Adv. Sci. Res., 13, 113–120, 2016
www.adv-sci-res.net/13/113/2016/
doi:10.5194/asr-13-113-2016

Advances in Science & Research
Open Access Proceedings

15th EMS Annual Meeting & 12th European Conference on Applications of Meteorology (ECAM)

Wind gust warning verification

Cristina Primo

Deutscher Wetterdienst, Offenbach am Main, 63067, Germany

Correspondence to: Cristina Primo (cristina.primo@dwd.de)

Received: 29 January 2016 – Revised: 11 May 2016 – Accepted: 28 June 2016 – Published: 18 July 2016

Abstract. Operational meteorological centres around the world increasingly include warnings as one of their regular forecast products. Warnings are issued to warn the public about extreme weather situations that might occur leading to damages and losses. In forecasting these extreme events, meteorological centres help their potential users in preventing the damage or losses they might suffer. However, verifying these warnings requires specific methods. This is due not only to the fact that they happen rarely, but also because a new temporal dimension is added when defining a warning, namely the time window of the forecasted event. This paper analyses the issues that might appear when dealing with warning verification. It also proposes some new verification approaches that can be applied to wind warnings. These new techniques are later applied to a real life example, the verification of wind gust warnings at the German Meteorological Centre ("Deutscher Wetterdienst"). Finally, the results obtained from the latter are discussed.

1 Introduction

Forecasting extreme events helps the public take action to prevent losses or disasters. Therefore, meteorological centers around the world increasingly include the provision of warnings of extreme events among their duties. Different warning systems or extreme event forecast strategies are currently implemented in many weather centers around the world. To improve these warnings systems and satisfy public demands there is a need to develop appropriate warning verification methods. These methods aim to provide information about the performance of a warning system and to compare different versions of it. As a result, there is a high demand for verification techniques for extreme weather events and warnings. This issue was pointed out by the Technical Advisory Committee Subgroup on Verification Measures in two meetings carried out in 2008 and 2009 at ECMWF. This has also been discussed continuously in recent verification meetings of the Joint Working Group on Forecast Verification Research (JWGFVR, www.wmo.int/pages/prog/arep/wwrp/new/Forecast_Verification.html). Some advances have been made in the field of extreme weather verification (Jolliffe and Stephenson, 2011; Ebert et al., 2013; Sansom, 2015). However, substantial progress is still needed in order to have sound and reliable weather warning verifications.

A warning is a forecast issued at a time t_0 (issue time) of an event that will occur from a starting time t_1 to an ending time t_2. The forecasted event will have a particular intensity, and will occur over an area of interest. The time window of occurrence (t_1, t_2) is warning-dependent. The difference between the starting time and the issue time is known as the lead time. This paper is focused on verifying wind warnings. In this case, the warnings are binary forecasts that predict: (a) whether the wind will exceed a pre-defined threshold; (b) when it will happen; and (c) over which region it is expected to occur.

With these considerations in mind, a warning is fully characterized by the intensity, the location, the time window when the severe weather is expected to happen, and the lead time. A warning is useful when the lead time is long enough to allow the user to take adequate actions. Provided that the lead time is long enough, a perfect warning is then a warning that has the correct intensity and is given for the right area during the correct time window. Ideally, a verification study should give information about the performance of these relevant aspects, i.e. lead time, intensity, correct timing and correct area. In verifying the latter aspects, two properties have to be considered: accuracy (did the warning predict the event in the right place, at the right time and with the right intensity?) and timeliness (was the warning given early in advance to allow for taking action and preventing damages or losses?).

Regarding accuracy, there are many scores defined to verify binary events that have been used in warning verification: False Alarm Ratio, Probability of Detection, Critical Success Index, etc. (Schaefer, 1990; Barnes et al., 2007). These measures have been used by many meteorological centers, such us Met Office (Sharpe, 2010), the Austrian National Weather Service (Wittman, 2009) or NOAA in the USA (Brotzge et al., 2013). The German Meteorological Service ("Deutscher Wetterdienst", DWD) uses these scores, among other warning verifications, to verify thunderstorm warnings and to compare different nowcasting systems (Wapler et al., 2012).

In the case of rare events, the rather low occurrence frequency makes the scores tend to zero. The problem of finding a good score for extreme events has been actively studied in the literature during the last decade. The Extreme Dependence Score (EDS, Stephenson et al., 2008) was presented as a new score to verify extreme weather that does not vanish for low base rate events. However, it depends on the base rate and can be increased by over-forecasting (Ghelli and Primo, 2009; Primo and Ghelli, 2009). New scores such as the Symmetric Extreme Dependence Score (SEDS, Hogan et al., 2009) or the Extremal Dependence Index (SEDI) and Symmetric Extremal Dependence Index have been introduced to improve the properties of the score (see Ferro and Stephenson, 2011 for a review). The behavior of these scores has been examined for extreme precipitation events (Nurmi, 2010; North et al., 2013). The results obtained for differentiating the performance of competing forecast systems for extreme events seem to be good, yet these scores have not been widely tested.

Another aspect to take into consideration is that many verification studies do not consider how much in advance the warning was issued. Instead, these studies only consider whether there is a warning in place at the moment when the event happened. However, as Wilson and Giles (2013) pointed out, the warnings have to be given to the public early enough so action can be taken to prevent damages or losses.

In order to account for this in the verification methodology, they introduced a new index for the simultaneous verification of accuracy and timeliness of weather warnings of the Canadian weather warning programme. This index accounts not only for the accuracy of the warnings by using the Extremal Dependency Index (EDI, Ferro and Stephenson, 2011), but also for the relation between the lead time of the warning compared to a maximum allowed lead time. In this index, those lead times which exceed twice the maximum lead time will not be given any credit. However, in order to have a meaningful limit, the EDI is recommended by their founders to be used only in calibrated systems (Ferro and Stephenson, 2011). This is indeed not a desirable property of a warning system since the cost of a missed event usually greatly outweighs the cost of a false alarm for most severe weather situations. Then forecasters would feel thus encouraged to overforecast severe events. Nevertheless, overforecasting strategies are not severely penalized unless excessive, because the EDI penalizes an additional false alarm much less than an additional miss, since adding a single false alarm produces a much smaller increment in the false alarm rate than adding a single miss does in the hit rate.

This paper analyzes how the wind warning verification is carried out at the German Weather Service. Accuracy and timeliness are analyzed separately, showing how the warning system performs for different lead times. The outline of the paper is as follows. Section 2 describes the data used in the study. Section 3 shows how observations and forecasts can be matched. Section 4 presents verification results on an hourly basis and Sect. 5 from an event-based point of view. Finally Sect. 6 presents the summary and conclusions.

2 Data

The German Weather Service ("Deutscher Wetterdienst", DWD) is developing a semi-automatic system to generate warnings, the so-called Automatic Status Generator (ASG; Schröder, 2013). The ASG is part of the AutoWARN project (Reichert, 2009; Reichert et al., 2015). This warning system combines data derived from model output statistics (Hoffmann, 2008) to produce warning proposals that are given to the forecasters. These warning proposals, thereafter referred to as automatic warnings, consist of polygons over Germany. These polygons are placed where an extreme event is expected to happen. These polygons contain all the relevant information about the warnings, including the intensity of the event, the starting time, the ending time and the area affected. Once the forecasters receive these warnings, they can modify them based on all information they currently possess and on their own experience, prior to producing the final warnings to be given to the public. Thus, the warning process is a two-step process: the semi-automatic part, derived from automatic warning proposals, and the final warning given by the forecasters. This paper does not attempt to discuss the generation of warnings made at the DWD, but to present and discuss possible ways of verifying and comparing warnings produced in the two step warning process chain. Our focus is on knowing whether the semi-automatic system is able to produce warnings that are as good as the final warnings given by the forecasters. Additionally, the development of such a verification methodology would also allow for comparing and verifying different versions of the semi-automatic system ASG for various warning criteria. In this paper we focus on wind gust warnings. In particular, six different warning categories are defined for wind, according to a wind gust threshold that has to be reached. These thresholds are (a) $14\,\mathrm{m\,s^{-1}}$ or more; (b) $18\,\mathrm{m\,s^{-1}}$ or more; (c) $25\,\mathrm{m\,s^{-1}}$ or more; (d) $29\,\mathrm{m\,s^{-1}}$ or more; (e) $33\,\mathrm{m\,s^{-1}}$ or more and (f) $39\,\mathrm{m\,s^{-1}}$ or more. A colored label is assigned to these warnings, going from yellow for high winds (wind gust

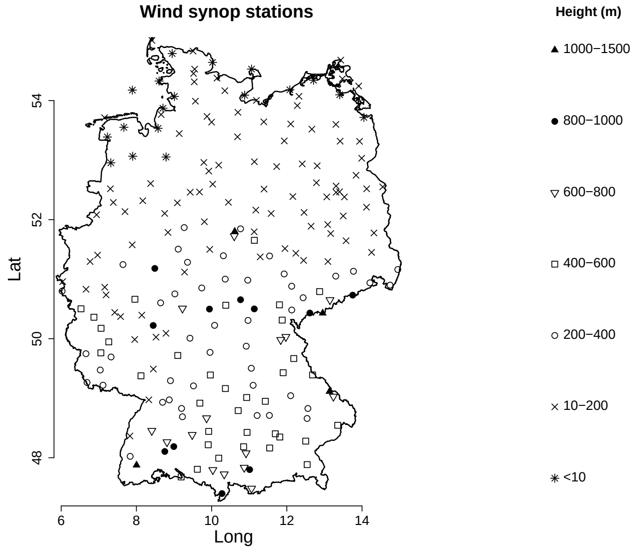

Figure 1. Map of synoptic stations over Germany. The symbols represent the altitude of the station in meters.

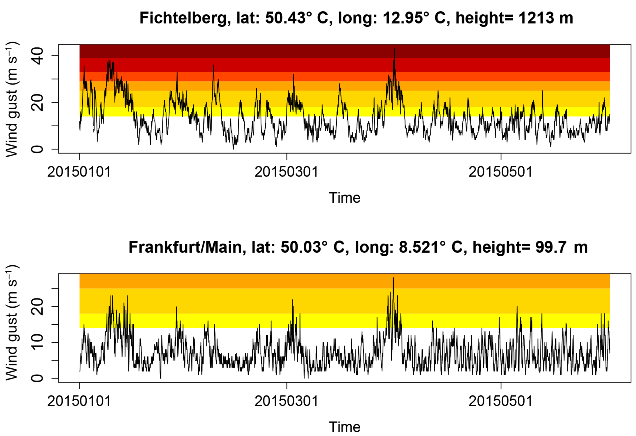

Figure 2. Temporal series from January to May 2015 of the wind gusts registered at two stations of Germany. The color bars represent the different categories of the wind warnings given by the German Meteorological Service ("Deutscher Wetterdienst"), going from high wind warnings (14 m s^{-1} or more, yellow) up to severe warnings (39 m s^{-1} or more, dark red).

above 14 m s^{-1}) up to dark red for severe warnings (above 39 m s^{-1}). The warnings are given a two-digit code in the DWD convention going from 51 up to 56, respectively. Some figures in this paper will refer to this naming convention.

Regarding the observations, the DWD is provided with a network of synoptic stations around Germany that report wind gusts on an hourly basis. Figure 1 represents the spatial distribution of the 226 synoptic stations around Germany used in this study. The symbols represent the altitude of these stations. For each station, we have a temporal series that is coded according to the different warning criteria. Figure 2 represents two temporal series of the observed wind gusts for two stations in Germany: Frankfurt, having an altitude of 99.7 m and where less severe events happened and Fichtelberg, with an altitude of 1213 m and the occurrence of some severe cases. Wind can change rapidly and the temporal series has many jumps up and down. However, the warning system is designed to avoid jumping from a warning at one hour to no warning the next and back to a warning again in the following hour. These warnings are accepted as good warnings when they start at the beginning of the storm and finish at the end, even though at some hours in the middle the intensity was not that high. The verification technique should take this into account, to avoid penalizing warnings which do not forecast correctly the internal jumps within a storm. Therefore, even though it is not advisable to process observations before use in a verification, in this case we can justify smoothing the observations to avoid jumps within a storm and to improve the representativeness of the data. In this study, the two warning systems follow the same criterion of avoiding issuing many warnings within the same storm, but rather try to issue one unique warning representing the maximum intensity. The observations are then smoothed according to the following criterion: every hour each observation is replaced by:

$$\widetilde{o}_{\mathrm{h}} = \min\{\max(o_{\mathrm{h}}, o_{\mathrm{h}-1}, o_{\mathrm{h}-2}), \max(o_{\mathrm{h}}, o_{\mathrm{h}+1}, o_{\mathrm{h}+2})\} \tag{1}$$

Where o_{h} is the observation at an hour h, $o_{\mathrm{h}-x}$ is the observation x hours before and $o_{\mathrm{h}+x}$ is the observation x hours later. This smoothing function does not affect the maximum of the time series but it reduces the magnitude of intensity minima within an event of higher intensity.

Figure 3 represents the coded series according to the warning criteria at the DWD of the hourly observations in the stations represented in Fig. 2.

3 Matching forecast and observations

The first issue found when verifying warnings against synoptic stations is the representativeness of the observations. Warnings are issued over areas, while the observations are at the particular location of the synoptic station. In this study, 226 synoptic stations distributed around Germany are taken into account. However these do not cover all the areas where warnings are issued. There are two alternatives when dealing with such a problem. On the one hand, one could focus on the warnings, checking stations that lie in warned areas and defining a strategy to produce the contingency table (hits, misses, false alarms and correct rejections). For example a hit is defined when one of the synop stations within the warned area exceeds the threshold; or more strictly when all the stations within the warned area exceed the threshold. Here the threshold refers to the minimum wind gust that needs to happen to have a warning (e.g. 14, 18 m s^{-1}, etc.). On the other hand, one could focus on verifying all stations by checking if there has been a warning issued in the area where the station

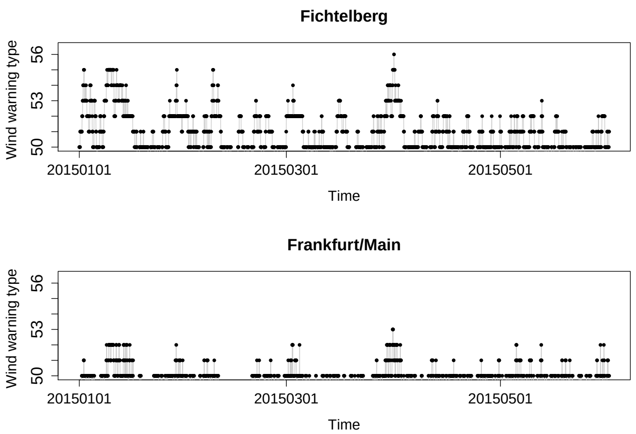

Figure 3. Categorized series of the hourly observed wind gust in Frankfurt and Fichtelberg (Germany) between January and May 2015 according to the warning criteria of the German Meteorological service. Categories from 51 to 56 correspond to winds above 14, 18, 25, 29, 33, 39 $\mathrm{m\,s^{-1}}$ respectively. A new category named 50 was defined to account for the false alarms in which there is a warning but only winds between 8 and 14 $\mathrm{m\,s^{-1}}$ were observed.

lies. In this case, spatial correlations may influence the results if we have strong variations in station density between different areas. Nevertheless, we are more interested in verifying if our system was able to forecast what was observed, rather than if observations concur with what our model warned. In the first case misses are more penalized than false alarms whereas in the second false alarms are more penalized than misses. From the point of view of users, warnings may be given in uninhabited places, whereas places with a high population always have a synoptic station nearby. Therefore, we go for the second approach and run verification for every synoptic station, comparing with what it was warned for that point.

4 Hourly verification

As a first attempt to verify warnings at the DWD, hourly observations were compared to the warnings given during the last hour for that particular point. Two different systems are compared, the warnings proposed by the semi-automatic system alone and the final warnings given by the forecasters. From these forecast-observation pairs the contingency tables are obtained for each warning system and two scores are computed: the hit rate and the false alarm ratio. Figure 4 shows the hourly verification of these two warning systems during the period January to May 2015 for three different lead times: 1, 3 and 6 h ahead. For lead times higher than one, hourly observations are compared to the warnings given during the last sixty minutes before the respective full hour (e.g. lead time 3 refers to warnings given between three and four hours ago). Grey bars represent the semi-automatic system ASG, where the colour indicates the severity of the event (darker colours refer to higher intensities of an event) and white colours represent the warnings given by the forecasters. Vertical lines show confidence limits computed by bootstrap. Hourly observations with intensity 14, 18, 25, 29, 33 and 39 $\mathrm{m\,s^{-1}}$ happened 71 654, 23 641, 3474, 1103, 349 and 68 times respectively, which represents a base rate equal to 10.19, 3.37, 0.50, 0.16, 0.05 and 0.01 %. This verification aims to show if the automatic system is able to perform at least as well as the forecasters. The figure shows that ASG is able to have a similar or better number of hits. For warnings of lower categories one hour ahead, it tends to over forecast, but in general, the number of false alarms remains at least as good as the forecasters. Three hours in advance, the semi-automatic system is able to increase the hits while decreasing the number of false alarms. For six hours in advance, the differences between the systems are even higher for both scores. However, this is probably because forecasters tend not to issue warnings so long in advance.

5 Event-based verification

An hourly verification has some issues. The observed and forecast time windows often differ in time (one can start earlier than the other or vice versa) or in length (one can last longer than the other or vice versa). In cases in which the observed and forecast time windows do not overlap exactly, for example due to a small mismatch in time, we will have a period in which the event is missed, and another period in which the forecast was a false alarm. These failed periods correspond to a unique event, although they are considered independently. Thus, this particular warning will have a double penalty in the verification process; the mismatch will be counted both as a short false warning given when nothing was observed at all and as a short event completely missed. The verification method should account for these different cases, because it is worse for a system to miss warning an event at all rather than just having a small mismatch in time. In this sense, a verification represents the system better if it is event-based. Thus, if a warning is defined over a time window, the verification should be made over those windows, rather than hour by hour. Sharpe (2015) shows an approach to verify warnings at the Met Office in the UK based on an event definition. This objective verification introduces a flexible way of considering small mismatches in time or intensity – Hence it does not use the standard 2×2 contingency table but introduces other categories. For example, in addition to the strict hit, other categories such as early hit, late hit or low hit are considered. This avoids obtaining disappointing results due to customers increasingly requiring warnings to be issued for small areas. However, the use of these new categories renders scores improper because the verification thresholds are allowed to depend on whether or not a warning was issued. If these extra categories are considered then they

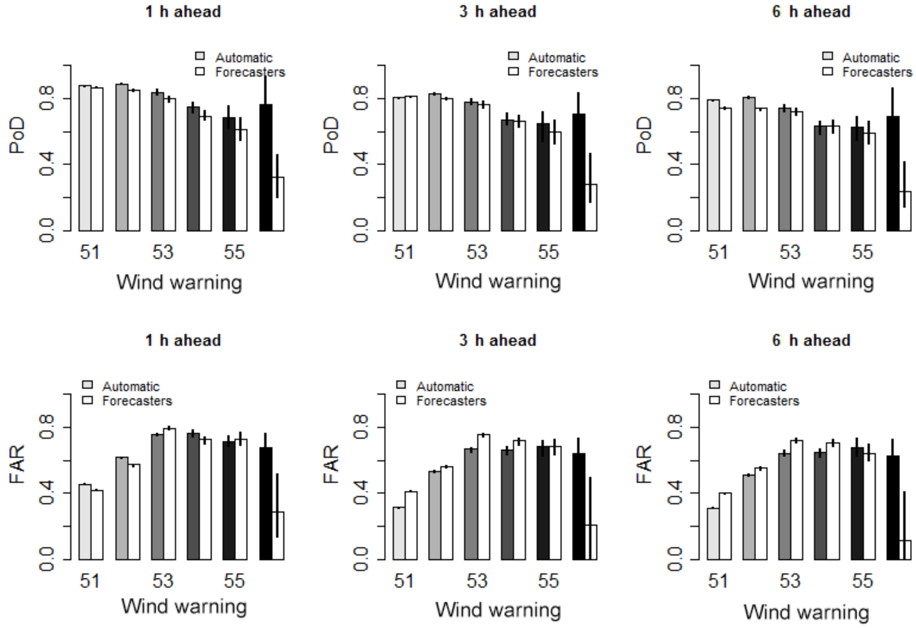

Figure 4. Probability of Detection (first row) and False Alarm Ratio (second row) for three different lead times: 1 h (first column), 3 h (second column) and 6 h (third column), for two different warning systems: the semi-automatic system (ASG, gray bars) and the warnings given by the forecasters (white bars). Every plot represents six different wind warning categories, going from category 51 up to 56 that correspond to wind gusts above 14, 18, 25, 29, 33 and 39 $m\,s^{-1}$ respectively. The period considered covers the first five months of 2015. The verification of the warnings is made on an hourly basis. Confidence intervals obtained by bootstrap are also included

must also be considered for cases where a warning was not issued, which might increase the number of missed events.

There is also another issue: how to define the pairs observed event-warning. A criterion is needed to define the contingency table, for example, to know what is considered as a hit. When we have observed events and warnings whose time length differ, we have different possibilities: it may happen that the time window of a warning covers more than one observed event, or vice versa, the time window of an observed event covers more than one warning. Therefore, the relation between forecast warnings and observed events is not bijective (one to-one correspondence, a forecast does not imply one and only one observed event). There could be only one observed event and one warning, but the time windows do not match. One possible option in this case is to define a threshold that corresponds to the minimum percentage of the observed event that has to be warned to be considered as a hit. Those cases in which there is a warning during the observed event, but the duration of the warning does not reach the percentage threshold must be considered a miss.

In addition, as pointed out in the hourly verification, a conditioned verification can also be done. For example, one could choose to verify the warnings given when an event was observed, or on the contrary, to check what is observed when the system gives a warning. The difference between these two points of view lies in the importance we give to the missed events or to the false alarms. If we focus on the observations, we will not know what happens when nothing was observed and thus the false alarms are not penalized. In contrast, if we focus on the warnings, we will not know what happens between two warnings and thus we will not give importance to the misses. In an operational weather centre, misses are more penalized than false alarms. Therefore, we will focus on the observed events and check what the warning system warns during those events. In any case, both perspectives miss part of the contingency table. This is because a warning system produces warnings, but does not produce "non-warnings" when an observed event was missed.

We have decided to verify observed events. However, ignoring the false alarms encourages hedging (Jolliffe, 2008), and the verification results could be easily improved just by increasing the number of warnings, because the false alarms are not penalized. Hedging is a non-desirable property and it should be penalized by the verification method. Thus, it is recommended to consider also the false alarms. In this study, those warnings issued during an observed event, but not covered by the hits or misses because they are of a higher category, are considered as false alarms. A new extra category, wind above 8 $m\,s^{-1}$, was created just to account for those cases in which there was a warning (i.e. forecast of a wind above 14 $m\,s^{-1}$) but the observed wind was not above 14 $m\,s^{-1}$ but only above 8 $m\,s^{-1}$. Bearing in mind that warning systems tend to overforecast to avoid missing an event, these false alarms provide a way to penalize overforecasting.

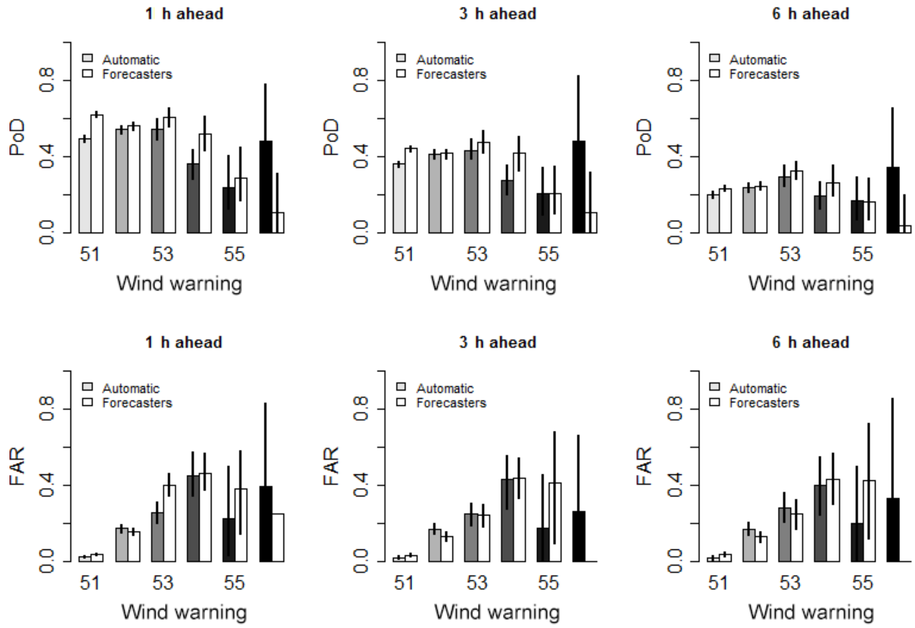

Figure 5. Probability of Detection (first row) and False Alarm Ratio (second row) for two different warning systems: the semi-automatic system (ASG, gray bars) and the warnings given by the forecasters (white bars). Three different lead times are considered: 1, 3 and 6 h ahead. Every plot represents six different wind categories, going from category 51 up to 56 that correspond to wind gusts above 14, 18, 25, 29, 33 and 39 $\mathrm{m\,s^{-1}}$ respectively. Vertical lines represent confidence intervals obtained by bootstrap. The period considered covers the first five months of 2015. The verification of the warnings is made based on events.

However, false alarms may be underestimated since warnings cannot be matched with non-observed events since the non-observed event duration is unclear.

Example

In our case, one of the questions we want to answer is if the warnings were able to forecast the maximum intensity of the observed events, distinguishing the six categories defined in the previous section (from warnings for 14 $\mathrm{m\,s^{-1}}$ or more up to 39 $\mathrm{m\,s^{-1}}$ or more). Thus, as a first attempt at event-based verification, we choose the following criterion: we allow a hit whenever the observed event's intensity was warned for at least one hour during the event duration. The observed events are computed by looking for maximum wind speeds observed in our period (from January to May 2015). To obtain these maxima, a moving average of windows of seven hours is computed, to avoid localized maxima of very short duration. Then, the maxima wind speed of the resulting series, define the observed events. The duration is defined according to the number of hours in which this maximum wind occurred. For those hours, we check if there was a warning and what category it was. Following this we fill the contingency table. Both systems, automatic and forecasters, will lead to contingency tables of the same total size.

Figure 5 shows the results obtained from this event-based verification for three different lead times: 1, 3 and 6 h ahead. Vertical lines represent confidence intervals obtained by bootstrap. For winds of 14 $\mathrm{m\,s^{-1}}$ or more (element 51), forecasters are still better, but in other cases, such as winds above 25 $\mathrm{m\,s^{-1}}$ (element 53), the higher hit rate is due to overforecasting since a higher false alarm ratio is seen. The number of observed events with intensity equal to 14, 18, 25, 29, 33 and 39 $\mathrm{m\,s^{-1}}$ happened 9086, 5585, 3558, 588, 212, 88 and 29 times respectively. In this verification all the events are considered equally relevant in the verification. However, event duration has an impact on the performance of the system. Thus, Fig. 6 shows the results after repeating the verification by including only those events that last more than three hours to avoid these short convective situations. In this case, Fig. 6 shows an improvement of ASG compared to its performance forecasting short events, since it has higher probability of detection. Something has to be improved for ASG warnings of events lasting one or two hours. This issue has already been corrected in a later ASG version and work is continuously in progress to improve the semi-automatic system. In any case, this study shows that the automatic system at the DWD achieves a performance comparable to forecasters for wind gust warnings.

6 Discussion and future work

Warnings have become a standard product in meteorological centres since they help the public prevent major disasters and minimize costs or losses. Therefore, verification methods

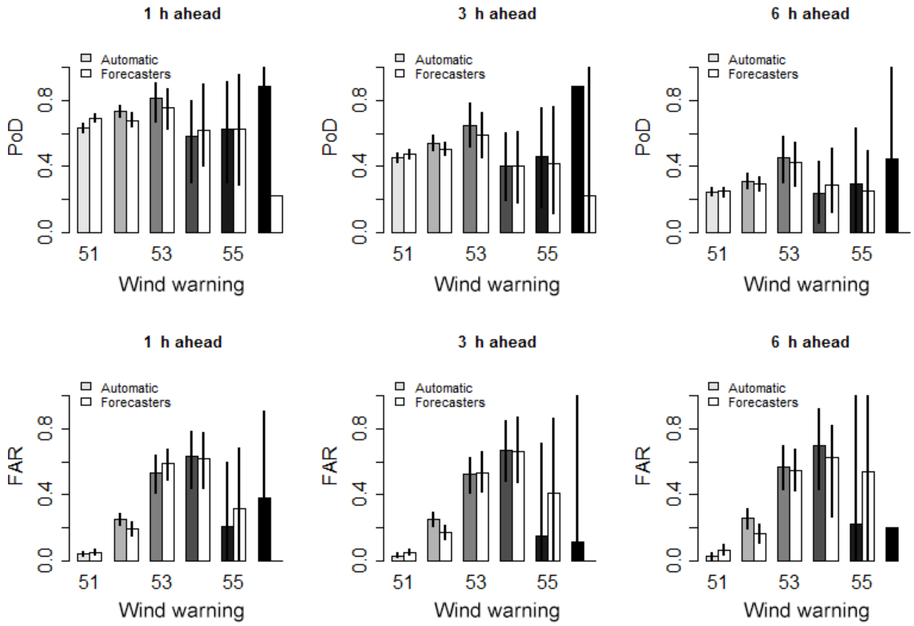

Figure 6. Similar to Fig. 5, but only events with a duration longer than three hours are taken into account.

need to adapt to the fact that warnings forecast rare events and they are given for a time window rather than for a particular time unit. Thus, a verification strategy has to be defined to match observations with forecasts and to clarify how to treat those warnings that do not overlap exactly an observed event, but are misplaced in time.

This study describes the issues relating to wind warning verification and reviews the current state of warning verification methods. Some verification approaches implemented at the DWD are presented to compare warnings coming from a semi-automatic warning system and the final warnings proposed by the forecasters. Results show that the semi-automatic system performs similarly to the forecasters, even though some issues need to be solved for very short observed events. Work is already in progress and new versions of the warning system (Automatic Status Generator, ASG) have been developed to solve this problem. In addition, research is ongoing to propose new verification techniques that solve the limitations of the current ones and better describe the quality of the warning system.

Spatial issues may also need to be considered for new studies. For example, a spatial tolerance can be allowed to match observations with warnings that are within a certain radius. This would help deal with the issue that observations are at point locations while warnings cover areas. Stratification by regions could also help assess whether altitude impacts on verification results.

7 Data availability

Data used in this study are not publicly available, but they are archived in the German Meteorological Service archive. Please, contact "Deutscher Wetterdienst" upon availability.

Acknowledgements. The author would like to thank Guido Schröder for providing the warning data and fruitful discussions about warnings and the semi-automatic system at DWD. This work was supported by the AutoWARN project at the Deutscher Wetterdienst.

Edited by: P. Nurmi
Reviewed by: two anonymous referees

References

Barnes, L. R., Gruntfest, E. C., Hayden, M. H., Schultz, D. M., and Benight, C.: False alarms and close calls: a conceptual model of warnings accuracy, Weather Forecast., 22, 1140–1147, 2007.

Brotzge, J. A., Nelson, S. E., Thompson, R. L., and Smith, B. T.: Tornado Probability of Detection and Lead Time as a Function of Convective Mode and Environmental Parameters, Weather Forecast., 28, 1261–1276, 2013.

Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Nurmi, P., Gill, P., Göber, M., Joslyn, S., Brown, B., Fowler, T., and Watkins, A.: Progress and challenges in forecast verification, Meteorol. Appl., 20, 130—139, doi:10.1002/met.1392, 2013.

Ferro, C. A. T. and Stephenson, D. B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, Weather Forecast., 26, 699–713, doi:10.1175/WAF-D-10-05030.1, 2011.

Ghelli, A. and Primo, C.: On the use of the extreme dependency score to investigate the performance of an NWP model for rare events, Meteorol. Appl., 16, 537–544, 2009.

Hoffmann, J. M.: Entwicklung und Anwendung von statistischen Vorhersage-Interpretationsverfahren für Gewitternowcasting und Unwetterwarnungen unter Einbeziehung von Fernerkundungsdaten, Promotionsarbeit im Fach Meteorologie, University of Berlin, 2008.

Hogan, R., O'Connor, E. J., and Illingworth, A. J.: Verification of cloud-fraction forecasts, Q. J. Roy. Meteor. Soc., 135, 1494–1511, 2009.

Jolliffe, I. T.: The impenetrable hedge: a note on propriety, equitability and consistency, Meteorol. Appl., 15, 25–29, 2008.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practioner's Guide in Atmospheric Sciences, Wiley, 292 pp., 2011.

North, R., Trueman, M., Mittermaier, M., and Rodwell, M.: An assessment of the SEEPS and SEDI metrics for the verification of 6 h forecast precipitation accumulations, Meteorol. Appl., 20, 164–175, 2013.

Nurmi, P.: Experimentation with new verification measures for categorized QPFs in the verification of high impact precipitation events – an ECMWF initiative, Proceedings, 3rd WMO International Conference on Quantitative Precipitation Estimation and Quantitative Precipitation Forecasting and Hydrology, 18–22 October 2010, Nanjing, China, 222–226, 2010.

Primo, C. and Ghelli, A.: The effect of the base rate on the extreme dependency score, Meteorol. Appl., 16, 533–535, 2009.

Reichert, B. K.: AutoWARN – Automatische Unterstützung der Herausgabe von Unwetterwarnungen, Promet, 35, 98–103, 2009.

Reichert, B. K., Glashoff, J., Hess, R., Hirsch, T., James, P., Lenhart, C., Paller, J., Primo, C., Raatz, W., Schleinzer, T., and Schröder, G.: The Decision Support System AutoWARN for the Weather Warning Service at DWD, EMS Annual Meeting Abstracts, 12, EMS2015-221-2, 15th EMS & 12th ECAM Conference, 7–11 September 2015, Sofia, Bulgaria, 2015.

Sansom, P.: Advances in forecast verification, Weather, 70, 14-14, doi:10.1002/wea.2445, 2015.

Schaefer, J. T.: The critical success index as an indicator of warning skill, Weather Forecast., 5, 570–575, 1990.

Schröder, G.: Entwurf Fachkonzept. ASG – Erzeugung von Warnvorschlägen für AutoWARN 2.0, Deutscher Wetterdienst, Internal Publication, Info available at: http://www.dwd.de/DE/forschung/wettervorhersage/met_fachverfahren/unterstuetzung_warnprozess/autowarn/autowarn_node.html,ASG (last access: 15 July 2016), 2013.

Sharpe, M.: Verification of weather warnings, UK Met Office Internal Report, Met Office: Exeter, UK, 9 pp., 2010.

Sharpe, M. A.: A flexible approach to the objective verification of warnings, Meteorol. Appl., 23, 65–75, doi:10.1002/met.1530, 2015.

Stephenson, D. B., Casati, B., Ferro, C. A. T., and Wilson, C. A.: The Extreme Dependency Score: a nonvanishing measure for forecasts of rare events, Meteorol. Appl., 15, 41–50, 2008.

Wapler, K., Goeber, M., and Trepte, S.: Comparative verification of different nowcasting systems to support optimisation of thunderstorm warnings, Adv. Sci. Res., 8, 121–127, doi:10.5194/asr-8-121-2012, 2012.

Wilson, L. J. and Giles, A.: A new index for the verification of accuracy and timeliness of weather warnings, Meteorol. Appl., 20, 206–216, doi:10.1002/met.1404, 2013.

Wittmann, C.: Evaluation of Severe Weather Warnings at the Austrian National Weather Service, Working Group for the Cooperation between European Forecasters, Newsletter, Vol. 14, 12–15, available at: http://www.euroforecaster.org/newsletter14/wittmann.pdf (last access: 15 July 2016), 2009.