keil@dlr.de

DOI: 10.1175/MWR3457.1

motely sensed imagery in order to evaluate forecasts have been made by several researchers using techniques originally developed for image processing. Feature-based verification techniques have been applied on satellite-observed precipitable water fields (Hoffman and Grassotti 1996), and on radar observations (Brewster 2003a,b; Zepeda-Arce and Foufoula-Georgiou 2000; Casati et al. 2004; Venugopal et al. 2005). In such object-oriented approaches the forecast quality depends on the models' ability to reproduce the multiscale spatial structure and space–time dynamics of the observed weather systems. Usually, an evaluation on different spatial scales is performed to assess the phase error, while an intensity-scale evaluation gives the amplitude error of the forecast. Ebert and McBride (2000) applied such an approach on entities labeled contiguous precipitation areas based on daily rain gauge observations. Their method determines displacement errors and other parameters for contiguous regions that can be decomposed according to the sources of error (e.g., displacement, pattern, etc.). A similar method has recently been developed by Davis et al. (2006a) matching forecast and observed rain-area pairs based mainly on the separation of their centroids relative to the sum of their sizes. Davis et al. (2006b) extend this method to measure spatiotemporal errors by identifying features that are continuous in both space and time in the forecast or in the observations.

Given the uncertainties of precipitation forecasting, point- and time-specific prediction of precipitation intensity is in practice nondeterministic, especially during the warm season. Numerous studies (e.g., Bright and Mullen 2002; Yuan et al. 2005, and references therein) suggest that the ensemble approach could improve short-range weather forecasts, especially precipitation forecasting. In ensemble prediction systems the inherent observational uncertainty, model error, and the chaotic, nonlinear behavior of atmospheric dynamics can be incorporated, providing a range of scenarios with information about the forecast uncertainty [see the Network of European Meteorological Services (EUMETNET) Web site at http://srnwp.cscs.ch/]. However, ensemble forecasting at high spatial resolutions generates a multitude of highly localized and episodic phenomena, and renders a comprehensive comparison of individual ensemble members with observations even more challenging, underpinning the need for validation methods based on the patterns of weather objects.

In the present paper a forecast quality measure is presented that crucially builds on the pattern information contained in the imagery of the *Meteosat-7* geostationary satellite. The aim of this study is to demonstrate a technique for estimating the displacement error, and to examine whether the displacement-based algorithm provides a reasonable error measure, by applying it in the framework of a regional ensemble system to select and rank individual ensemble members based on this technique, and compare the ranking with a subjective evaluation of forecast quality.

In the next section the forecast quality measure and its application on an idealized feature are presented. This is followed by a description of the regional forecasting system, the generation of synthetic satellite images within the mesoscale model, and the observational data in section 3. The application of the pyramid matching technique on a case study is presented in section 4. Finally, conclusions are drawn and an outlook is given in section 5.

2. A forecast quality measure

The objective evaluation of forecast quality is performed using the pyramidal matching algorithm, which was originally developed to detect and track cloud features (e.g., convective clouds, contrails) in satellite imagery (Mannstein et al. 2002; Muller et al. 2007; Zinner et al. 2007, manuscript submitted to *Meteor. Atmos. Phys.*, hereafter ZMT). The pyramid matching algorithm computes a vector field (optical flow) that deforms, or morphs, an image into a replica of another image by seeking to minimize an amplitude-based quantity (e.g., correlation coefficient, mean squared error) at different scales within a fixed search environment. The vector field is computed using the following steps.

1) The two images are coarse grained by averaging 2^F pixels onto one pixel element (F is called the subsampling factor); this is the topmost pyramid level (lowest resolution).
2) A displacement vector at each pixel element location is computed by translating one image within the range of ± 2 pixel elements in all directions, and choosing the displacement that gives the minimum squared difference in a local region centered on the pixel element. The extent of the local region is defined by a Gaussian kernel (ZMT) with compact support on a five pixel by five pixel region centered on the location of the original pixel element.
3) This vector field is then applied to the original image to generate an intermediate image that accounts for the large-scale (topmost pyramid level) motions.
4) The intermediate image is then coarse grained by averaging 2^{F-1} pixels to generate pixel elements at the next pyramid level, and a motion vector field is determined as in step 2 above, which can be regarded as a correction to the vector field computed

at the topmost pyramid level. This process is repeated on successively finer scales, until the full resolution of the image is reached.

5) The final displacement vector field is determined as the sum of the vector fields determined at each of the individual levels. This displacement vector field is used to construct the final morphed image.

Before the pyramid matching is carried out, the observed and synthetic satellite images are preprocessed to a suitable form. First, the observed image is interpolated to the model grid. Second, for the application considered here, where IR imagery is used as a proxy to locate precipitating cloud features, a threshold brightness temperature is applied to mask out the land surface and shallow nonprecipitating cloud.

The subsampling factor F, which defines the topmost pyramid level and thus the maximum distance over which features can be matched, must be chosen based on the physical phenomenon of interest. A forecast and an observed feature that are closer than this distance will be considered to be the same feature, but displaced in space (and a displacement vector will be computed), whereas features separated by larger distances will be assumed to be unrelated.

The mean magnitude of the displacement vector field gives an indication of the forecast quality, provided that there are features in the fields available to be matched. However, in a case where a forecast cloud feature cannot be matched because the observed one is outside of the search environment (or not forecast at all), which subjectively would be a forecast failure, the mean displacement vector is zero and gives no indication of forecast quality. A perfect forecast would also give a zero mean displacement. Therefore, a second quantity is needed to account for such cases. A large value of the mean squared difference of the observed and morphed images in a local region indicates such a forecast failure. Combining both quantities, the following local forecast quality measure (LFQM) is proposed:

$$\mathrm{LFQM} = \max(c_1 \cdot \mathrm{DIS}, c_2 \cdot \mathrm{LSE}).$$

In a case where the forecast and observed features can be matched, the magnitude of the displacement vector DIS characterizes the forecast quality, whereas when matching features cannot be found within the search distance, the local squared error (LSE) between both fields is used. The displacement is scaled with the maximum possible displacement (the corner-to-corner size of the region that is tested at the topmost pyramid level; see step 2 of the algorithm above),

$$c_1 = \mathrm{DIS}_{\max}^{-1} = \left(\sqrt{2} \cdot 2^{F+2}\right)^{-1},$$

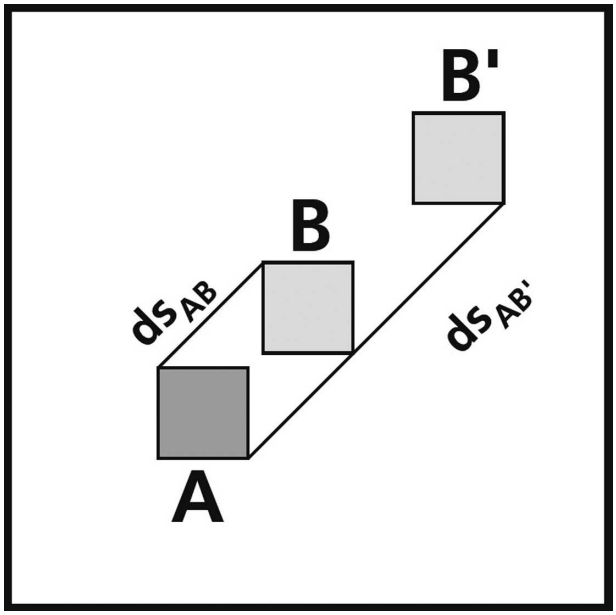

FIG. 1. Schematic illustrating the setup of an idealized case: A represents the "observed" feature, B (B′ etc.) the displaced "model forecast" feature. The pyramidal image matcher is morphing the increasingly separated ($\mathrm{ds}_{\mathrm{AB}}$) feature B (B′ etc.) toward the fixed feature A.

and the LSE with the maximum squared observed brightness temperature difference,

$$c_2 = (\mathrm{BT}_{\max} - \mathrm{BT}_{\min})^{-2}.$$

In this way the error assigned when no feature is available to match within the search distance (one image has $\mathrm{BT}_{\min}$, indicating cold cloud, while the other has $\mathrm{BT}_{\max}$, indicating no cloud) is approximately equal to the error associated with a displacement equal to the maximum search distance. The LFQM can then be averaged over a verification area A to give an overall forecast quality measure:

$$\mathrm{FQM} = \frac{1}{A}\sum_{A} \mathrm{LFQM}.$$

The measure FQM attains zero for a "perfect" forecast; that is, the terms DIS and LSE are both zero.

The behavior of the pyramid matching algorithm is illustrated for a simplified case. Consider two identical features, say two squares (20×20 pixels each), originally at the same location, then increasingly separated from each other. Such an idealized case is schematically depicted in Fig. 1. Let the subsampling factor F be 4 (i.e., one pixel element at the coarsest scale of the pyramid contains 16×16 pixels), and let the finest scale be the original pixel resolution. The largest search distance corresponds to 42 pixels at the original resolution using

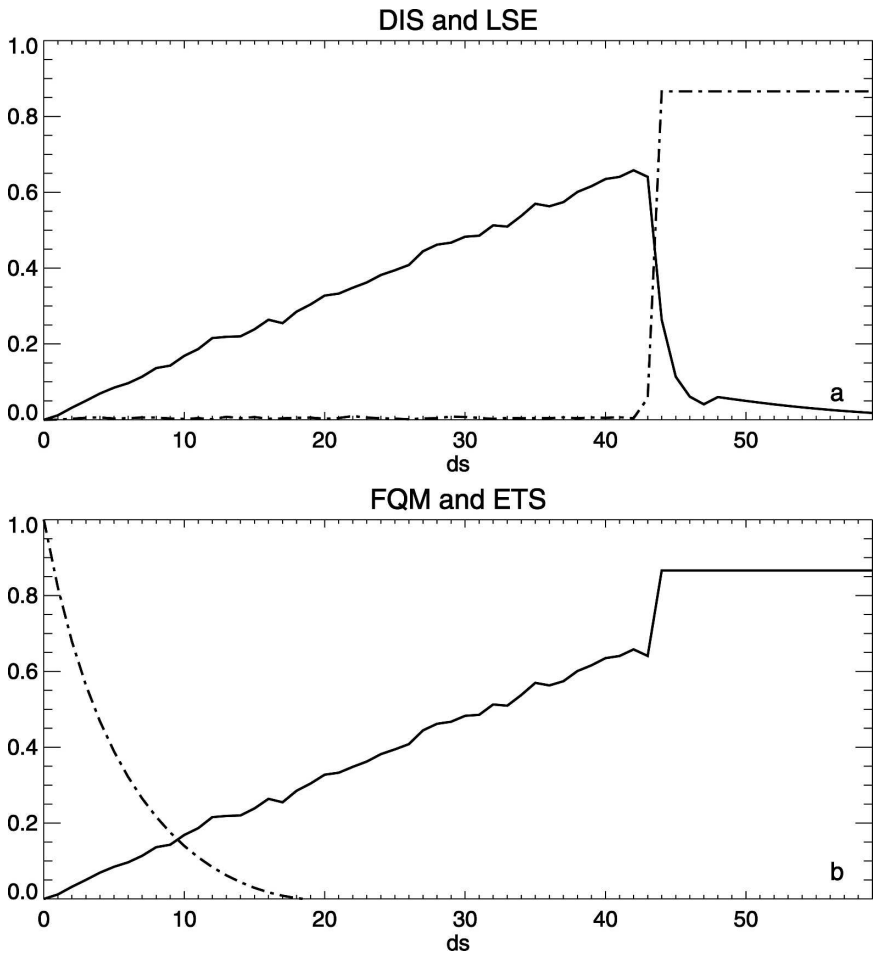

FIG. 2. Four measures as a function of separation distance (in pixels) for the simplified case in Fig. 1: (a) the mean displacement DIS (solid) and the LSE (dashed–dotted) of the observed and morphed forecast features and (b) the novel measure FQM (solid) and the ETS (dashed–dotted) as a reference measure.

a 5×5 Gaussian kernel. The values of DIS, LSE, FQM, and a conventional measure, the equitable threat score (ETS; Ebert et al. 2003) are depicted in Fig. 2 for incremental separations of both features up to 60 pixels. The mean magnitude of the displacement vector field steadily increases with increasing separation distance up to the largest search distance, $\mathrm{DIS}_{\max} = 42$, where it sharply decreases when no match between the two features can be accomplished. At this distance, the LSE of the fixed feature A and the morphed feature B abruptly rises from close to zero to a constant value. Combining both parameters in the FQM gives an almost monotonically increasing curve with increasing separation distance. When the features are separated by more than the maximum search distance ($\mathrm{ds} > 42$), the LSE component of FQM masks the failing of the matching marked by the step in FQM. The DIS component does not go exactly to zero, since when there is no feature B within the matching distance, the algorithm tries to shrink feature A by producing a convergent vector field. The normalization factors in DIS and LSE could be tuned to produce an FQM that increases to exactly unity for distances larger than $\mathrm{DIS}_{\max}$, at least for this particular idealized problem, but this sacrifice of simplicity is unlikely to bring any measurable advantage when the FQM is applied to complex observed images.

Consider in particular two different scenarios with the feature B separated by 25 and 35 pixels, respectively, from the fixed feature A. Using FQM allows an objective distinction in quality between both scenarios; FQM attaining 0.5 and 0.7, respectively. A smaller displacement results in a lower value of FQM, which agrees well with human intuition. Conventional scores like the BIAS or ETS fail to describe the location errors in this example. The BIAS attains 1 for all separations (not shown), whereas the ETS decreases from 1 (when

both features are identical, ds = 0) down to 0 at a distance equal to the feature size (ds = 20; Fig. 2b). At larger distances there is no information contained in the conventional score ETS. Traditional measures like BIAS and ETS fail to capture such differing location errors. In contrast, the new measure FQM is able to incorporate the advantages of distance- and amplitude-based measures thus allowing an objective evaluation of forecast quality based on image comparison.

3. The ensemble forecasting system and satellite data

a. The regional ensemble system

The new forecast quality measure is applied to regional ensemble forecasts generated using the Consortium for Small Scale Modeling Limited-area Ensemble Prediction System (COSMO-LEPS; Molteni et al. 2001; Marsigli et al. 2001). In COSMO-LEPS, the global European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) provides initial and boundary conditions for the Deutscher Wetterdienst's (DWD) high-resolution nonhydrostatic Lokal-Modell (LM; Steppeler et al. 2003). A brute-force approach that uses every member of the global ensemble is likely to be inefficient, however, since much of the variability in the global ensemble may be confined outside of the domain of the limited-area model. Therefore, global forecasts that are similar in the target region are clustered, and only a single representative set of boundary conditions is used for each cluster. Marsigli et al. (2005) found that most of the variability in the 51-member ECMWF EPS for a region centered on central Europe can be retained by as few as 10 members. Probabilistic forecasts are routinely generated by assigning to each LM integration a weight proportional to the population of the cluster from which the representative member (providing initial and boundary conditions) is selected (Marsigli et al. 2005).

For the present study, COSMO-LEPS is configured as follows: one 51-member ECMWF EPS T255L40 experiment (about 80-km horizontal resolution with 40 vertical levels; operational suite in 2005) started at 1200 UTC 8 July 2002 provides the initial and boundary conditions (6 hourly) for the limited-area model. Following the operational COSMO-LEPS procedure, a cluster analysis is performed to determine 10 clusters (out of 51 global EPS forecasts) that are similar in the target region (Europe) in the forecast range of interest (+24 to +36 h) based on the horizontal wind components (u, v), geopotential (Φ), and specific humidity (q) at three pressure levels: 500, 700, and 850 hPa.

Subsequently, LM experiments (+72 h forecast range) are performed for each representative member using DWD's operational LM domain (corresponding to the area shown in Fig. 3; 325 × 325 grid points with 35 vertical levels), with a horizontal resolution of 7 km. The LM prognostic variables are the three wind components, temperature, pressure perturbation, specific humidity, cloud liquid water, cloud ice, rain, and snow. The sedimentation flux of rain and snow is the product of the effective fall velocity and the density and is treated using a three-dimensional semi-Lagrangian advection scheme (Baldauf and Schulz 2004). Moist convection is parameterized after Tiedtke. For comparison purposes, a deterministic forecast is run starting just before the convective period at 0600 UTC 9 July 2002 and driven with deterministic forecast data of the global model.

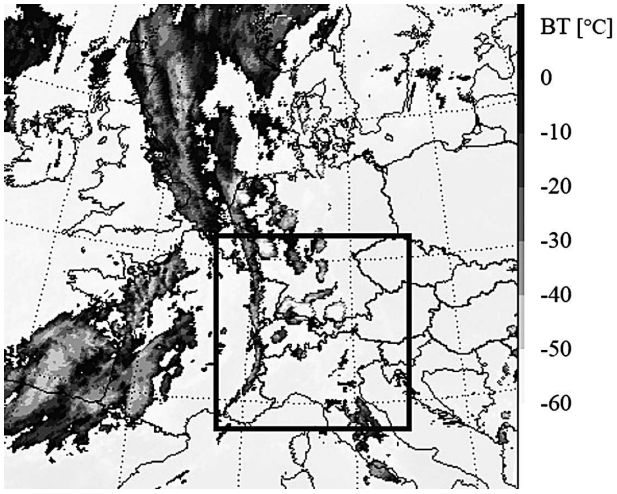

FIG. 3. Observed *Meteosat-7* IR image projected onto the LM domain at 1600 UTC 9 Jul 2002. The black rectangle denotes the area displayed in Fig. 4 for which the ranking (see Table 1) is done.

b. Synthetic satellite imagery

A forward model is needed to project model states into observation space. In the LM, synthetic satellite imagery is generated using the fast radiative transfer model for the Television Infrared Operational Satellite (TIROS) Operational Vertical Sounder (RTTOV-7), which allows for fast simulation of brightness temperatures (BTs) for various satellite radiometers [e.g., the *Meteosat-7* Visible and Infrared Imager (MVIRI) and the *Meteosat-8* Spinning Enhanced Visible and Infrared Imager (SEVIRI); Saunders et al. 1999]. The input variables provided by LM are atmospheric profiles of temperature and specific humidity, various cloud properties (cloud cover, cloud liquid water, cloud ice), the

specific content of snow, and surface properties (skin temperature, temperature and specific humidity at 2 m, land–sea mask). The LM is configured to include precipitating snow crystals in the computation of synthetic satellite imagery leading to better agreement among the observed and simulated upper-tropospheric clouds (Keil et al. 2006).

c. Satellite data

Satellite data from the geostationary satellite *Meteosat-7* are used in the present study.

Meteosat-7 is currently positioned at the equator on the Greenwich meridian. The satellite's spatial resolution is 5 km at nadir and about 8 km at 50°N, which is comparable to the current resolution of the LM. The infrared window channel (IR; 10.1–13.0 μm) is sensitive to cloud amount throughout the troposphere. The uncertainty caused by calibration error amounts to 2–4 K (Köpken 2001) and is significantly smaller than the cloud signals investigated in the present study.

4. Prefrontal summertime convection in Bavaria: A case study

To give an initial idea of the behavior of the system with real meteorological data, a case study will now be presented. The performance of the pyramid matching algorithm will be examined in detail for a single image, and then the resulting FQM for a set of ensemble forecasts valid at this time will be compared to a subjective ranking of forecast quality. Finally, the time evolution of the FQM at hourly intervals over a 36-h period for the various ensemble members will be considered, to see if there is evidence of a persistence of quality that could potentially be useful in a forecast context.

a. Subjective evaluation of ensemble members

Ahead of an eastward-propagating cold front, prefrontal convection developed in the northern Alpine region in the afternoon of 9 July 2002. *Meteosat-7* IR imagery shows the cloud signature of two strong convective cells north of the Alpine chain at 1600 UTC (Fig. 3). The western cell was reinforced by Alpine orography, while the eastern cell was initiated in the northern Alps of southern Bavaria. The elongated, north–south-oriented, cloud band across eastern France marks the cold front. Two hours later, both convective cells merged and formed a mesoscale convective system covering all of Bavaria, with a diameter of approximately 400 km. This particular episode represents a typical case of mesoscale convective systems in the northern Alpine region (Hagen et al. 2000).

The corresponding model-forecast synthetic IR imagery of the regional ensemble system is displayed on a subdomain (900×900 km^2) at 1600 UTC in Fig. 4. Figures 4a–j show the 10 individual members, and Fig. 4k shows the short-range deterministic forecast (see section 4d). Most of the clusters forecast some cloud associated with the cold front in the western half of the domain, while there are large differences with respect to the prefrontal convection. Visual intercomparison of the observed and synthetic IR imagery shows that the ensemble generally underestimates the cloud amount with considerable differences among the individual members. Clearly, clusters 5 and 8 are able to reproduce the convection and the corresponding cloud signature in the region to some extent, while others fail entirely in forecasting convective activity and even the frontal cloudiness (e.g., cluster 9).

An important first test of the proposed FQM will be to verify that it captures these subjective differences. As a control, a small survey was performed, where eight research scientists were asked to rank the 10 ensemble members, based on their agreement with the observed image in location structure and amount of cloud (Table 1). The mean rank correlation between pairs of colleagues is high (0.82), confirming a good agreement among themselves and pointing toward a clear ranking of the clusters.

b. Objective evaluation

Since the primary interest in the present study is to quantify mesoscale position errors of convective storms that may occur in similar synoptic environments, the pyramid matcher has been applied with the following configuration: (i) observed and forecast imagery are both projected on a subdomain covering 128×128 LM grid points (see Fig. 4), and (ii) the IR brightness temperature threshold is set to -20°C (i.e., only middle- and upper-tropospheric cloud structures shall be considered). To confine the matching to subsynoptic distances, the coarse-grain pixel elements are defined to contain eight by eight LM grid points per pixel element (subsampling factor $F = 3$), so that the maximum search distance $\mathrm{DIS}_{\max}$ extends to about 300 km. The coarse-grained image at the topmost pyramid level consists of 16×16 pixel elements. It is to be expected that the results will be somewhat sensitive to this choice of maximum search distance. If it is chosen too small, the error measure will only respond to nearly coincident features, while if it is too large, the matching will connect physically unrelated features. The distance over which convective systems share the same synoptic environment is determined in midlatitudes by the Rossby radius of deformation, which is the length scale over

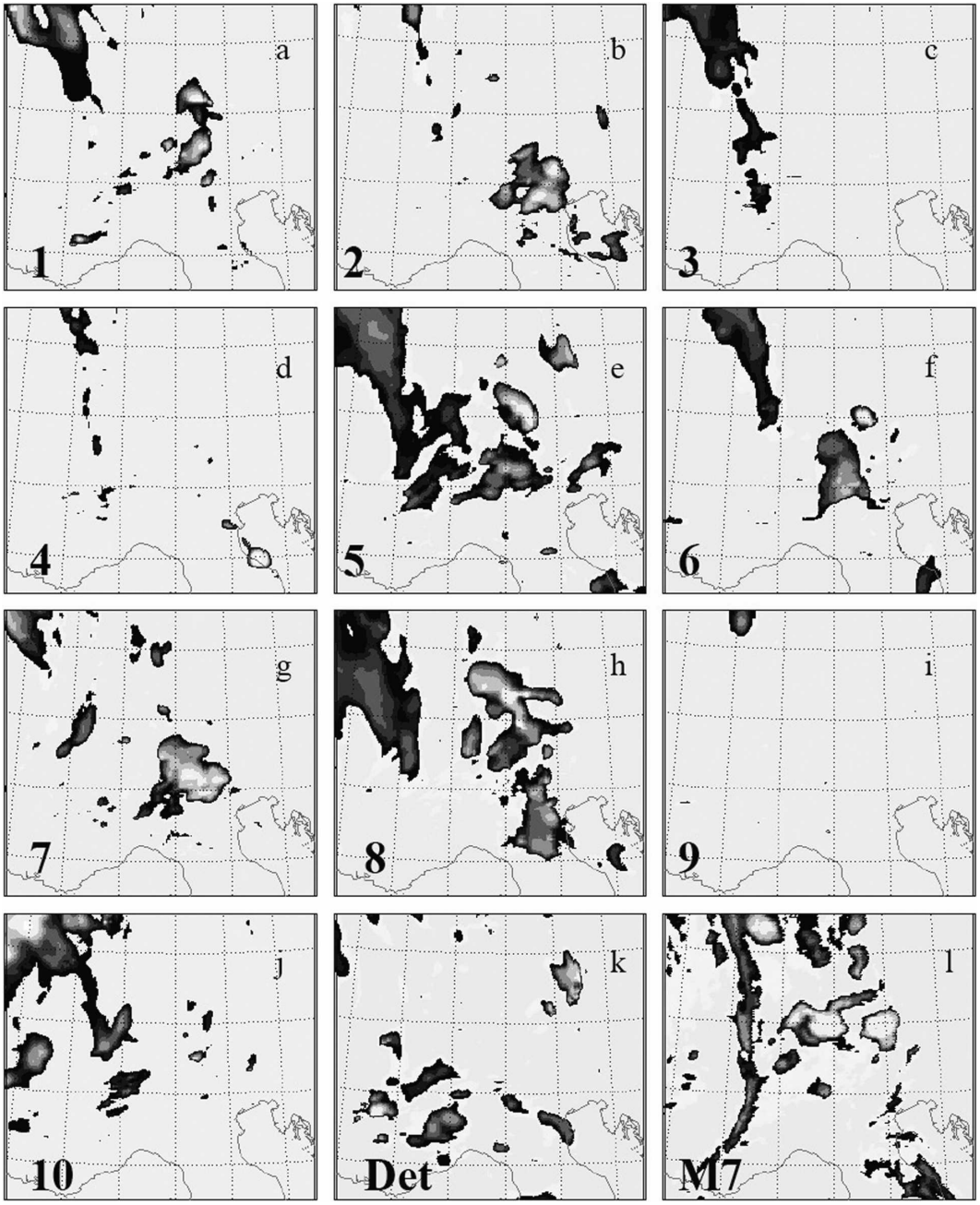

FIG. 4. Forecast IR synthetic satellite imagery at 1600 UTC 9 Jul 2002 (LM +28 h forecast range): (a)–(j) the individual members of clusters 1–10, (k) the short-term deterministic forecast, and (l) the *Meteosat-7* observation for comparison.

which significant horizontal temperature gradients can be maintained (by geostrophic balance). A fixed maximum distance of 300 km was chosen in this paper, and the results of the sensitivity tests where this distance was altered will be noted in section 4e. The next logical step would be to use a local radius of deformation based on the total vorticity. This would yield a shorter distance near features like fronts and prevent mismatches such as prefrontal convection being matched to embedded convection within the front. Ideally, one would use an anisotropic maximum distance that would allow longer matching distances in the direction away from the front.

An example sequence to help visualize the function-

TABLE 1. Ranking of the 10 clusters at 1600 UTC 9 Jul 2002 according to (i) the subjective visual evaluation of eight forecasters, (ii) the objectively calculated forecast quality measure FQM, and (iii) the cluster population (number of members per cluster).

Rank	1	2	3	4	5	6	7	8	9	10	Corr
Forecasters	5	8	1	6	7	10	3	2	4	9	0.81
FQM	5	8	6	7	1	10	4	4	9	2	0.92
Population	3	2	5	6	4	7	10	1	8	9	0.05

ing of the algorithm is presented in Fig. 5. In the top panel of both rows in Fig. 5 the observed BT, the forecast BT, and the displacement vector field at different scales for cluster 8 at 28-h forecast range are displayed. First, the topmost pyramid level with coarse-grain imagery from the Meteosat, the LM, and the displacement vector field superimposed on the forecast LM image are depicted (Figs. 5a–c). The displacement arrows show an inhomogeneous, partly converging vector field indicating a northeastward displacement of the cloud feature in the image's western part, a southward displacement in the central north, and a westward displacement near the southeastern boundary. Second, at the next smaller scale (one pixel element contains four by four LM grid points) the observed and the forecast BT field, in which the displacement resulting from the coarse grain is already applied, are compared. The resulting vectors are superimposed on the forecast image in Fig. 5f, pushing the cloud field in the image's center to the southwest. The final result of applying the pyramid matcher is depicted in the third row in Fig. 5, which shows the observed image (Fig. 5g), the forecast super-

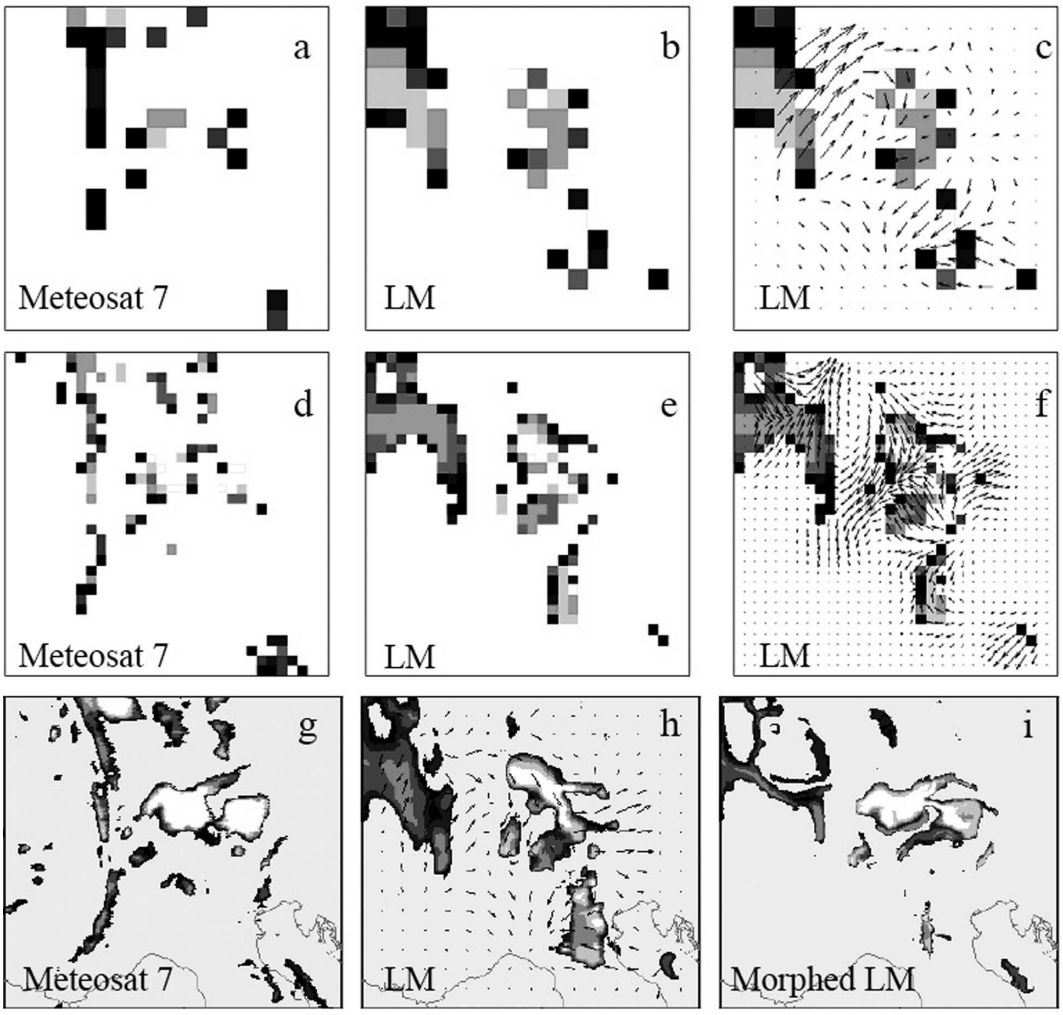

FIG. 5. Sequence of differently grained (a), (d) observed and (b), (e) forecast BT fields; (c), (f) are (b), (e) superimposed with the displacement vector field for cluster 8 at 28-h forecast range. (g) The observed BT field, (h) the forecast BT field superimposed with displacement vectors summarized over all pyramid levels, and (i) the morphed forecast BT field displayed at full resolution.

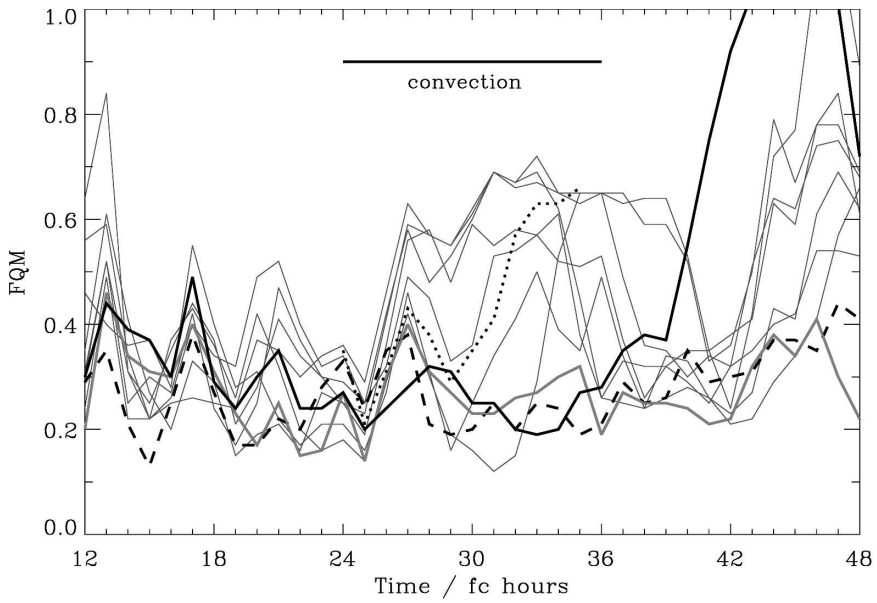

FIG. 6. Time series of FQM for the 10 ensemble members and the short-term deterministic forecast extending from 0000 UTC 9 Jul to 1200 UTC 10 Jul 2002 (forecast range +12 h until +48 h). Highlighted are ensemble members 5 (dashed), 7 (solid gray), and 8 (solid black), which are referred to in the text. For comparison the short-term deterministic forecast (initialized at 0600 UTC 9 Jul with a forecast range of +6 to +18 h) is also displayed (dotted). The time interval corresponding to the life cycle of the convective cell discussed in the text is marked.

imposed with the displacement vectors summarized over all scales (Fig. 5h), and the morphed forecast field (Fig. 5i). The vectors determined at the largest scale dominate the overall displacement, moving the frontal clouds in the western part toward the northeast and the prefrontal convective clouds in the image's center toward the south. Converging displacement vectors shrink the forecast cloudiness south of the Alps considerably. Moreover, it is evident that pixel elements close to the image boundary (for instance near the cold front in the northwestern part) are kept fixed. This boundary zone encompassing two pixel elements at the largest scale; hence, a frame of 16 model grid points is excluded in the calculation of the FQM.[1]

The application of the pyramidal image matcher on the cloud pattern at 1600 UTC and the calculation of FQM allow an objective ranking of the 10 clusters. Comparison of the human ranking and the objective one shows that the new measure provides a good error measure (Table 1). The rank correlation between the average human and the objective ranking attains 0.92, confirming the consistent results of both rankings. In contrast, the ranking based on the COSMO-LEPS cluster population (number of members per cluster) shows no correlation with the other rankings for this episode of strong convection. This last result is perhaps to be expected, since the cluster population is based on synoptic information over central Europe, and is not necessarily well correlated with the cloud information in a local region evaluated in the FQM.

c. *Time series of forecast quality*

Next, the time evolution of the FQM for the various ensemble members is considered, to see if there is evidence of a persistence of quality over some forecast period. On 9 July 2002, convection initiated north of the Alps at about 1400 UTC (+26 h forecast). Ten hours later (+36 h forecast), the mesoscale convective system starts to move out of the subdomain, while a new cloud system enters from the west. The temporal evolution of the FQM is displayed for a 36-h interval (+12 to +48 h forecast range) in Fig. 6. Apparently, this period can be subdivided into three distinct episodes: a preconvective period with only scattered cloudiness in the subdomain (+12 to +24 h forecast range), a convective period marked by considerable spread of FQM (+24 to +36 h forecast range), and a succeeding postconvective period with the advent of a new synoptic-scale weather system. The lack of prominent cloud features during the preconvective period precludes any well-defined ranking.

[1] The computational requirements of the algorithm are small. On a PC, this computation takes less than a minute per forecast–observation pair.

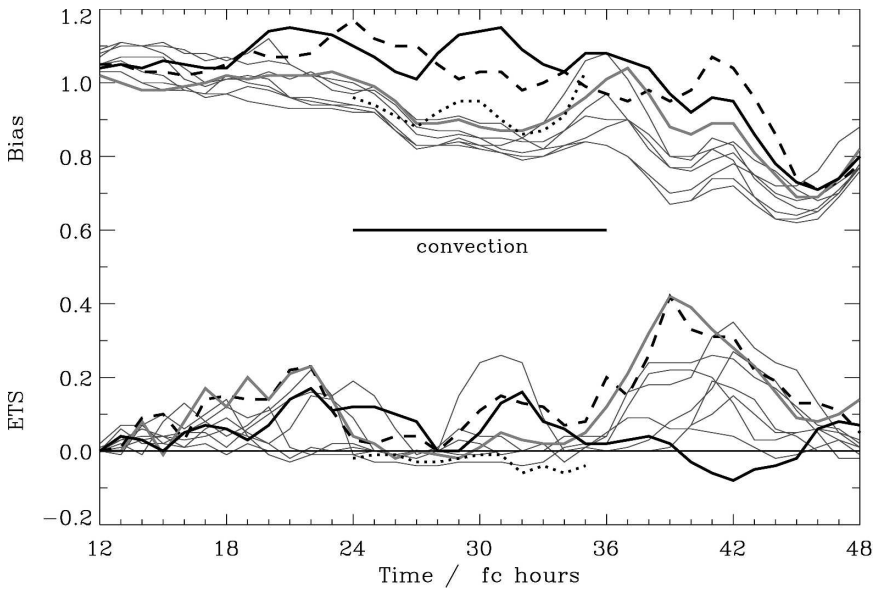

FIG. 7. Same as in Fig. 6 but for conventional (top) BIAS and (bottom) ETS scores.

However, this is clearly possible within the convective period when clusters 5 and 8 outperform the other ensemble members (small FQM values). This is consistent with the previous discussion for the snapshot at 1600 UTC. Likewise, forecasts of clusters identified as "bad" members at 1600 UTC remain bad throughout the convective period. While there is considerable persistence in FQM visible in Fig. 6 through the convective period, there is little apparent relation to forecast quality before or after that period. The new synoptic-scale weather system moving into the domain from 0000 UTC onward developed into the damaging Berlin storm on 10 July 2002 (Gatzen 2004). The cloud signature of this second storm is best captured by clusters 5 and 7, while the forecast skill of cluster 8 strongly decreases (Fig. 6).

Figure 7 shows the time variation of the conventional scores BIAS and ETS of all ensemble members. Apparently, most of the forecasts underestimate the cloud occurrence (BIAS < 1) through the convective period. The LM is known to have a negative bias in cloud amount (Keil et al. 2006), and the increase of the negative bias with time seen in Fig. 7 presumably represents a relaxation toward the model climatology, albeit heavily modulated by the various cloud features in the small region used to compute the score at any given time. The ETS values are generally poor (ETS < 0.4). The ETS attains larger values for clusters 5 and 7 during the second storm, which is in agreement with the new measure (small FQM). Consider, for example, the various scores at 1600 UTC (with corresponding imagery shown in Fig. 4). At this 28-h forecast range the BIAS indicates that clusters 5 and 8 predict the relative amounts of cloudiness correctly whereas the ETS indicates a forecast failure of all members with respect to the exact location. In contrast, the proposed FQM incorporating an amplitude- and a distance-based component allows a distinct ranking of the different forecasts.

d. Forecast quality of ensemble members versus a short-range deterministic forecast

Generally, a short-term deterministic forecast is thought to exhibit a better forecast quality than forecasts started at earlier times due to the more recent data that have been incorporated into the data assimilation process. The question of whether a short-term deterministic forecast (started at 0600 UTC 9 July; same as in the LM configuration) has a higher forecast skill than individual ensemble members (started at 1200 UTC 8 July, i.e., 18 h earlier) can be assessed using the new measure.

In the short-term deterministic forecast (0600 UTC +18 h forecast range) a large convective cell is present at 1600 UTC (Fig. 4k). Visual comparison with the observations indicates an eastward displacement of the convection of about 100 km. Since this is within the search distance, the matching algorithm is able to morph the misplaced forecast cloud toward the observed cloud features. The resulting FQM of the deterministic forecast is 0.38, a value that indicates medium forecast quality compared with the 10 ensemble members (Table 2 and Fig. 6).

TABLE 2. FQM of "best" ensemble member, deterministic forecast, and "worst" ensemble member for different search distances at 1600 UTC 9 Jul 2002.

Subsampling factor F	4	3	2	1
Maximum search distance (km)	600	300	150	75
"Best" member	0.24	0.18	0.23	0.36
Deterministic forecast	0.37	0.38	0.44	0.52
"Worst" member	0.57	0.58	0.58	0.59

e. Dependence of FQM on maximum search radius

Results were also examined for other search distances (Table 2). Ensemble members that fail to produce a feature even within the largest radius (worst members) result in the same FQM for all search distances (0.57–0.59). The best members, which produce clouds in the vicinity of the observed system (good forecast), give a smaller FQM that is largely independent of search radius (0.18–0.23), except at the smallest radius of 75 km. The deterministic forecast is an example of an intermediate situation, where some features can be matched if a large search radius is used (producing an intermediate FQM of 0.37–0.38), but less matching is possible with smaller search radii (FQM decreasing to 0.52). This behavior is expected, and underlines the fact that there is no intrinsically optimal choice for the maximum search radius. As noted previously the choice in this paper of 300 km is motivated by the physical argument that convective features separated by synoptic distances (500 km or more) form in different environments, and are unlikely to be simply displaced.

5. Conclusions and outlook

A novel forecast quality measure FQM is proposed based on application of the pyramid matching algorithm to observed and model-forecast satellite imagery. The new measure is composed of the displacement necessary to match both fields (to measure differences in location), and the local squared difference of the observed and morphed forecast brightness temperature fields (to account for cloud features in one field that cannot be matched to any feature in the other field).

The FQM has been applied within the framework of a regional ensemble system to evaluate the individual members and to rank them. The displacement-based error measure allows an objective evaluation of forecast quality based on image comparison. The new method provides a plausible measure of forecast error, which is consistent with subjective rankings, as shown for a typical episode of prefrontal summertime convection in Bavaria. The cluster population that is operationally used to weight the individual members in COSMO-LEPS exhibits a poor indication of local skill on 9 July 2002. Using the displacement-based error measure, individual ensemble members compare better with the observations than does a short-term deterministic forecast throughout the convective period. In contrast, traditional measures like BIAS and ETS fail to capture differing location errors of convective cloudiness. Importantly, "good" ensemble members remain good throughout this convective episode.

If this case is typical, such a persistence of skill could suggest that a selected best member, or a probability distribution based on weighted ensemble members, may be a useful tool for short-range forecasting, particularly if, as for the case shown here, the best member is superior to more recent deterministic forecasts, or the ensemble mean. It should be noted that although the short-range deterministic forecast presented in this paper used more recent synoptic information (provided by the more recent initial global analysis), the recent satellite data used in the image matching were not assimilated, and a deterministic forecast using these data might be better. It will not necessarily be better, however, since the deterministic forecast may suffer from an initial spinup period, in contrast to the fully spunup ensemble members, and may not benefit greatly from an attempt to assimilate small-scale cloud information. Experience suggests that information about convective precipitation may not be retained by the model for long if the background forecast does not have the right conditions to support the convection (Leuenberger and Rossa 2007). It does however suggest a hybrid approach, where the best ensemble member is selected as background for a short-range data assimilation–forecast cycle. This possibility is currently being investigated. However, this is speculative and has to be proven in many other cases.

The pyramidal image matching method has also been used as the basis of a cell-tracking method (ZMT), which could be used to identify timing errors, similar to Davis et al. (2006b). This possibility is currently being explored.

A systematic evaluation of the performance of the regional ensemble system using the proposed measure and the suggested ensemble refiltering approach is planned during the 3-month period of the Convective and Orographically induced Precipitation Study (COPS) field experiment that took place in the summer 2007 in a low-mountain area in southwestern Germany–eastern France, which is characterized by high summer thunderstorm activity and particularly low skill of numerical weather prediction models (Wulfmeyer et al. 2005). This will allow for a more reliable comparison between the various error measures, in-

cluding the establishment of climatological baseline values of FQM based on comparing random images, and an assessment of the utility of the displacement-based measure for capturing predictability information from the ensemble.

Acknowledgments. We gratefully acknowledge Hermann Mannstein (DLR) for providing the pyramid matching algorithm. The authors thank Andrea Montani and Chiara Marsigli (both ARPA-SIM) for their guidance on running COSMO-LEPS. The numerical experiments have been performed using ECMWF and DWD resources. Meteosat data are copyrighted by EUMETSAT. This work has been conducted within the DAQUA project founded by the Deutsche Forschungsgemeinschaft (DFG).

REFERENCES

Baldauf, M., and J.-P. Schulz, 2004: Prognostic precipitation in the Lokal-Modell of DWD. *COSMO Newsletter,* No. 4, Consortium for Small Scale Modeling, 177–180. [Available online at http://cosmo-model.cscs.ch/public/various/newsLetters/newsLetter04/chp9-7.pdf.]

Brewster, K. A., 2003a: Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part I: Method description and simulation testing. *Mon. Wea. Rev.,* **131,** 480–492.

——, 2003b: Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part II: Application to a severe storm outbreak. *Mon. Wea. Rev.,* **131,** 493–507.

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

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.,* **11,** 141–154.

Davis, C. D., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.,* **134,** 1772–1784.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.,* **134,** 1785–1795.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

——, U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc.,* **84,** 481–492.

Gatzen, C., 2004: A derecho in Europe: Berlin, 10 July 2002. *Wea. Forecasting,* **19,** 639–645.

Hagen, M., H.-H. Schiesser, and M. Dorninger, 2000: Monitoring of mesoscale precipitation systems in the Alps and the northern Alpine foreland by radar and rain gauges. *Meteor. Atmos. Phys.,* **72,** 87–100.

Hoffman, R. N., and C. Grassotti, 1996: A technique for assimilating SSM/I observations of marine atmospheric storms: Tests with ECMWF analyses. *J. Appl. Meteor.,* **35,** 1177–1188.

Keil, C., A. Tafferner, and T. Reinhardt, 2006: Synthetic satellite imagery in the Lokal-Modell. *Atmos. Res.,* **82,** 19–25.

Köpken, C., 2001: Monitoring of METEOSAT WV radiances and solar stray light effects. Research Rep. 10, EUMETSAT/ECMWF Fellowship Programme, 46 pp.

Leuenberger, D., and A. Rossa, 2007: Revisiting the latent heat nudging scheme for the rainfall assimilation of a simulated convective storm. *Meteor. Atmos. Phys.,* doi:10.1007/s00703-007-0260-9.

Mannstein, H., H. Huntrieser, and S. Wimmer, 2002: Determination of the mass flux in convective cells over Europe. *Proc. 2002 EUMETSAT Meteorological Satellite Conf.,* Dublin, Ireland, EUMETSAT, 264–269.

Marsigli, C., A. Montani, F. Nerozzi, T. Paccagnella, S. Tibaldi, F. Molteni, and R. Buizza, 2001: A strategy for high-resolution ensemble prediction. Part II: Limited-area experiments in four Alpine flood events. *Quart. J. Roy. Meteor. Soc.,* **127,** 2095–2115.

——, F. Boccanera, A. Montani, and T. Paccagnella, 2005: The COSMO-LEPS mesoscale ensemble system: Validation of the methodology and verification. *Nonlinear Proc. Geophys.,* **12,** 527–536.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.,* **77,** 2637–2649.

Molteni, F., R. Buizza, C. Marsigli, A. Montani, F. Nerozzi, and T. Paccagnella, 2001: A strategy for high-resolution ensemble prediction. Part I: Definition of representative members and global-model experiments. *Quart. J. Roy. Meteor. Soc.,* **127,** 2069–2094.

Muller, J.-P., M.-A. Denis, R. D. Dundas, K. L. Mitchell, C. Naud, and H. Mannstein, 2007: Stereo cloud-top heights and cloud amount retrieval from ATSR2. *Int. J. Remote Sens.,* **28,** 1921–1938.

Saunders, R., M. Matricardi, and P. Brunel, 1999: An improved radiative transfer model for assimilation of satellite radiance observations. *Quart. J. Roy. Meteor. Soc.,* **125,** 1407–1425.

Steppeler, J., G. Doms, U. Schättler, H. W. Bitzer, A. Gassmann, U. Damrath, and G. Gregoric, 2003: Meso-gamma scale forecasts using the nonhydrostatic model LM. *Meteor. Atmos. Phys.,* **82,** 75–96.

Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.,* **110,** D08111, doi:10.1029/2004JD005395.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Wulfmeyer, V., A. Behrendt, C. Kottmeier, and U. Corsmeier, Eds., cited 2005: COPS: Convective and Orographically-induced Precipitation Study. [Available online at http://www.uni-hohenheim.de/spp-iop/documents/051109_COPS_SOD_final.pdf.]

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.,* **133,** 279–294.

Zepeda-Arce, J., and E. Foufoula-Georgiou, 2000: Space–time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.,* **105,** 10 129–10 146.