Meteorol. Z. (Contrib. Atm. Sci.), Vol. **28**, No. 3, 203–213 (published online January 21, 2019)

7IVMW

Application of an object-based verification method to ensemble forecasts of 10-m wind gusts during winter storms

Philipp Zschenderlein[1,2], Tobias Pardowitz[2,3*] and Uwe Ulbrich[2]

[1]Karlsruhe Institute of Technology, Germany
[2]Freie Universität Berlin, Germany
[3]Hans-Ertel-Centre for Weather Research, Berlin, Germany

(Manuscript received October 27, 2017; in revised form November 15, 2018; accepted November 26, 2018)

Abstract

The object-based method SAL (Structure, Amplitude and Location) was adapted for investigating the errors of forecasts of extreme 10-m wind gusts associated with winter storms in Germany. It has been applied to a statistically downscaled version of the 51 member ECMWF (European Centre for Medium Range Weather Forecasts) operational ensemble forecast. The horizontal resolution of both downscaled data and of the German weather service's operational analysis data used for verification is 7 km. Forecast errors are subdivided in terms of storm intensity, location and extent. After identifying a set of storm events, objects of moderate and intense 10-m wind gusts were identified with a local percentile-based threshold (90th percentile for moderate and 98th percentile for intense gust objects). Depending on the intensity of the storm, the gust objects differ in terms of size, shape and intensity. The characteristics of the ensemble forecasts of 10-m wind gusts can basically be assessed in two different ways. Individual forecast members can be evaluated with respect to the location, intensity and extent of the gust field, and then address the ensemble characteristics by the score distributions. Alternatively, the gust fields' location, intensity and extent can be evaluated by directly using the ensemble mean forecast instead of the individual members. The results of the identified set of storms clearly indicate a high case-to-case variability in the predictability of 10-m wind gusts objects, particularly when focusing on the structure of intense wind gust objects. It is found, that the gust fields' location and overall intensity can be better estimated from the ensemble mean forecast, compared to the individual forecast members. From a forecaster's perspective this means, that a storms' location and intensity can be well estimated by considering the ensemble mean wind forecasts. Considering the structure of the gust objects, results are different. While for longer lead times, there also seems to be a benefit from applying ensemble averaging, at short lead times the ensemble mean forecast performs equally or worse than most of the individual forecast members. The amplitude error is often the smallest component of the three error types. The findings are particularly relevant when deriving warning information, by giving guidance to forecasters when interpreting ensemble forecasts for severe storms.

Keywords: spatial verification, object-based verification, winter storms, ensemble forecasts

1 Introduction

Forecast verification serves the developers of the operational weather forecasting systems by providing information on model characteristics and forecast deficits. It also assists forecast users in their interpretation of forecasts and the knowledge about inherent uncertainties (Davis et al., 2006a).

Besides the classical grid-point based verification methods, the emergence of high resolution models and convection permitting regional models motivated the development of new spatial verification methods during the last decade, primarily focusing on quantitative precipitation forecasts (Ebert and McBride, 2000; Casati et al., 2004; Davis et al., 2006a; Keil and Craig, 2007; Wernli et al., 2008). The main motivation for the emergence of these methods is well described by the so-called double penalty problem (Jolliffe and Stephenson, 2012). In case of high resolution predictions, extreme wind gusts might well be forecasted in terms of their structure, amplitude and timing, however with a slightly incorrect position. Classical grid-point based verification measures would classify this forecast as very poor, i.e. many misses and false alarms (double penalty problem). Using object-based (spatial) approaches, forecast errors may instead be assessed in terms of the object characteristics directly, i.e. characterizing and comparing structure, intensity and location of objects. An illustrative example of the double penalty problem and the failure of classical verification measures can be found in Davis et al. (2006a). An overview of already existing spatial verification methods and their classifications is given in Ebert (2008) and Gilleland et al. (2009). Spatial verification techniques are not only applied to precipitation fields (Fox et al., 2016; Gofa et al.,

*Corresponding author: Tobias Pardowitz, Hans-Ertel-Centre for Weather Research, Institute of Meteorology, Freie Universität Berlin, Carl-Heinrich-Becker-Weg 6–10, 12165 Berlin, Germany, e-mail: tobias.pardowitz@met.fu-berlin.de

2017; Wernli et al., 2008), but also to other meteorological variables. Weniger and Friederichs (2016) applied the SAL technique to spatial fields of cloud cover and spectral radiances. Mittermaier et al. (2016) used the MODE technique (Davis et al., 2006a; Davis et al., 2006b) to verify the position, timing and intensity of jet cores, surface highs and lows. It is therefore worth applying object-based verification techniques to wind gusts, which is done in this study. As an important extension to the above-mentioned methods, it should be mentioned that it might be desirable to include the temporal development in order to represent errors in the timing of a forecasted event (Gilleland et al., 2009; Zimmer and Wernli, 2011; Mittermaier et al., 2016).

The object-based methods give an additional and different perspective on forecast quality and predictability. By applying the SAL technique, the present study complements a recent investigation, revisiting the synoptic-scale predictability of winter storms (Pantillon et al., 2017).

These different views on the predictability of storms and respective forecast errors are justified by the large variety of users of weather forecasts with broadly different information needs. While point-wise forecasts are certainly needed in a wide range of applications, event-based forecasts and warnings are essential for users, particularly from logistics and aviation.

Damages to man-made and natural structures depend critically on the strength of local wind gusts making it extremely important to derive accurate storm forecast and warning information from numerical weather prediction systems. There are numerous studies linking the occurrence of local severe wind gusts with damages to buildings both in deterministic approaches (Klawa and Ulbrich, 2003; Heneka and Ruck, 2008; Donat et al., 2011) as well as probabilistic approaches (Heneka and Hofherr, 2010; Prahl et al., 2012; Pardowitz et al., 2016). It has been found that maximum sustained wind gusts can serve well to describe the patterns of damage occurrences. However, results may differ with regards to the exact dependence of storm damages on increasing wind speeds. Generally this dependence is found to be strongly non-linear with several studies assuming a cubic dependence of wind speed, as described originally in Klawa and Ulbrich (2003). Another common notion is that damages occur for severe wind situations. Even though this finding might vary depending on the actual damage data, it has been found that exceedances of the 98$^{\text{th}}$ percentile of local wind speeds serves as an appropriate threshold to identify potentially damaging wind situations (Klawa and Ulbrich, 2003; Leckebusch et al., 2008).

Considering the importance of high resolution ensemble forecast systems, the application of object-based verification methods might be particularly important. For grid-point based forecasts, the interpretation of ensembles in terms of occurrence probabilities for certain events is well established (Anderson, 1996; Broecker and Smith, 2008). For spatial objects or features such as storms and wind gusts, the interpretation and analysis of ensemble forecasts is less obvious. Information about the intensity, extent and location can, for example, be derived from individual forecast members and can be evaluated by analysing the ensemble distributions of such features. Instead of using all the individual forecast members, the spatial verification method can alternatively be applied to the ensemble mean forecast. However, one should consider that averaged fields may not be a physically consistent representation of a storm event due to, for example the smoothing of the wind field. The limitations of interpreting the ensemble average have been described in several studies (Cheung, 2001; Surcel et al., 2014). It can be noted that by definition, the ensemble mean provides smoother fields and can thus provide guidance on the large-scale flow properties (Cheung, 2001). However, for the small scale characteristics in the case of extreme events probabilistic approaches should be chosen (Cheung, 2001; Surcel et al., 2014).

In this study, we adapt and apply an object-based verification method to ensemble forecasts of 10-m wind gusts associated with winter storms in Germany. For this reason, we employ the so-called SAL (Structure, Amplitude and Location) technique by Wernli et al. (2008) which is an object-based quality measure, originally developed for the verification of quantitative precipitation forecasts. By applying it to ensemble forecasts, we investigate the forecast uncertainty which is expressed in terms of object features, by offering a complementary view of the forecast errors as expressed in typical grid-point measures. Such object-based views might ultimately contribute to an improvement in the understanding of underlying processes and their representation in forecast models. Additionally, we demonstrate how to make use of the ensemble information with object-based methods, in order to analyse how the results of this method can differ using either the individual members or the ensemble mean only.

The remainder of this paper is structured as follows. Section 2 gives an overview of the data and methodology, which was used in this study. Results obtained by means of the object-based verification methodology are described in Section 3, followed by Section 4 giving conclusions that can be drawn from this study.

2 Data and methods

2.1 COSMO-EU analyses

As a surrogate for ground truth, we use the analyses operationally conducted by the German Weather Service (DWD) from the COSMO-EU Model (Consortium for small-scale Modelling). The period of available data is from 1 January 2006 to 31 December 2011. COSMO-EU was developed at the DWD and was used as a regional model for Europe with a horizontal resolution of 7 km (Schulz and Schättler, 2014; Doms

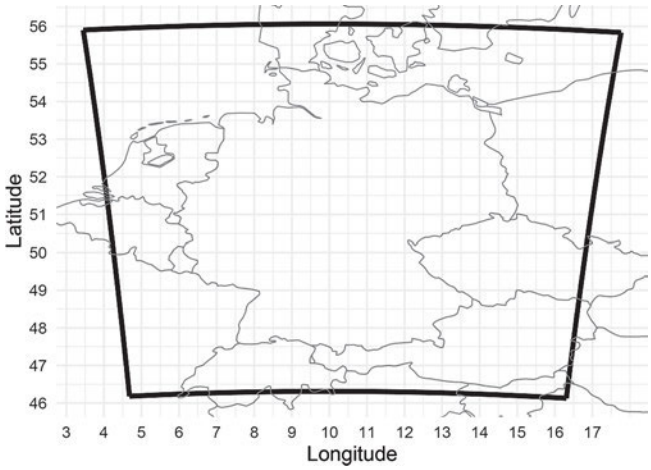

Figure 1: Verification domain considered, as given by the domain for which the statistical downscaling of ensemble forecasts has been performed.

and Baldauf, 2015). It is non-hydrostatic and used for numerical weather predictions. The COSMO model is based on the primitive equations and has 40 vertical layers from 10 m to 24 km in height. The investigated variable in this study is the daily maximal wind gust, which is calculated as the daily maximum of the hourly maximal wind gusts at 10 m height. The wind gusts are estimated from the simulated variables using diagnostic schemes as described in Schulz and Heise (2003) and Schulz (2008), including a parametrization scheme for turbulent gusts and convective wind gusts. Final estimates of the hourly maximum wind gusts are calculated as the maximum of both gusts. We only used a part of the COSMO-EU domain, because the forecast dataset was available for a smaller region only (Figure 1).

2.2 Downscaled ECWMF ensemble forecasts

In this study we employ forecast data from the ECMWF EPS (Ensemble Prediction System) which is statistically downscaled to a grid with the same horizontal resolution as the COSMO-EU analyses. A detailed description of the employed downscaling methodology and the dataset can be found in Pardowitz et al. (2016). As input for the statistical downscaling, 6 hourly output of instantaneous 10 m wind speed of the 50 perturbed ECMWF ensemble members is used, which was operationally produced between November 2000 and January 2010. Each forecast is integrated over 15 days, but the horizontal resolution is reduced after forecast day 10. We thus confine all analyses to the first ten forecast days during which the resolution is kept constant. Specifics of the ECMWF-EPS can be found in various references (Molteni et al., 1996; Palmer et al., 1998; Leutbecher and Palmer, 2008; Buizza et al., 1999; Palmer et al., 2009).

The ECMWF-EPS forecasts are statistically downscaled to the COSMO-EU-resolution of approximately 7 km, as described in Kruschke (2015). The output of the statistical downscaling method is the daily maximum wind gust at 10 m height, which is used in this study. The statistical downscaling is done with a multiple linear regression method. This method assesses the relationship between the surface gusts of the higher to the coarser scale of the corresponding ECMWF-EPS forecast (Kruschke, 2015). By choosing skilful predictors of the COSMO-EU model, a regression equation is created. These predictors are more specifically the EPS surface winds scaled by the respective climatological 98th percentiles. Due to changes in the EPS resolution a subsequent interpolation (first order conservative) to the coarsest EPS resolution (T159) is performed (compare Pardowitz et al., 2016). More details on the statistical downscaling method can be found in Pardowitz et al. (2016) and Kruschke (2015).

Compared to the typical dimensions of winter wind storms, the investigated model domain is relatively small, which might have an influence on the verification results presented in Section 3. However, since the statistically downscaled EPS forecasts are only available for the domain indicated in Figure 1, no systematic analysis of this dependence can be performed.

2.3 Storm identification

As a prerequisite for the application of an object-based verification, an identification of a storm set needs to be done. This identification is based on near surface wind gusts (10 m height) from the COSMO-EU analysis data, also serving as the verification reference in this study.

Due to usually stronger wind speeds at higher altitudes or above sea surfaces, the daily maximum wind gusts were normalized with their local 98th percentile. "Local" implies that the percentile was calculated at each grid point and for the whole available period 2006–2011. The outcome of this identification procedure will be referred to as the *normalized wind gusts*.

The storm identification was restricted to the winter season (October to March) for the period 2006–2010, where both the forecast and analysis data were available. A similar criteria to that used in the study of Leckebusch et al. (2008) was used to perform the identification of a set of winter storms. At least an area equivalent to 400 km × 400 km in the COSMO-EU analysis needs to exceed the local 98th percentile of daily maximum wind gusts. This corresponds roughly to 10 % of the considered verification domain (compare Figure 1). The outcome of this identification procedure will be referred to as the *storm set analysis* (compare Figure 2) and forms the basis for the object-based verification described in section 2.5. By means of this criteria, 82 storm days were identified during the period of January 2006 to January 2010 within the COSMO-EU analyses.

2.4 Storm occurrence verification

Similar to the storm identification described in section 2.3, we additionally identified storms in the forecast data for the storm occurrence verification. The criterion

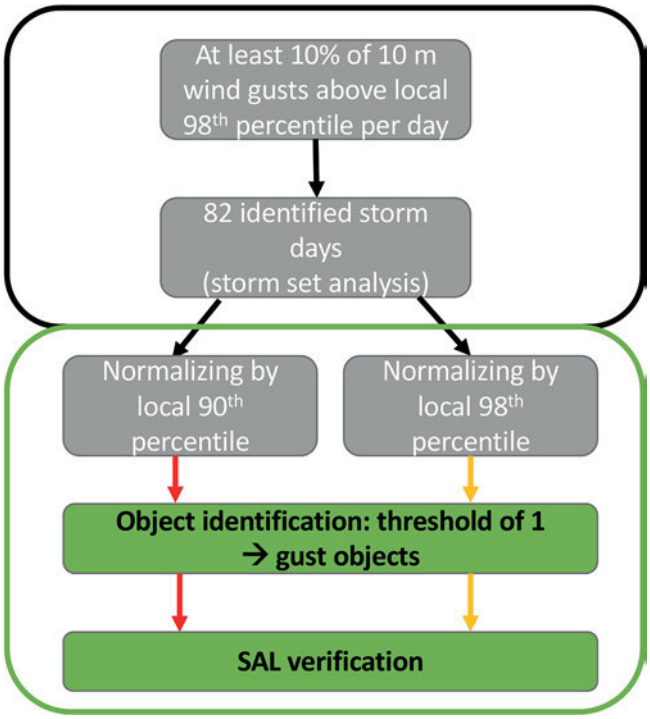

Figure 2: Schematic depicting the storm identification from the COSMO-EU analysis and the SAL verification procedure. The upper panel describes the storm identification (section 2.3). The object-based verification (lower panel), as described in section 2.5, is performed on the basis of the wind gusts normalized by the local 90th and 98th percentile, respectively. These objects are referred to as the *gust objects*.

was the same as for the analysis data, i.e. at least an area equivalent to 400 km × 400 km needs to exceed the local 98th percentile of daily maximum wind gusts of the forecast. With this storm set in the forecast and analysis data, a storm occurrence verification is performed simply by comparing whether a storm was observed (regardless of its location and intensity in the verification domain) and whether the storm was forecasted or not (i.e. dichotomous yes/no forecast). For this purpose, the contingency table can be derived, including hits (event observed and forecasted), misses (event observed and not forecasted), false alarms (event not observed but forecasted) and correct rejects (event neither observed nor forecasted). Typical quantities to address the quality of such forecasts are the false alarm ratio (FAR) and the hit rate (H). FAR is the ratio of false alarms compared to the number of events forecast, ranging between 0 (no false alarms) and 1 (only false alarms). H is the fraction of observed events which have been forecasted correctly ranging from 0 (no event correctly forecasted) to 1 (all events correctly forecasted).

2.5 Applying SAL to gust objects

For the whole *storm set analysis* of 82 identified storm days (Figure 2 upper panel) within the investigation period, the errors regarding the extent, location and intensity of the relative 10 m daily maximum wind gusts were addressed by means of the SAL method proposed in Wernli et al. (2008). Note that the object-based verification was only based on storms which were observed. Before applying the SAL technique, the wind gusts were normalized with their local 90th, or alternatively 98th percentiles of the analysis and the forecast, respectively. Model biases are therefore excluded. Note that the percentiles for the forecast are calculated for each forecast lead time separately, hence the value of the respective percentiles can differ between the forecast lead times. Then, all grid points with values above the equal to the threshold value of 1 are considered part of one of the gust objects which contribute to the storm event (Figure 2 lower panel). If many contiguous grid points exceed the threshold, they belong to one object. If these grid points are separated from each other, they refer to different objects. The SAL calculations were based on the normalized wind gusts and their identified gust objects (Figure 2 lower panel). These moderate and intense 'gust objects' (90th and 98th percentile respectively) are, in other words, objects of relative daily maximum wind gusts above a certain threshold. This nomenclature follows that of Wernli et al. (2008), where they used 'precipitation objects'.

Comparing gridded forecasts and observations for a specified area, the SAL method calculates three error components with respect to the object's structure, amplitude and location. The amplitude error A (see eq. 2 in Wernli et al., 2008) represents the normalized difference of the domain-averaged relative wind gusts, irrespective of the objects identified. Values of A are within −2 (relative wind gusts are observed but not forecasted) and +2 (relative wind gusts are forecasted but not observed). Positive (negative) values of A denote an overestimation (underestimation) of the forecasted relative wind gusts. The location error L (see eq. 4 and 6 in Wernli et al., 2008) consists of two parts (L_1 and L_2). L_1 represents the distance of the normalized difference between the center of mass of the relative wind gusts of the whole domain. Also for this score, an object identification is not needed. As an example, an L_1 error of 0.1 within a domain size of 1000 km equals a displacement of the center of mass of 100 km. L_2 is the normalized difference of each gust objects' center of mass from the center of mass of the whole domain. Location errors are only positive, ranging from 0 to +2; the individual ranges for L_1 and L_2 are 0 to +1. The last component is the structure error S (see eq. 9 in Wernli et al., 2008), which compares the difference of the gust objects' "volume". Values are within -2 and +2, with positive (negative) values denoting objects that are too large and flat (small and peaked). With respect to wind gusts, positive S errors indicate a too large gust field, and/or a too smooth gust distribution in the forecast. For example, convection in cold fronts may be underestimated. Negative S errors indicate a too small gust field and/or a too large variability of gusts within. For the calculation of the SAL components, we used the *R* software package *SpatialVx* (Gilleland, 2015).

In the following, two ways of processing the ensemble information were applied. Firstly, the individual ensemble members were analysed in terms of the forecast errors of predicted gust objects. Forecast uncertainty can then be expressed in terms of the distribution of error parameters (structure, location and amplitude). Secondly, gust fields of the ensemble mean are considered and corresponding error parameters are directly inferred. Of course in the latter, no information about the ensemble uncertainty can be derived. The ensemble mean was calculated from the raw fields of the ensemble members and then normalized with the local 90^{th} and 98^{th} percentile, respectively. The percentiles of the ensemble mean were calculated for each forecast lead time separately, just as for the individual ensemble members. Hence, the resulting errors with higher forecast lead time are not a lead-time dependant bias, but an effect of the higher uncertainty in the forecast model.

3 Results

3.1 Case study – winter storm 'Quinten'

As an example of the application of the SAL method to gust forecasts we consider the winter storm 'Quinten', which affected southern parts of Germany on 10 February 2009. Figure 3 (a) shows the daily maximum wind gusts from the COSMO-EU analyses for this date with high values in the southwestern parts of the considered domain. Also clearly visible are orographic dependences, especially around the southern Rhine valley, with particularly high gusts in the Vosges area of France and the Black Forest in Germany, and rather low values in the valley itself. Figure 3 (b) shows the gust objects for local percentiles (exceedances of the 90^{th} percentile in grey, the 95^{th} in blue, the 98^{th} in dark blue and the 99^{th} in green). Although still not fully homogeneous, this representation allows for a much better detection of a coherent structure of the gust field. Also in this representation, the North Sea is clearly excluded from the gust object, although in terms of absolute gust speeds considerably high values of about 20–25 ms^{-1} are present. However, in these areas, such wind gusts are frequent and are not to be considered as a part of the gust object.

Figure 3 (c,f,i) shows the ensemble mean of forecast gust speeds for 10 February 2009 for lead times of one, four and seven days. The overall structure of high gust speeds in southwestern regions of the domain can be found accordingly, however absolute values of gust speeds may vary for certain regions due to local orographic influences on the gust field. Such differences are disregarded when considering local percentile exceedances (Figure 3 (d,g,j)). Note that the percentiles are computed using the ensemble mean values for each forecast lead time separately. The gust objects, as identified in the analyses, can clearly be identified within the ensemble forecasts at all considered lead times. However, intensity, extent and location do vary.

Figure 3 (e,h,k) summarizes results from the SAL method applied to the gust objects (exceedances of the local 98^{th} percentile). Here, grey histograms show the distributions of structure, amplitude and location errors for the 51 individual ensemble members, while the black dashed vertical line indicates the errors for the ensemble mean forecast. For a lead time of one day, all individual members showed rather similar and small errors in all three SAL parameters. The structure error is found to be slightly positive, indicating a gust field which is too large and flat. Amplitude error is found to be positive also, which indicated that the field average of the relative wind gusts is higher compared to the analyses, which also indicated that the overall intensity is overestimated. With growing lead time, individual ensemble members diverge, leading to a growing spread in the structure, amplitude and location errors. Interestingly, the amplitude and location error for the ensemble mean exhibited rather stable and low values, which indicated good forecast quality with respect to these parameters. The structure error for the ensemble mean forecast however seems to grow considerably, being about -1.2 at a lead time of seven days. The definition of the structure error in Wernli et al. (2008) states that negative structure errors are associated with too small and too peaked (in our case too gusty) forecasts. However, fields in the ensemble mean are generally smoothed. Therefore, negative structure errors in the ensemble mean imply a too small forecasted gust object. Generally, a negative structure error together with a small amplitude error can result if (i) the storm (i.e. gust) field is too small and (ii) in some areas of the storm field the gust velocities are overestimated and in other areas underestimated compared to the verification reference. The small size of the forecasted gust object is nicely visible comparing for example the dark-blue colours in Figure 3 (b) and (j). Of course, the results for this case study cannot be generalized; a more systematic analysis will be discussed in the following sections.

3.2 Storm occurrence verification results

Figure 4 shows results of all storm events identified in analyses and forecasts (storm definition: at least 10 % of the verification domain above the local 98^{th} percentile; compare sections 2.3 and 2.4). For a lead time of one day, the FAR shown in Figure 4 (top) is about 0.2 (about 20 % of cases in which a storm is forecasted are false alarms) with results for individual ensemble members varying by about ±0.08 (8 %).

The hit rate (in Figure 4 (bottom)) for a lead time of one day is about 0.6 (60 % of all observed storms were forecasted correctly), again results varying for individual members by about ±0.09 (9 %). With growing lead time, FAR grows and at the same time H decreases. For a lead time of nine days, individual ensemble members show an increased FAR of about 72 % ± 8 %, while the hit rate decreases to about 23 % ± 7 %. This means that for a lead time of nine days, more than 70 % of all cases

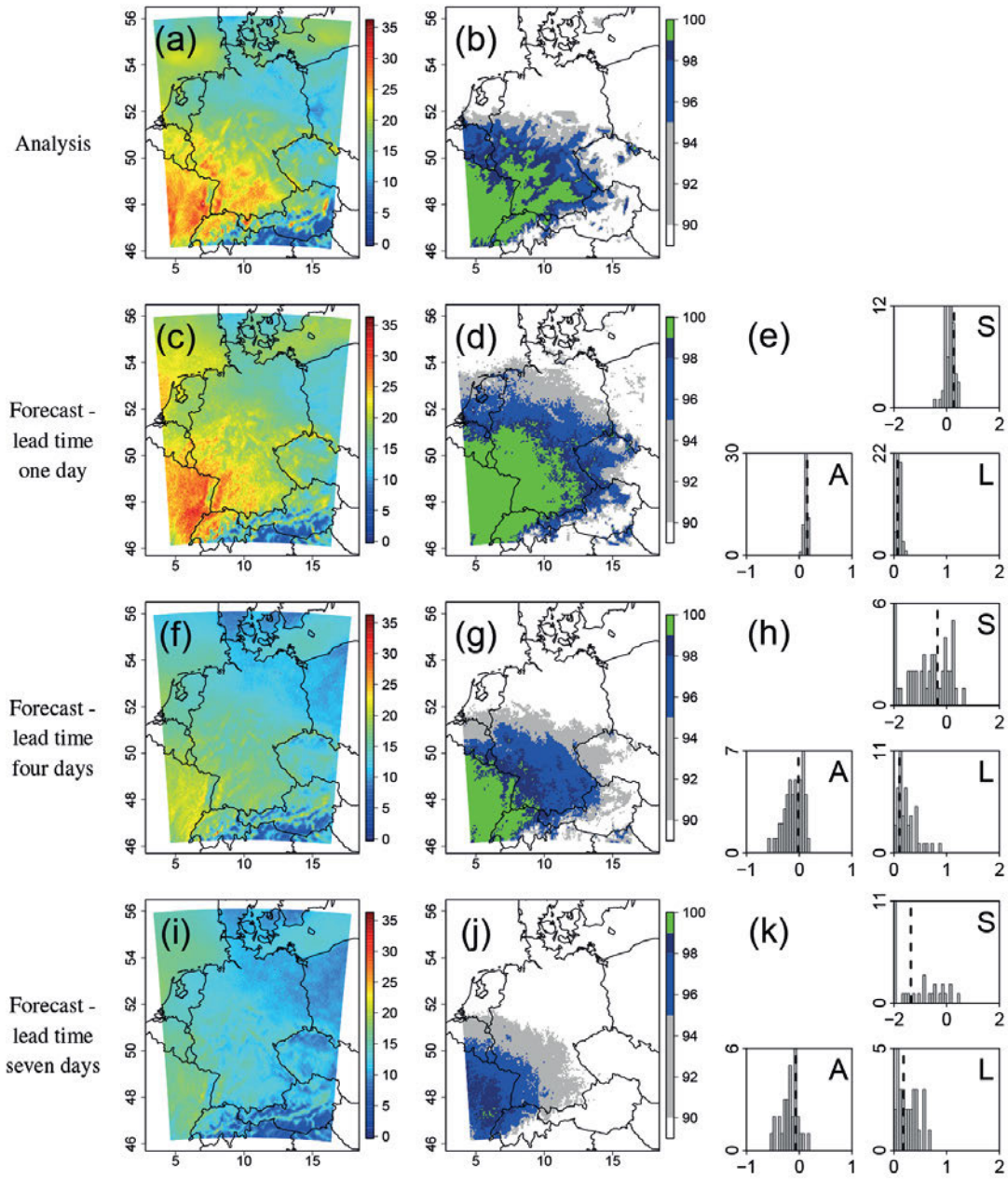

Figure 3: Case study of winter storm 'Quinten' (10 February 2009). The first column depicts absolute maximum wind gust speeds, the second column depicts exceedances of the different thresholds considered (90th percentile in grey, the 95th in blue, the 98th in darkblue and the 99th in green) and the third column show the SAL results for three different forecast lead times. Daily maximum wind gusts from COSMO-EU analysis (a,b); ensemble mean forecasts one day lead time (c,d), four days lead time (f,g) and seven days lead time (i,j). Resulting distributions of structure, amplitude and location errors for lead times one, four and seven days (using the 98th percentile as threshold) (e,h,k). Distribution of ensemble members are shown as grey histograms while result for the ensemble mean forecast is indicated by the vertical dashed line.

of a storm forecast are false alarms and at the same time, only 23 % of all observed storms were forecasted.

The forecast quality with respect to both FAR and H increases considering the ensemble mean forecast (red dots in Figure 4). For a lead time of nine days, the FAR is reduced by about 15 % to about 58 % and at the same time the hit rate is increased by about 17 % to 40 %. Thus, a clear benefit of using the ensemble mean forecast can be identified with respect to the forecast of storm occurrences within the considered domain.

Of course, other definitions of a storm event can be applied here. A lower threshold focuses on larger storm areas, whereas a higher threshold focuses on areas of extreme and damaging wind speeds. The results differ with respect to the absolute values of FAR and H. When using a lower threshold (90th percentile), the FAR is found to be lower and at the same time the hit rate is found to be higher. Similarly for a higher threshold (99th percentile) FAR is higher and hit rate is lower. Regardless of which threshold is used to identify the storm events,

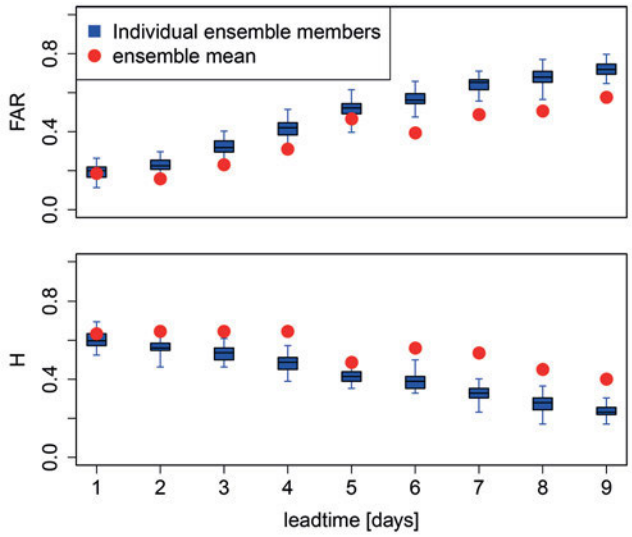

Figure 4: Storm occurrence verification results for all identified storms in the analysis and forecasts. Criterion for the detection of a storm event is that 10 % of the considered domain exceed the local 98th percentile of wind speeds. Results for the false alarm ratio (FAR, top) and the hit rate (H, bottom) in dependence of the lead time are shown for individual members as blue box plots and the ensemble mean as red dots.

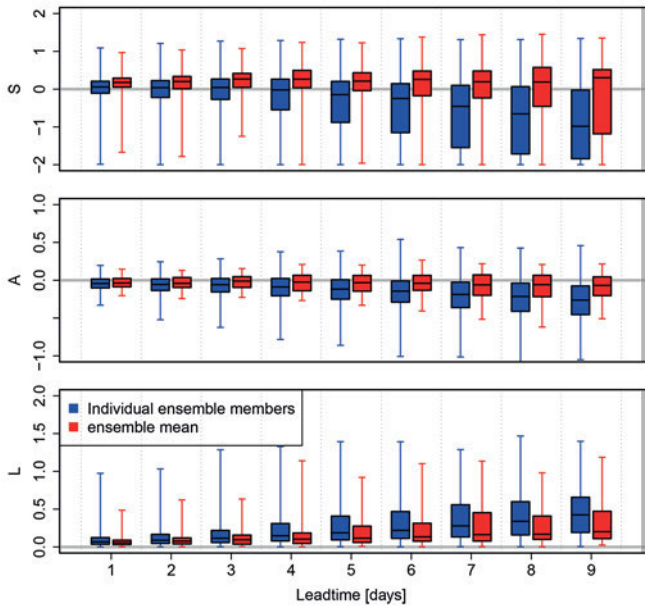

Figure 5: Distributions of structure error (top), amplitude error (middle) and location error (bottom) in dependence of lead time for individual members (blue box plots) and the ensemble mean (red box plots) for the 82 storm identified storm events. The wind gusts in the analysis and the forecast were normalized by their local 90th percentile and a threshold of 1 was used for object identification (i.e. all grid points above the local 90th percentile will be identified as gust object).

it was found that analysing the ensemble mean forecast yields a lower (better) FAR as well as a higher (better) hit rate compared to analysing all individual ensemble members.

3.3 Object-based verification results

Figure 5 shows results for the object-based SAL method applied to the 82 storm events (see *storm set analysis* in Figure 2), which are identified in the COSMO-EU analyses. For each of the storm events, the SAL method is calculated for individual ensemble members as well as for the ensemble mean forecast. As a threshold for a gust object within the SAL method, the 90th and 98th local percentiles are used. Figure 5 shows results for the lowest threshold, i.e. rather large objects of high relative wind gust speeds. Blue boxes in Figure 5 show the distribution of the structure error (top figure), amplitude error (middle figure) and location error (bottom figure) when considering the individual member forecasts for each of the 82 storm events.

For a lead time of one day, structure errors are found to be rather small, however individual member forecasts for individual storm events range from −2 (forecast gust objects much too small and peaked) up to 1 (forecast gust objects much too large and flat). In the definition of the SAL method in Wernli et al. (2008), negative structure errors can be interpreted as too small and too peaked objects. In this study, we used as forecast model a statistically downscaled model. Of course, this model is not able to resolve convective processes compared to dynamically downscaled atmospheric models. Therefore, negative structure errors imply that the forecasted gust objects are in general too small, or in other words, the winter storm has a too small size in the verification domain. Another possibility is that several too small objects are forecasted. There is a slight tendency towards positive structure errors at lead times of one day. Amplitude errors are rather small and mostly negative at short lead times. This means that the forecasts tend to produce lower relative wind gusts (in other words: a lower storm intensity) on average over the considered domain. Also, variations for individual storm events and forecast members are comparatively small. Location errors are found to be rather small for the first few days, bearing in mind however that a location error of 0.1 implies a displacement of the centre of mass of about 100 km in a domain size of 1000 km (if only one gust object is identified in the verification domain). The whiskers in Figure 5 (bottom) indicate that for individual storm events and forecast members location error can be up to 1, even for short lead times of one day. Both amplitude and location error steadily increase with growing lead times, i.e. the amplitude error grows to more negative values (stronger underestimation of average storm intensity) and the location error to more positive values, indicating a larger offset of storm centres with respect to the analysed gust fields. While the structure error for short lead times shows a tendency toward positive values (too large and flat storm fields), the tendency reverses with increasing lead time. Thus, at lead times beyond five days there is an increasing tendency toward smaller gust fields within the individual ensemble member fore-

casts. At a lead time of nine days, this is expressed by the fact that 75 % of individual forecasts for all 82 storm fields showed a negative structure error.

The amplitude errors of the individual ensemble members are negative for all lead times. Amplitude errors become larger (i.e. more negative) with increasing lead time. They are consistently smaller when the ensemble mean gust forecasts (based on all 82 storm forecasts) for all lead times are evaluated instead of the individual ensemble members (red box plots in Figure 5, middle). For individual storm events positive amplitude errors are found. The difference is particularly large for higher lead times. At first glance, it is not intuitive that for longer forecast lead times the individual members most frequently underestimate the gusts while for part of the storms the ensemble mean overestimated the gust speed (Fig. 5, middle). To shed more light on this feature, we determined for one grid point in the verification domain (50° N, 10° E) the distribution of the whole individual members and the ensemble means' gust velocities for different forecast lead times (not shown). For short forecast lead times, both distributions have nearly the same shape, but for longer forecasts lead times, the distributions deviate from each other. The ensemble mean's skewness for higher lead times is reduced compared to the distribution of the respective ensemble members. Thus, the values of the percentiles in the ensemble mean forecast are decreased compared to the percentiles of the individual ensemble members. While the percentiles of the individual ensemble members are relatively stable for the forecast lead times, the skewness of the ensemble means gust values at individual grid points and for individual storms decreases with lead time. Considering now one of the individual storms, the absolute ensemble mean gust velocities are usually relative low, meaning an underestimation of the mean of observed gust velocities. As we compute the forecast error in terms of deviations from the percentile value (i.e. the value determined from the model climatology for each forecast lead time), the amplitude errors are reduced. Hence, although the absolute gust speeds can be low, the amplitude error compared to the model climatology (i.e. the percentiles) for each lead time is relatively low (Figure 5, middle). Therefore, our results should be interpreted in the way that the amplitude is relatively well forecasted compared to the model climatology of the ensemble mean, but are generally underestimated in an absolute sense.

The location errors are considerably reduced when considering the ensemble mean forecast (Figure 5, bottom). This result is expected because the number of objects reduce in the ensemble mean due to the smoothing. That reduces the L_2 component and therefore the whole location error. Also, L_1 of the ensemble mean is reduced compared to the L_1 errors of the ensemble members under the premise of no far outliers and that the observed center of mass of the ensemble mean is located within the area spread by the center of masses of the ensemble members. Again, for longer lead times this reduction is particularly large. However, for the location error

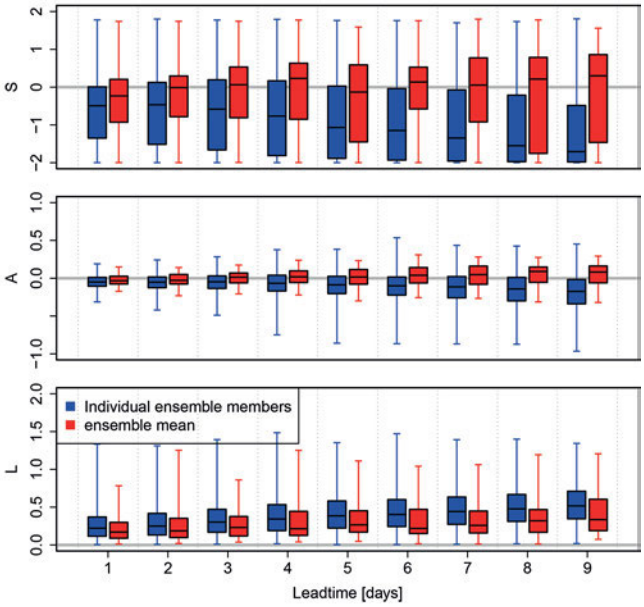

Figure 6: Same as Figure 5, despite that the wind gusts were normalized by their local 98th percentile.

there are individual storms for which the ensemble mean forecast shows large location errors. Interesting results can be found for the structure error (Figure 5, top) when considering the ensemble mean forecast, which is different to the individual ensemble member forecasts, in this case there is a consistent tendency towards positive structure errors (i.e. too large and flat gust objects). This can be explained by the fact that the ensemble averaging of wind gusts implies a smoothing of the gust fields. While individual forecast members may contain gust objects at different locations, the ensemble mean forecast field contains smoothed gust objects which are both too large and too flat at the same time. In terms of absolute values, however, there seems to be a tendency that structure errors for the ensemble mean are small compared to the individual ensemble member forecasts. Yet, particularly for large lead times, there is a large variation of resulting structure errors for each of the 82 storm events. However, a structure error of an ensemble mean can be difficult to interpret, because it can contain gust structures, which are physically inconsistent.

The results discussed above apply to the 90th percentile as threshold for an 'object'. By using the 98th percentile, the considered objects are smaller, ultimately containing only regions of severe wind gusts. The results obtained by using the 98th percentile as threshold are shown in Figure 6. In comparison, certain similarities can be found, as well as some differences.

Results for the location errors (Figure 6, bottom) are qualitatively similar to the previous results, however, with a considerably larger variation. This indicates in this case, that the results depend much more on the individual storm considered. In general, the result is confirmed that location errors are smaller when considering the ensemble mean forecast, particularly for longer

lead times. Amplitude errors (Figure 6, middle) for gust objects defined by the 98th percentile are again mostly negative, again with increasingly more negative values at higher lead times. It was mentioned in section 2.5 that the amplitude error does not depend on the object identification threshold. The differences in the amplitude errors in Figures 5 and 6 are due to the different normalizations. Before applying SAL, the gust fields were normalized with the local 90th and 98th percentiles (Figure 2). Therefore the normalized values differ between the two fields which affects the amplitude error. Therefore, the differences in the amplitude error for the different percentiles stem from our approach. Another approach which uses absolute wind speeds instead of normalized wind speeds in forecast and analysis would lead to the same amplitude errors for two thresholds, because these thresholds are not considered in the calculation of the amplitude component. Interestingly, however, the ensemble mean forecasts show a tendency towards positive amplitude errors, which may also be explained by the reduced percentiles for higher lead times for the ensemble mean. Considering the structure error (Figure 6, top) it can be noted that there is a clear tendency towards negative structure errors already at short lead times (75 % of individual forecasts are negative at a lead time of one day). This means that already for short lead times, the forecasts have a tendency towards too small gust objects which again increases with growing lead time. Considering the ensemble mean forecast, the smoothing effect can be confirmed by a tendency to higher values for the structure error. Interestingly for short lead times (where the location of gust objects is very close to each other within individual forecasts) the structure error, however, remains negative in most cases. For longer lead times, when the location of gust objects diverge, the smoothing effect is much stronger and finally leads to positive structure errors in most cases. The variation of structure errors given by the box plots show that in the case of a higher object threshold (98th percentile) the results are very different for the individual 82 storm events considered. These tendencies might hence not be reflected when considering individual storm events.

The results described above are deduced by considering a fixed storm set of 82 storms. It might be argued however that results may be sensitive to the identification thresholds. To investigate this sensitivity of results we additionally defined three independent storm sets by modifying the identification criterion. For the first storm set at least 5 % and less than 10 % of gridboxes need to exceed the local 98th percentile (31 storms). The second requires at least 10 % and less than 20 % of the grid boxes to exceed the local 98th percentile (35 storms) and for the third at least 20 % of the grid boxes need to exceed the local 98th percentile (47 storms). We find that the conclusions derived from the previous analysis remain unaltered. However, there are some general differences worth mentioning. In particular, the structure error generally varies much more for the set of small storms and the location error seems to be generally larger for the set of small storms. For both the set of small and large storms it is confirmed that the location and amplitude errors are smaller when considering the ensemble mean wind field compared to individual ensemble members. However this effect is found to be smaller when considering smaller storm events. Also, the tendency towards smaller structure errors for the ensemble mean wind field at larger lead times can be found irrespective of the storm set considered. However, considering small storm events, the large variation in the structure error implies that this is not consistently found for individual storm events.

4 Summary and conclusions

In this study, the SAL method of Wernli et al. (2008), originally developed for the verification of quantitative precipitation forecasts, was adapted to the verification of 10 m daily maximum wind gusts associated with winter storms in Germany. For this we applied the SAL method to relative wind gusts, i.e. normalized gusts with respect to its local 90th and 98th percentiles. This normalization is done to reduce orographic influences and for each forecast lead time separately. The method was applied to statistically downscaled ECMWF EPS forecasts for a set of 82 storm events identified in the period from 2006 to 2010. Ensemble forecasts are evaluated for lead times ranging from one to nine days. All grid points exceeding the local 90th and 98th percentile respectively were identified as a 'gust object'. Hence, the analysis can either be focused on larger gust objects with high gust speeds (90th percentile) or smaller gust objects with severe and potentially damaging wind speeds (98th percentile).

The results first showed a consistent increase of all three error types (structure, amplitude and location) with growing lead time. However, it was shown that the errors are strongly dependent on the specific event considered and do vary considerably. This was shown to be particularly true for the gust objects defined by a higher threshold (98th percentile). As a basic verification it has been investigated whether or not storm events (regardless of their location and intensity) were correctly forecast within the considered area. Results indicate that by means of the ensemble mean forecast, the FAR can be reduced and the hit rate can be increased in comparison to individual forecast members. The spatial verification showed a clear benefit of using the ensemble mean forecast for the amplitude and location errors. That means that the estimation of a storm's location and average intensity is (in most cases) more precise when done on the basis of the ensemble mean forecast instead of using an individual forecast. The structure errors for the ensemble mean were also reduced. However, ensemble averaging leads to smoothing of wind fields, which can lead to the worsening of results in some cases. This was particularly found for the larger extent objects (using

the 90th percentile as a threshold) at short lead times. In this case the ensemble mean forecast performed considerably worse in comparison to individual forecasts. To conclude, this study demonstrated that the object-based verification SAL method, could be well adapted to investigate forecast errors for gusts associated with winter storms in Germany.

However, it should be noted that the results might depend on the choice of the verification domain. The size of the domain should be at least the size of the objects under consideration, this means in the present case about 1000 km, which is the typical length scale of winter storm events in mid-latitudes. On the other hand, the domain size should not be too large, so that multiple storm events are not present at the same time. The SAL method in this case would not make a distinction between the two storm events (other than the CRA method in Ebert and McBride (2000), which includes an object identification and matching). Interpretation of location and structure error would thus be very difficult in this case, since they do not refer to an individual winter storm. Due to the limited availability of forecast data, which was restricted to the domain shown in Figure 1, the sensitivity to the choice of the verification domain could not be tested. Further studies on the basis of regional model forecasts available for a larger domain should certainly include such sensitivity tests.

Additionally, further research should include comparing these results to the results obtained with other object based verification methods, such as MODE (Davis et al., 2006a). It might also be of interest to explicitly assess errors in the timing of storms (i.e. onset and duration), since this is very relevant for issuing appropriate warnings. Timing errors can have impacts on the spatial verification results (Zimmer and Wernli, 2011). Applying the SAL method to one day before and after the actual day would therefore give an estimate of the timing error. Zimmer and Wernli (2011) applied the SAL method some hours before and after an hourly accumulated precipitation forecast to estimate the timing error. They noted that the approach gives a better insight into the quality and deficits of the precipitation forecasts with short accumulation times. In our case, the "accumulation" time was one day, i.e. daily maximum wind gusts. We therefore have indirectly taken a slight timing error (less than one day) into account. Future studies doing the spatial verification of hourly maximum wind gusts can better estimate the timing error using the approach of Zimmer and Wernli (2011).

The results presented in this study are based on a statistical downscaling of wind gusts from the ECMWF ensemble prediction system and may differ to dynamical models. The wind fields considered here can be considered to be physically consistent on the synoptic scale, however spatial details within the fields of local wind gusts, e.g. convection in cold fronts might not be represented, which may be due to the lack of physical consistency in the downscaling method. It may be of particular interest to compare the presented results with results based on dynamical high resolution ensemble forecasts, which generate physically consistent forecasts. Additionally, it would be worthwhile to increase the temporal resolution, because a winter storm has its maximum intensity normally within a few hours.

Furthermore, the present study investigated how to make use of ensemble forecasts in the case of wind gusts associated with winter storms. Two ways of processing the ensemble information were used. Firstly, the individual ensemble members were analysed in terms of the forecast errors of predicted gust objects. Forecast uncertainty can then be expressed in terms of the distribution of error parameters (structure, location and amplitude). Secondly, ensemble averaged gust fields were considered and corresponding error parameters were directly inferred. Of course, in the latter, no information about the ensemble uncertainty could be derived. Results indicated that the information about a storm's location and its overall intensity (amplitude component) could be well derived from the ensemble mean forecasts. This means that the forecaster will find good guidance with these parameters by considering ensemble averaged gust fields, if using the statistically downscaled model. However, the results suggested that the structure might not be well represented in all cases. While for longer lead times it seems to be beneficial to consider the ensemble averaged wind fields, at short lead times this might not be the case. The smoothing of wind fields might in such cases lead to an overestimation of the affected areas with less distinct wind peaks. In particular, this means for short lead times it may be favourable to consider individual ensemble members to extract accurate information on the storm's structure. In order to give good guidance to forecasters on the optimal interpretation of ensemble predictions, the presented findings are particularly relevant for the task of deriving accurate warning information from ensemble prediction systems.

Acknowledgments

This research was carried out in the Hans-Ertel-Centre for Weather Research. This research network of Universities, Research Institutes and the Deutscher Wetterdienst is funded by the BMVI (Federal Ministry of Transport and Digital Infrastructures). We are grateful to the German Weather Service (DWD) and the ECMWF for providing access to the EPS data. We are also thankful for the helpful comments of two reviewers who undoubtedly helped to improve the manuscript.

References

Anderson, J.L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. – J. Climate **9**, 1518–1530.

Broecker, J., L.A. Smith, 2008: From ensemble forecasts to predictive distribution functions. – Tellus A **60**, 663–678.

Buizza, R., M. Milleer, T.N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. – Quart J. Roy. Meteor. Soc. **125**, 2887–2908.

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

Cheung, K., 2001: A review of ensemble forecasting techniques with a focus on tropical cyclone forecasting. – Meteor. Appl. **8**, 315–332.

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

Davis, C., B. Brown, R. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. – Mon. Wea. Rev. **134**, 1785–1795.

Doms, G., M. Baldauf, 2015: A description of the Nonhydrostatic Regional COSMO-Model – Part I: Dynamics and Numerics. – Consortium for Small-Scale Modeling, 158 pp.

Donat, M.G., T. Pardowitz, G.C. Leckebusch, U. Ulbrich, O. Burghoff, 2011: High-resolution refinement of a storm loss model and estimation of return periods of loss-intensive storms over germany. – Nat. Hazards Earth Syst. Sci. **11**, 2821–2833.

Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. – Meteor. Appl. **15**, 179–202.

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

Fox, N.I., A.C. Micheas, Y. Peng, 2016: Applications of Bayesian Procrustes shape analysis to ensemble radar reflectivity nowcast verification. – Atmos. Res. **176-177**, 75–86.

Gilleland, E., 2015: SpatialVx: Spatial Forecast Verification. – R package version 0.2-4.

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

Gofa, F., D. Boucouvala, P. Louka, H.A. Flocas, 2017: Spatial verification approaches as a tool to evaluate the performance of high resolution precipitation forecasts. – Atmos. Res. **208**, 78–87.

Heneka, P., T. Hofherr, 2010: Probabilistic winter storm risk assessment for residential buildings in germany. – Nat. Hazards **56**, 815–831.

Heneka, P., B. Ruck, 2008: A damage model for the assessment of storm damage to buildings. – Eng. Struct. **30**, 3603–3609.

Jolliffe, I.T., D.B. Stephenson, 2012: Forecast Verification: A Practitioner's Guide in Atmospheric Science. – Wiley and Sons.

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

Klawa, M., U. Ulbrich, 2003: A model for the estimation of storm losses and the identification of severe winter storms in germany. – Nat. Hazards Earth Syst. Sci. **3**, 725–732.

Kruschke, T., 2015: Winter wind storms: Identification, verification of decadal predictions and regionalization. – Ph.D. thesis, Institute of Meteorology, Dept. of Earth Sciences, Freie Universität Berlin, Germany.

Leckebusch, G.C., D. Renggli, U. Ulbrich, 2008: Development and application of an objective storm severity measure for the northeast atlantic region. – Meteorol. Z. **17**, 575–587.

Leutbecher, M., T.N. Palmer, 2008: Ensemble forecasting. – J. Comput. Phys. **227**, 3515–3539.

Mittermaier, M., R. North, A. Semple, R. Bullock, 2016: Feature-based diagnostic evaluation of global NWP forecasts. – Mon. Wea. Rev. **144**, 3871–3893.

Molteni, F., R. Buizza, T.N. Palmer, T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. – Quart. J. Roy. Meteor. Soc. **122**, 73–119.

Palmer, T.N., R. Gelaro, J. Barkmeijer, R. Buizza, 1998: Singular Vectors, Metrics, and Adaptive Observations. – J. Atmos. Sci. **55**, 633–653.

Palmer, T., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. – Technical Memorandum **42**.

Pantillon, F., P. Knippertz, U. Corsmeier, 2017: Revisiting the synoptic-scale predictability of severe European winter storms using ECMWF ensemble reforecasts. – Nat. Hazards Earth Syst. Sci. **17**, 1795–1810.

Pardowitz, T., R. Osinski, T. Kruschke, U. Ulbrich, 2016: An analysis of uncertainties and skill in forecasts of winter storm losses. – Nat. Hazards Earth Syst. Sci. **16**, 2391–2402.

Prahl, B.F., D. Rybski, J.P. Kropp, O. Burghoff, H. Held, 2012: Applying stochastic small-scale damage functions to german winter storms. – Geophys. Res. Lett. **39**, L06806.

Schulz, J.P., 2008: Revision of the turbulent gust diagnostic in the COSMO Model. – COSMO Newsletter **8**, 17–22.

Schulz, J.P., E. Heise, 2003: A new scheme for diagnosing near surface convective gusts. – COSMO Newsletter **3**, 221–225.

Schulz, J.P., U. Schättler, 2014: Kurze Beschreibung des Lokal-Modells Europa COSMO-EU (LME) und seiner Datenbanken auf dem Datenserver des DWD. – Deutscher Wetterdienst (Stand: 13.06.2014), 81 pp.

Surcel, M., I. Zawadzki, M.K. Yau, 2014: On the Filtering Properties of Ensemble Averaging for Storm-Scale Precipitation Forecasts. – Mon. Wea. Rev. **142**, 1093–1105.

Weniger, M., P. Friederichs, 2016: Using the SAL Technique for Spatial Verification of Cloud Processes: A Sensitivity Analysis. – J. Appl. Meteor. Climatol. **55**, 2091–2108.

Wernli, H., M. Paulat, M. Hagen, C. Frei, 2008: SAL – A novel quality measure for the verification of quantitative precipitation forecasts. – Mon. Wea. Rev. **136**, 4470–4487.

Zimmer, M., H. Wernli, 2011: Verification of quantitative precipitation forecasts on short time-scales: a fuzzy appraoch to handle timing errors with SAL. – Meteorol. Z. **20**, 95–105.