MAUSAM, **66**, 3 (July 2015), 625-634

551.509.5

Human estimates of warning uncertainty: Numerical and verbal descriptions

TOBIAS PARDOWITZ, THOMAS KOX, MARTIN GÖBER and ALEXANDER BÜTOW*

Freie Universität Berlin, Institut für Meteorologie, Berlin, Germany and

Hans-Ertel-Centre for Weather Research, Optimal Application of Weather Forecasts Branch

**Hans-Ertel-Centre for Weather Research, Optimal Application of Weather Forecasts Branch and*

Deutscher Wetterdienst, Offenbach, Germany

e mail : tobias.pardowitz@met.fu-berlin.de

सार – मौसम चेतावनियों की अनिश्चितता अधिकांशतः केवल पठनीय रूप (उदाहरणतः कल अपराह्न में गर्ज के साथ तूफान आने की संभावना है) में दी जाती है। अतः भाषा में दी जाने वाली चेतावनियों की अनिश्चितता को संख्यात्मक अनिश्चितता के साथ जोड़ा जाना चाहिए। दो प्रश्न उठते हैं : क्या मानव पूर्वानुमानकर्ता अनिश्चितता का आकलन कर सकते हैं और इसे मौखिक रूप से कितना अच्छी तरह किया जाता है।

बर्लिन शहर के संबंध में प्रचंड मौसम की घटनाओं के होने की संभाव्यता के विषयपरक और सांख्यिकीय पूर्वानुमानों की जाँच की गई। बर्लिन में गर्ज के साथ तूफान और पवन झोकों >14 m/s की घटना के लिए संभाव्यता के मानव द्वारा किए गए आकलन सांख्यिकीय पूर्वानुमान की तुलना में विश्वसनीय पाए गए और उनमें उल्लेखनीय कौशल दिखा। इसके अतिरिक्त एक प्रचालनात्मक चेतावनी रिपोर्ट में चेतावनी की अनिश्चितता का मौखिक वर्णन वर्गीकृत किया गया और उसकी वस्तुनिष्ठ जाँच की गई। परिणाम दर्शाते हैं कि पूर्वानुमानकर्ता वास्तविक रूप में इससे जुड़ी अनिश्चितता के बारे में जानते है, परंतु उसे मौखिक रूप से व्यक्त करते हैं। संचार व्यवस्था में सुधार लाने के उद्देश्य से और प्रचंड मौसम सूचना से जुड़ी भाषिक अनिश्चितता से उठने वाले भ्रमों को कम करने के लिए पूर्वानुमान में कम और सुपरिभाषित मौखिक वाक्य होने चाहिए जो पूर्वानुमान की अनिश्चितता बता सकें। अनिश्चितता के संख्यात्मक से मौखिक वर्णन से शब्दों के मौलिक रूप से भिन्न प्रयोग का पता चला जब गर्ज के साथ तूफान और पवन झोंको >14m/s की चेतावनियों की तुलना गर्ज के साथ तूफानों के मामले में प्रयोग किए गए शब्द 'स्ट्रॉंगर' के साथ की गई। इससे पता चलता है कि उपयोगकर्ताओं को संभाव्यता की सूचना की जगह जोखिम की सूचना दी जानी चाहिए।

ABSTRACT. The uncertainty of weather warnings is mostly expressed only in textual form (*e.g.*, "thunderstorms are possible tomorrow afternoon"). Thus linguistic uncertainty might be added to the numerical uncertainty of the warnings. Two questions arise: can human forecasters estimate the uncertainty and how well is this done in verbal terms.

Subjective and statistical forecasts of the probability of the occurrence of severe weather events for the city of Berlin were verified. Human estimates of the probability for the occurrence of thunderstorms and wind gusts > 14 m/s in Berlin were found to be reliable and possess significant skill in comparison to the statistical reference forecast. Additionally, the verbal description of warning uncertainty in an operational textual warning report was classified and objectively verified. Results indicate that forecasters actually are aware of the inherent uncertainty, yet express this by means of a multitude of verbal terms. In order to improve the communication and reduce confusions arising from linguistic uncertainty inherent to severe weather information, forecasts should thus contain few and well defined verbal phrases expressing forecast uncertainty. Relating numerical to verbal descriptions of uncertainty revealed a fundamentally different usage of wording when comparing warnings of thunderstorms and wind gusts >14 m/s, with "stronger" wording used in case of thunderstorms. This might indicate that risk information rather than probability information is communicated to the users of the considered warning information.

Key words – Weather warnings, Subjective probability estimates, Reliability, Linguistic uncertainty, Warning verification.

1. Introduction

As with all weather forecasts, weather warnings are uncertain (Wilson and Giles, 2013). There are various sources of the uncertainty of weather forecasts, *e.g.*, in physical understanding, observations, models and their code or limited communication capacity (National Research Council, 2006). In addition, weather warnings are intentionally biased, because the need to make a binary decision to cope with a hazard leads to over forecasting to be "on the safe side".

How is the uncertainty of the warnings experienced by and visible to the user? There are two ways: firstly, there is the daily experience, *i.e.*, people who regularly receive warnings notice that not all warnings are correct. Weissmann *et al.* (2014) show that people are aware of this uncertainty, but have a tendency to underestimate it.

Secondly, warnings are issued via various media with an inclusion of some kind of uncertainty. In Germany, the official weather warning system of the German Weather Service (DWD) is organised in a three step process, each step involving very different designations of uncertainty. Early warning information is given by a 7-day forecast of potential weather hazards ("Wochenvorhersage Wettergefahren"). It includes information about expected severe large-scale weather events with qualitative statements about forecast uncertainty using verbal statements only. The terms possible ("möglich"), likely ("wahrscheinlich") and very likely ("sehr wahrscheinlich") are used by default. None of these terms are related to an explicit numerical value. Secondly, an alert or watch (the regional warning report - "Regionaler Warnlagebericht") is issued up to 12 hr before an expected event. These forecasts are provided at least four times a day 48 with different reports for the whole country and twelve regions respectively. The regions represent the larger German states or a combination of smaller ones. The regional warning report contains a plethora of unspecified uncertainty terms and will be the main focus of this study. Ultimately, warnings are issued on county level ["(Un) Wetterwarnung"]. On occasion, terms describing spatial or temporal constraints (*e.g.*, "locally") denote some uncertainty.

At the moment, all of the above mentioned uncertainty information is given verbally. Yet this verbal information is a further source of uncertainty, often called "linguistic uncertainty" (Regan *et al*., 2002) or "linguistic imprecision" (Morgan and Henrion, 1990).

The linguistic uncertainty in weather warnings are best shown by comparing numerical and verbal descriptions of uncertainty. In a questionnaire survey with participants from the emergency service community (fire fighters, relief forces, policemen and civil authorities) in Germany, Kox *et al.* (2014) asked participants to assign numerical values to the verbal statements used in DWD's 7-days forecast: "Imagine the national weather service states the advent of an upcoming storm in your region with the indications 'possible'/'likely'/'very likely'. Which of the following probabilities would you associate to this forecast?" The values 0% to 100% were predefined in 10% steps for each statement. Results (Fig. 1) show that attributions scatter widely and cover almost the whole range of numerical probabilities. The other most dominate feature is the overlap between the attributions. With exception of the very low probabilities (0% to 10%), every term overlap every other term (Kox *et al*., 2014). Rogell (1972) and Sink (1995) came to similar results in their studies with laypeople, while a study by Murphy and Brown (1983a) showed an overlap in interpretations of different terms of cloudiness.

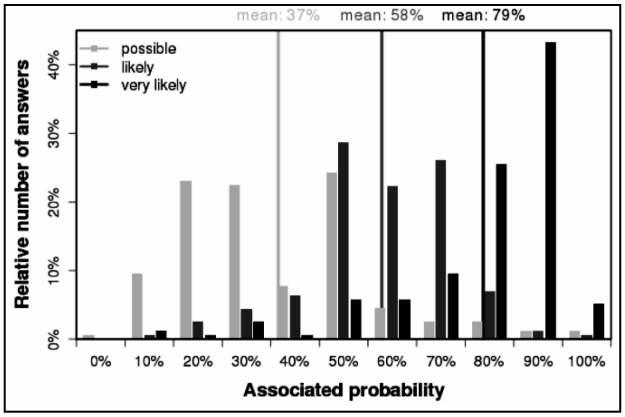


Fig. 1. Numeric associations to verbal expressions of uncertainty used in DWD's 7-days forecast stated by members of German emergency services. Mean values: 36.5% (possible), 57.8% (likely), 78.7% (very likely), n = 157 (Figure based on Kox *et al*., 2014)

Regan *et al.* (2002) discriminate linguistic uncertainty into:

- Under specificity,
- Vagueness,
- Ambiguity,
- Context dependence.

The example above shows very well the different subtypes of linguistic uncertainty: Firstly, the scattering of the three terms shows that they all are highly under specific and thus a subject to major variability in interpretation. Under specificity occurs when there is an unwanted generality and the terms do not provide the desired degree of specificity (Regan *et al*., 2002).

Secondly, the terms are vague, expressed by the overlap of the three terms. Vagueness arises because our language, including much of the scientific vocabulary, permits borderline cases (Regan *et al*., 2002). As shown, the terms possible, likely, and very likely can mean something totally different from one person to another.

TABLE 1

Characteristics of the three types of forecasts verified

	Regional warning report	Numerical probability estimate	Warn MOS
Type	Textual	Numerical	Numerical
Issued by	Human	Human	Machine
Issued when	10 local time	Just after regional warning report	10 local time (summer 11)
Domain	Berlin & Brandenburg as a whole	Berlin	Berlin
Temporal resolution	Words like morning, evening, in the night, …	6 hourly up to +30 hr	6 hourly
Target audience	Public	DWD Internal Test Product	DWD internal operational guidance
Verification period	06/2010-05/2014	02/2013-05/2014	02/2013-05/2014

Other examples are the descriptions of weather events: The term "rain" is vague because some people might read heavy showers, while others read prolonged rain. Regional differences might also occur as "rain" might mean something totally different to someone living in Berlin than someone living in Cherrapunjee.

Ambiguity, uncertainty arising from the fact that a word can have more than one meaning and it is unclear which meaning is intended (Regan *et al*., 2002), is often confused with vagueness and seen as similar by some authors. However, as Bueno and Colyvan (2012) point out the term vagueness should be used for borderline cases only.

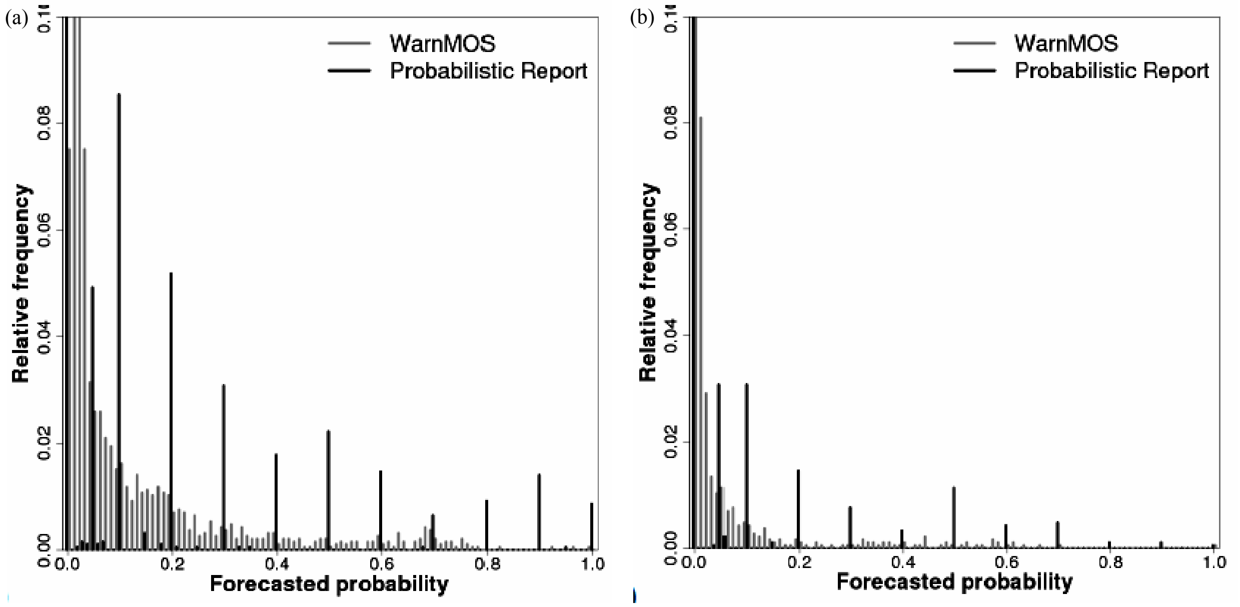
Some terms might be vague and context dependent as well. For example the term "There is a 60% probability of precipitation tomorrow" is hard to grasp without a context or reference class. People might not understand the correct interpretation that there will be rain in a specified place and time on 60% of days like tomorrow. They might instead misinterpret the term to mean the percentage of time it will rain tomorrow or the percentage of the affected area on which it will rain (Gigerenzer *et al*., 2005). Such misinterpretations result from the confusion between probability and variability. Providing the reference class "days like tomorrow" can help understanding (Spiegelhalter *et al*., 2011). But once the context is specified, a term's vagueness described above might still remain.

Furthermore, uncertainty can arise on both the senders' and the receivers' side. It is almost a cliché that science uses numbers to express uncertainty, and lay people use verbal expressions (Handmer and Proudley, 2007). But while the weather forecast is generated on the basis of numerical prediction systems, the communication of warnings is mostly verbal. While the example above showed the linguistic struggles end-users of a weather warning are facing, Erev and Cohen (1990) point to a new challenge when switching the perspective: While most people prefer to receive uncertainty information numerically, they prefer to express the same information verbally. This paradox seems even more bizarre since both the receivers and conveyors of information know that the information is uncertain and want the most efficient communication.

Another challenge forecasters are facing is that the more distinct probability levels they want to express, the harder it is to do it with verbal terms. On the one hand, it is hard to use verbal terms without losing precision (Erev and Cohen, 1990). On the other hand, using verbal terms could avoid the problem of having to reach consensus on a particular estimate or range (Patt and Schrag, 2003).

Uncertainty in warnings could not only be expressed in verbal terms, but also as numerical estimates. Yet are human forecasters able to make reliable numerical estimates of uncertainty? Here weather forecasters are the most frequently used example for experts, who can provide calibrated forecasts (Murphy and Winkler, 1977), one reason being the opportunity to give regular feedback through verification and thus to obtain calibration over time (Griffin and Brenner, 2008). This has been shown for frequent events, *e.g.*, or the probability of (any) precipitation (PoP), but little is known for rare and high impact events. Here we tested the ability of forecasters to reliably estimate numerically the uncertainty of gust and thunderstorm warning information and investigated the relationship between numerical and verbal appraisals of uncertainty.

The following sections (*i*) describe the three forecasts data sets evaluated in this study; (*ii*) specify the analysis of verbal terms and the verification methodology; (*iii*) describe the results; (*iv*) discuss the results, 120 draw conclusions and make recommendations.



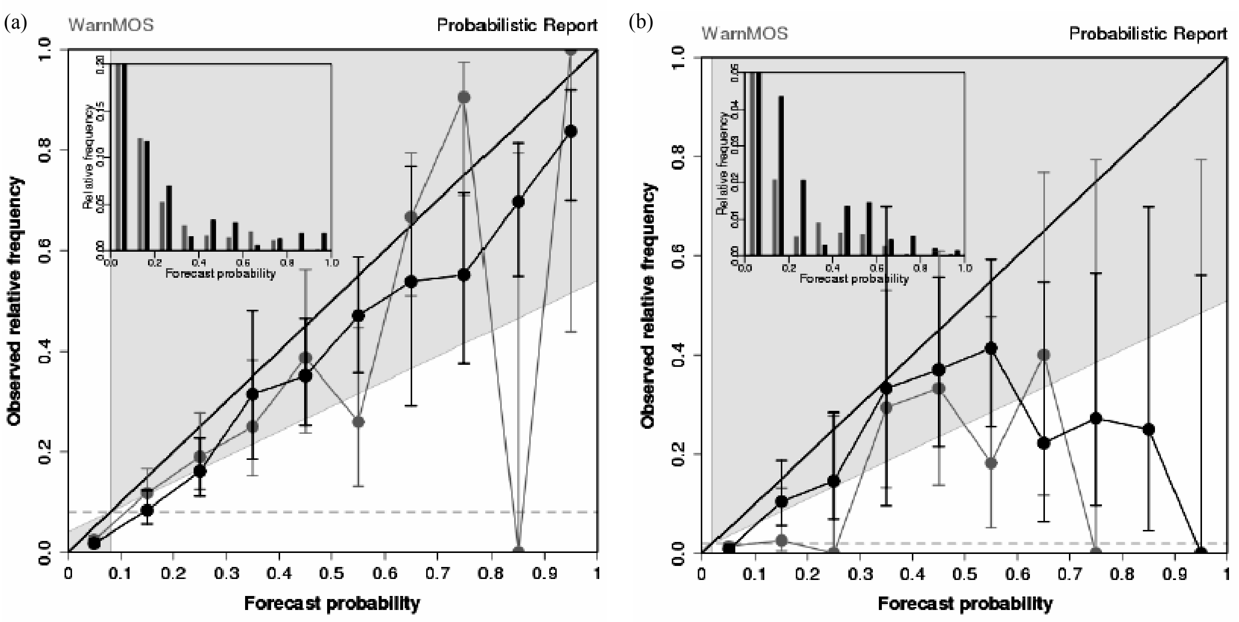
Figs. 2(a&b). Distribution of forecasted numerical probability estimates for wind gusts > 14 m/s (a) and thunderstorms and (b) Black bars denote human forecasts (Probabilistic Report) and machine estimates (WarnMOS) are shown in grey. Total number of forecasts is N = 1924

2. Data

Three different data sets have been evaluated in this study (Table 1). The operational regional warning report, corresponding to the second stage within the warning strategy of the German weather service was evaluated. It contains a verbal description of severe weather events likely to occur within the region of Berlin and Brandenburg (30.000 km^2, with Berlin right in the middle of Brandenburg, which allows no differentiation of the two areas for short range forecasts) within the upcoming 24 hours. Regional warning reports are being issued at least 4 times a day in 6 hourly intervals with additional reports being issued if necessary. In this study, the reports issued at around 1000 local time were analysed with a validity period lasting till midday of the subsequent day. Furthermore, a new test forecast product was analysed, which is issued by DWD's regional office in Potsdam for the city of Berlin (900 km^2). This test product is issued by the same forecaster just after the issue of the operational regional warning report mentioned above. It contains numerical estimates of the probability of occurrence of certain severe weather events including snowfall, thunderstorms, wind gusts (with different thresholds) as well as prolonged and heavy rainfall. Probabilities are specified for 5 time intervals covering the following 30 hours, namely 1200-1800 UTC and 1800-2400 UTC of the current date and 0000-0600 UTC, 0600-1200 UTC and 1200-1800 UTC of the consecutive day. For comparison, numerical probability forecasts from a statistical forecast system (WarnMOS, Hoffmann, 2008) initialised at 0900 UTC (1000 local time, in summer 1100) comparable to the human forecasts were considered. Additionally, observational data were used for verification purposes. Hourly wind gust measurements were used, available for three measurement stations within Berlin and 15 stations distributed over Brandenburg. Maximum gusts were calculated for the 6 hour periods corresponding to the validity periods of the forecasts to identify wind gusts exceeding the threshold of 14 m/s at least once within any of the periods. Lightning observations from a detection network (Betz *et al*., 2009) were evaluated with respect to whether at least one lightning with an intensity above 8000 Ampere had been observed within one district in the Berlin-Brandenburg area within the 6 hour forecast periods.

3. Methods

Within the regional warning reports, the verbal descriptions of uncertainty associated with the occurrence of thunderstorms and wind gusts > 14 m/s were classified with respect to the validity period as well as the phrases used to express the uncertainty. Regarding the reference time, a set of terms exist for which an “intended time” is defined in local time, which can be translated into the 6 hour UTC time segments. With these specifications, the terms “In the morning” (“Am Morgen”, 0000-0600 UTC), “forenoon” (“Vormittag”, 0600-1200 UTC), “noon” (“Mittag”, 0600-1200 UTC), “afternoon” (“Nachmittag”, 1200-1800 UTC), “evening” (“Abend”, 1200-1800 UTC) and “in the night” (1800-0600 UTC) are translated into the corresponding 6 hr periods as specified in brackets (note that 0600 UTC is 8 o'clock summer time in Germany).



Figs. 3(a&b). Reliability diagrams for probability forecasts of wind gusts > 14m/s (a) and thunderstorms and (b) Given a forecast in a specified probability range, with a binning width of 10%, the observed relative event frequencies are shown on the y-axis. The sharpness diagram (inset) shows the binned distribution of forecast probabilities. Colouring is as in Fig. 2, with machine estimates in grey and human estimates in black. Shaded area highlights the area of positive Brier score against climatology. Error bars are given according to formula 7.67 of Wilks (2006) p327 with an error level α = 0.05

Less commonly used descriptions of the time reference, which are not strictly defined and thus might be interpreted individually by the forecasters were identified like "during the day" ("Im Tagesverlauf", 0600-1800 UTC), "second half of the night" ("Zweite Nachthälfte", 0000-0600 UTC) and "second half of the day" ("Zweite Tageshälfte", 0600-1800 UTC) for which an intended reference time has been subjectively defined which were translated into the corresponding UTC time segments specified in brackets. It should be noted however, that the analysis was done with the original German wording and further uncertainty may arise due to the translation. Verbal descriptions furthermore include spatial restrictions (*e.g.*, "localised" or "in northern parts") as well as temporal restrictions (such as "temporarily", "occasionally" or "frequently") often in combination with purely probabilistic expressions. The most frequent probabilistic expressions identified were: "can not be excluded", "are possible", "can occur", "have to be accounted for" and "are expected". Also, numerous expressions implying certainty about the occurrence of an event were used, summarized in the following by the expression "will occur". Additionally, reports containing no reference to one of the event categories were identified, implying that an event was excluded. Disregarding the spatial and temporal restrictions, the probabilistic terms were identified within the textual description and linked to the corresponding event type (thunderstorm or wind gusts) and the corresponding validity period to produce a data set to be objectively verified.

4. Results

4.1. *Verification of numerical probability estimates for the Berlin area*

The distribution of numerical probabilities differs markedly comparing human to machine estimates [Figs. 2(a&b)]. While human forecasters restrict probabilities to multiples of 10%, with exceptions mainly for low probabilities, machine estimates feature a continuum of probabilities. Furthermore, for both variables, wind gusts > 14 m/s [Fig. 2(a)] as well as thunderstorms [Fig. 2(b)], distributions of human forecast probabilities feature a local maximum at 50% with considerably lower populations for 40% and 60% (known as the "error of central tendency", describing the affinity for survey respondents to choose the centre of a scale). In addition, considerable differences were found for high probabilities. While machine estimates rarely reach high probabilities (particularly in the case of thunderstorms), human forecasts show much higher populations here. For wind gusts > 14 m/s, WarnMOS forecasted probabilities larger than (or equal to) 80% in only 0.2% of all cases, while humans forecasted these probabilities in 3.2% of the cases. Similarly, for thunderstorms only 0.05% of the predictions above or equal to 80% were made by the machine, compared to 0.27% of human forecasts.

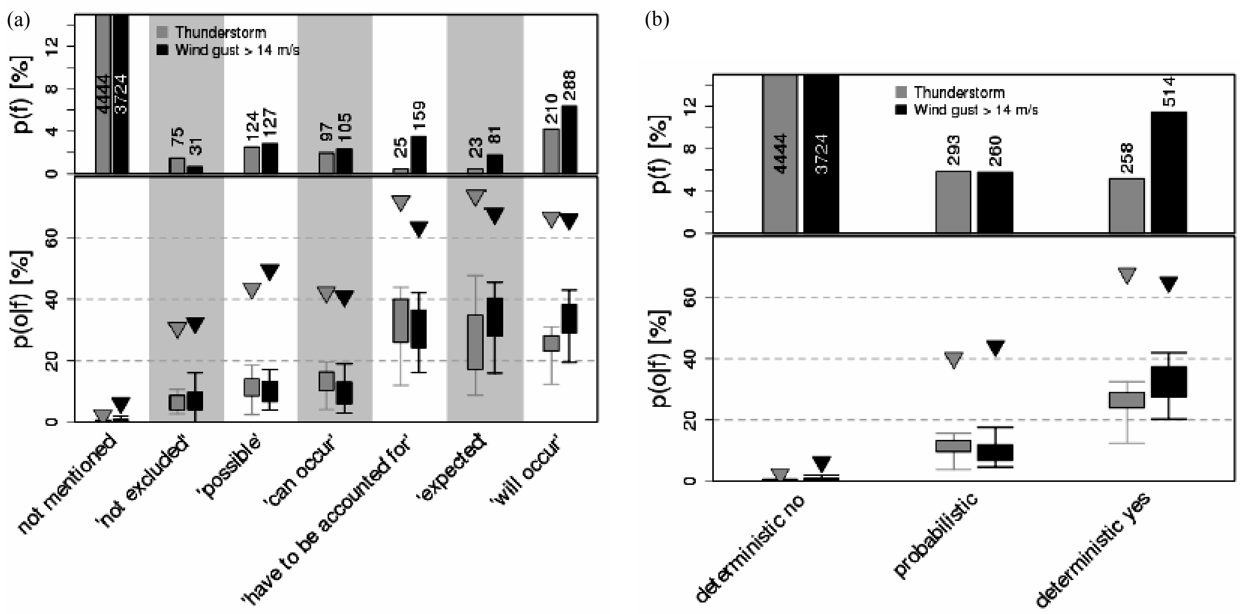
One key issue of probabilistic forecasting is whether the forecasts correctly represent the uncertainties associated with the forecasted event, which can only be assessed by considering a (large) set of forecasts. Given a forecast being in a specified probability range, the reliability diagram [Figs. 3(a&b)] shows the observed relative frequency of the forecasted event, assessed by counting the number of events observed in these cases and normalized with the number of forecasts (Jolliffe and Stephenson, 2011). In the case of forecasts for wind gusts > 14 m/s [Fig. 3(a)], observed relative frequencies correspond well to the forecasted probabilities for both human (black) as well as machine estimates (grey). Compared to the human forecasts the reliability diagram in case of the machine estimate is less monotonic, which is due to the fact mentioned before that high probabilities were rather rarely forecasted. *e.g*., in only one case a probability between 80% and 90% has been made by WarnMOS, a case in which no event had been observed at any of the considered measurement stations. Thus human probability estimates were found to be well calibrated, with observed frequencies being slightly lower than forecasted. This can partly be attributed to the fact that only three observation stations were available for the city of Berlin, while gusts above 14 m/s might have happened somewhere else in the forecast area. In the case of probabilistic thunderstorm forecasts by humans, good calibration is found for forecast probabilities below 50%, with worse calibration for machine forecasts. Since high probabilities are rare in the case of thunderstorm predictions, particularly in the case of machine forecasts, the small sample sizes however do not allow for a robust assessment of the calibration properties. In general, psychological research suggests that humans are often overconfident for forecasts of rare events (Griffin and Brenner, 2008).

The Brier Score $BS = \frac{1}{N}\sum_i (f_i - o_i)^2$ represents the mean squared probability deviation, with f_i being the i^{th} out of N forecasted probabilities and o_i being the corresponding observation which is either 1 if an event has been observed or 0 if not. The BS can be decomposed (Murphy, 1973) into a reliability term BS_{rel} describing the calibration (as discussed in terms of the reliability diagram), a resolution term BS_{res} expressing the ability of a forecast to discriminate between high and low event probabilities as well as an uncertainty term BS_{unc}, which is only dependent on the observed occurrence rate and is thus independent of the forecast performance. In case of machine estimates of thunderstorm probability BS_{rel}, BS_{res} and BS_{unc} were found to be 0.0035, 0.0017 and 0.0196 respectively, while for human estimates both BS_{rel} = 0.0045 and BS_{res} = 0.0039 are found to be higher, implying an improvement in resolution and a (somehow smaller) worsening in reliability. For wind gust forecasts, BS_{rel}, BS_{res} and BS_{unc} were found to be 0.0028, 0.0203 and 0.0741 respectively, while for human forecasts BS_{rel} decreases (improves) to 0.0015 and BS_{res} increases (improves) to 0.0281. Thus in both thunderstorm as well as wind gust forecasts, the increased resolution of the human probability forecasts leads to a reduction (improvement) of the BS, since this was achieved without degrading the forecast in terms of reliability. In terms of the Brier Skill Score (BSS = 1-BS/BS_{ref}), taking the machine estimate (WarnMOS) as a reference forecast (BS_{ref}), a reduction in mean squared probability deviations of 5.7% (BSS = 0.057) is found for the human thunderstorm forecasts and 16% (BSS = 0.161) in case of forecasts of wind gusts > 14 m/s.

4.2. *Verification of verbal probability expressions*

Similar to the assessment of reliability and resolution properties in the case of numerical probability estimates, the distribution of verbal probability expressions as well as the conditional relative frequencies of observed events - given a certain expression has been used in the forecast - were evaluated [Figs. 4(a&b)]. As shown in the upper panel of Figs. 4(a&b), most text reports do not contain any information on thunderstorm or wind gusts, which accounts for about 4000 six hour periods, *i.e*., about 85% of the sample. The probabilistic expressions identified in the texts were found to be used differently when comparing thunderstorm and wind gust forecasts. Most strikingly, these differences were found for the expression "have to be accounted for", which has been used in reference to wind gusts in 3.5% of the forecasts, whereas only 0.5% of thunderstorm forecasts contained this expression. While the terms "possible" and "can occur" were used with similar frequencies for gusts and thunderstorms, expressions implying rather high certainty (subjective rating of the original German wording) like "have to be accounted for" ("es muss mit ... gerechnet werden"), "are expected" ("zu erwarten") and deterministic expressions such as "will occur" ("werden auftreten") were used more frequently in reference to wind gusts.

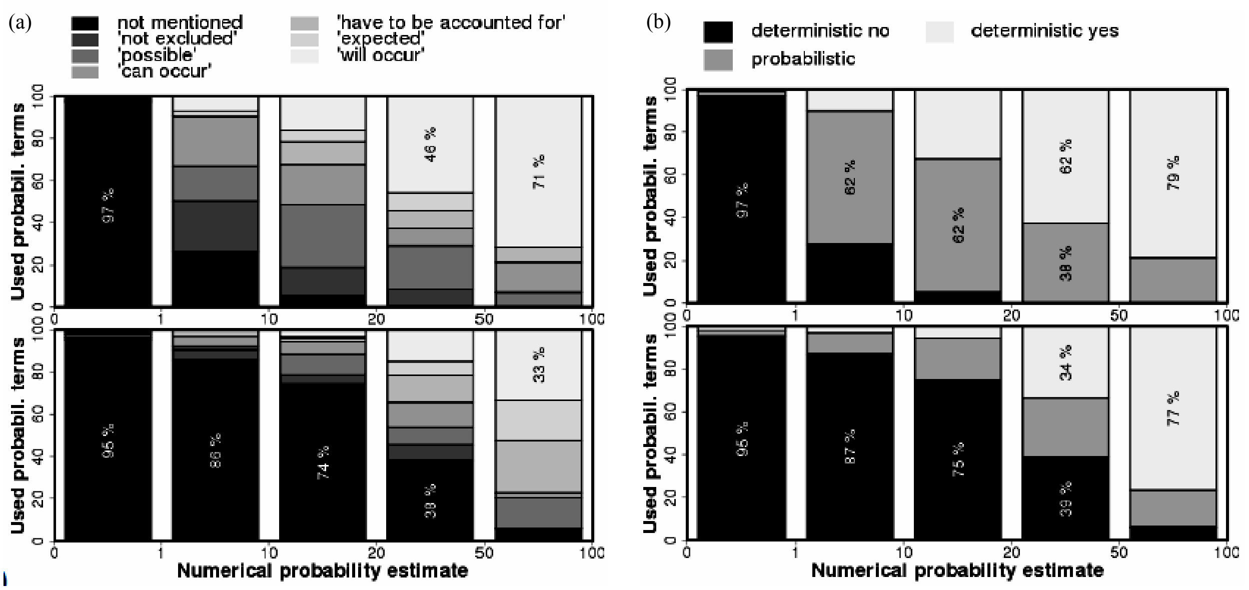
Given a forecast containing a certain uncertainty expression, observed event frequencies were assessed in a further step. Since the reference area of the forecast - different to the numeric probability forecasts for Berlin - is rather large, covering Brandenburg and Berlin, two different interpretations of the forecasts were taken. On the one hand the occurrence of an event can be realized if the corresponding threshold has been exceeded



Figs. 4(a&b). Forecasted event frequencies p(f) and conditional observed event frequencies p(o|f) given that a warning contained a specific uncertainty term (a) or an uncertainty term of the groups (b) grouping the expressions into probabilistic expressions ("not excluded", "possible", "can occur"), deterministic yes ("have to be accounted for", "expected", "will occur") and deterministic no (no mentioning within the report). Relative event frequencies p(o|f) are determined as threshold exceedances anywhere within the area of investigation (shown as triangles) or alternatively on the basis of individual stations (shown as box-plots). Period investigated was 06/2010-05/2014

at least at one of the measurement stations within the forecast area [shown by triangles in Figs. 4(a&b)]. In a different perspective (*e.g.*, a single person being at one time at one location only), each individual station can be evaluated to assess corresponding event probabilities. There is no definition whether only one of the two interpretations is intended by the forecast producers. There is also only anecdotal evidence about the interpretation by forecast users which suggests a local interpretation at least for thunderstorms. A mixture between the two interpretations is also frequently described by using terms to specify local restrictions like "... occur locally", "widespread", "in exposed areas".

For the set of 18 stations measuring wind gusts within Berlin and Brandenburg, the local interpretation led to a distribution of event probabilities which are depicted by the black Box-Whiskers in the lower panel of Fig. 4(a). Here, boxes represent the range between the 25% and 75% quantiles and whiskers show the minimum and maximum. Similarly for thunderstorm forecasts, distributions of the derived event probabilities for the 19 individual districts are shown as grey Box-Whiskers. As a result it can be found that if the forecast reads that wind gusts > 14 m/s "will occur" they actually occurred with a probability of about 65% anywhere in Brandenburg/ Berlin. At individual stations, an event was measured with a 20% to 40% chance, depending on the station, with half of the stations featuring an event in 30% to 38% of the cases. Furthermore, observed event frequencies were found to be only marginally different when comparing the expressions "have to be accounted for" and "are expected" with respect to both interpretations of the forecast. Likewise for the expressions "can not be excluded", "possible" and "can occur" similar event probabilities are derived. Interestingly, the diagnosed event probabilities agree very well comparing thunderstorm and wind gust predictions, with the largest deviations for forecasts at individual stations using deterministic expressions like "will occur". In this case, event probabilities at individual stations were found to be considerably lower for thunderstorms. On average an occurrence frequency of 25% was diagnosed compared to an average of 35% in case of wind gusts > 14m/s. However, probabilities for an event anywhere in the forecast area were found to agree well with a 65% chance in both cases. The agreement found for different event types might indicate that forecasters actually have a distinct appraisal of the inherent uncertainty. Yet they expressed this by means of multiple verbal expressions, revealed by a grouping of expressions with respect to observed event frequencies into categories of high event probabilities ("will occur", "are expected" and "have to be accounted for") and lower probabilities ("can not be excluded", "are possible" and "can occur"). Grouping together these expressions into those lower probability expressions (probabilistic) and



Figs. 5(a&b). Distribution of used uncertainty expressions (a) and groups of expressions (b) given a certain range of numerical probability estimate issued within the human forecasts. Top panels show the distributions in case of thunderstorm forecasts and bottom panels for wind gust forecasts, respectively. Period investigated was 02/2013-05/2014

high probability expressions (deterministic yes), Fig. 4(b) shows observed relative frequencies based on single stations to be about 10% in the probabilistic group and about 30% in the deterministic one. Considering events anywhere within the forecast area, occurrence frequencies were found to be close to 40% and 65% in the probabilistic and deterministic groups.

4.3. *Relation between numerical probability estimates and usage of verbal uncertainty expressions*

The numerical probability estimates as issued in the probabilistic report for Berlin and the verbal description of forecast uncertainty in the warning report for Berlin and Brandenburg are related by assessing the distribution of expressions used, given a forecasted probability within a specified range [Fig. 5(a), top panel for thunderstorm and bottom panel for wind gust >14m/s]. As expected, if the numerical probability is 0, in most cases no mention of either thunderstorm or wind gusts was found in the textual report. With increasing probability, the fraction of cases in which the events were not mentioned is gradually decreasing, while an increased use of probabilistic expression was found. For high probability estimates (>50%) the fraction of cases in which "will occur" was used increased up to 71% for thunderstorm predictions, and to 33% wind gusts >14m/s. Complementary, numerical probability estimates have been evaluated in dependence of the verbal phrase used (not shown). Even though a dependency of the mean probability can be diagnosed, results indicate broad distributions of the numerical estimates, irrespective of the verbal phrase used. Note that such finding does relate well to findings presented in Fig. 1, showing a similarly broad distribution of the users' association of numerical probabilities to verbal phrases. The most striking differences when comparing the distribution of expressions used for thunderstorms and for wind gusts, were found for intermediate probabilities. In about 75% of cases in which wind gusts were forecasted with a probability between 10% and 20%, they were not mentioned in the textual report. The remaining 25% were spread amongst the various probabilistic expressions, with few cases in which deterministic expressions were used. This is fundamentally different for thunderstorm forecasts, with only 5% of the cases not being mentioned in the report when the numerical estimate was between 10-20%. The remaining 95% again distribute well between probabilistic and deterministic expressions, with deterministic expressions used in more than 10% of these cases. In summary, a fundamentally different usage of expressions can be diagnosed. In case of intermediate probabilities the possibility for a thunderstorm is more often mentioned in the textual report compared to wind gusts above 14 m/s. Consistently, for intermediate and high probabilities, explicit deterministic expressions were used more frequently than for wind gusts to describe the chance for the occurrence of thunderstorm events.

5. Conclusions

Verification results for a new test forecast product introduced at a regional centre of the German weather service - numerical probability estimates for the occurrence of severe weather events in Berlin - have been presented. Human estimates of numerical forecast uncertainty were largely reliable, with forecast skill comparable to machine estimates. Furthermore it has been found that this skill is due to an increased resolution, implying a greater ability of the human forecaster to discriminate between situations of low and high event probabilities.

In addition, the usage of textual expressions of forecast uncertainty in an operational warning report has been objectively verified and in addition has been compared to the human numerical probability estimates. Firstly, we found that differing textual forecast expressions to be rather similar in terms of the frequency of observed events. To eliminate this overlap and vagueness of terms in weather warnings, it would be necessary to limit the forecaster's vocabulary to a small set of distinct words, with the cost of limiting the amount of information (Murphy and Brown 1983b). Secondly, a comparison of the operational usage of textual uncertainty terms and the human numerical probability estimates revealed that in cases with similarly forecasted numerical probabilities a large variety of verbal expressions were used. This is an example of the under specificity inherent in the warning communication. This kind of linguistic uncertainty can be treated by specifying the relationship between words and numbers and by sharply delineating categories (Murphy and Brown, 1983b; Carey and Burgman, 2008). Prominent examples of this are provided in the Weather Service Operational Manual of the US Public Weather Service (NOAA NWS, 1984), where numerical estimates of the Probability of Precipitation (PoP) are associated with distinctive verbal terms, or the quantification of verbal confidence descriptions in the IPCC process (Moss and Schneider, 2000). More recently the Australian Bureau of Meteorology (BOM, 2014) has changed the way to describe the chance of rain only with the terms "slight, medium, high or very high" and a percentage equivalent in parentheses.

However, since “no single representation suits all members of an audience” (Spiegelhalter *et al.*, 2011) and since people prefer numerical information for its accuracy, but prefer to use verbal statements to communicate uncertainty information to others, it is recommended to present both numerical and verbal uncertainty information in a warning. It would make sure that the receiver has the right information regardless of their requirements and needs (Kox *et al.*, 2014; Visschers *et al.*, 2009). Given that many people feel they understand the meaning of words better than numbers (Wallsten *et al.*, 1986) providing both numbers and words might be helpful, especially for forecasts of one-time events which may not have been witnessed before. When comparing the usage of verbal expressions of forecast uncertainty for thunderstorm and wind gust predictions, distinctive differences have been identified, with a clear tendency of using “stronger” (*i.e.*, with implied higher determinism) expressions in the case of thunderstorm forecasts given the same numerical forecast probability as in wind gust forecasts. This may result from the fact that forecasters do not only include purely probabilistic weather information into their description, but also have a focus on possible consequences of the event and thus the risk associated with this information. Assuming higher impacts of thunderstorm events compared to wind gusts above 14 m/s, the higher risk might be expressed by the “stronger” wording used. Such an implicit expression of risk is not uncommon, as the literature on risk perception shows (Patt and Schrag, 2003; Weber and Hilton, 1990) that people both interpret and use probabilities as information about the potential impact of an event as well. “People are more likely to choose more certain sounding probability descriptors (*e.g.*, likely instead of unlikely) to discuss more serious consequence events” (Patt and Schrag, 2003).

An example of the explicit use of risk information in weather warning is the UK Met Office, which issues warnings on the basis of a risk matrix combining the likelihood of an event with its potential impact (Neal *et al.*, 2014). It will be subject of further interdisciplinary research to introduce probabilistic and risk based warning information to the emergency services in Germany and thus investigate their perception and usage of this additional information.

Acknowledgements

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 thank the forecasters of DWD's regional office in Potsdam for issuing the probabilistic test forecast. The authors gratefully acknowledge helpful discussions with Sebastian Trepte from the DWD and Stefan Herzog from the Max-Planck Institute for Human Development, as well as Thomas Endrulat and Linda Jäckel from DWD's regional office in Potsdam. The authors thankfully acknowledge the helpful comments of Elizabeth Ebert and another anonymous reviewer.

References

Betz, H. D., Schmidt, K., Laroche, P., Blanchet, P., Oettinger, W., Defer, E., Dziewit, Z. and Konarski, J., 2009, “LINET - An international lightning detection network in Europe”, Atmospheric Research, **91**, 564-573.

BOM (Bureau of Meteorology), 2014, http://media.bom.gov.au/social/blog/440/clearing-up-the-patchy-rain-introducing-a-more-precise-forecast-language/(retrieved December 17th, 2014).

Bueno, O. and Colyvan, M., 2012, "Just What is Vagueness?", *Ratio*, **25**, 1, 19-33.

Carey, J. M. and Burgman, M. A., 2008, "Linguistic Uncertainty in Qualitative Risk Analysis and How to Minimize It", *Annals of the New York Academy of Sciences*, **1128**, 13-17.

Erev, I. and Cohen, B. L., 1990, "Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox", *Organizational Behavior and Human Decision Processes*, **45,** 1, 1-18.

Gigerenzer, G., Hertwig, R., van der Broek, E., Fasolo, B. and Katsikopoulos, K. V., 2005, "A 30% Chance of Rain Tomorrow': How Does the Public Understand Probabilistic Weather Forecasts?", *Risk Analysis*, **25**, 3, 623-629.

Griffin, D. and Brenner, L., 2008, "Perspectives on Probability Judgment Calibration", in "Blackwell Handbook of Judgment and Decision Making", Koehler, D.J. and Harvey, N. (eds.), Blackwell Publishing Ltd, Malden, MA, USA.

Handmer, J. and Proudley, B., 2007, "Communicating uncertainty via probabilities: The case of weather forecasts", *Environmental Hazards*, **7**, 2, 79-87.

Hoffmann, J., 2008, "Entwicklung und Anwendung von statistischen Vorhersage-Interpretationsverfahren für Gewitternowcasting und Unwetterwarnungen unter Einbeziehung von Fernerkundungsdaten", Ph.D thesis, Freie Universität Berlin, p205. http://www.diss.fu-berlin.de/diss/receive/FUDISS_ thesis _000000004192.

Jolliffe, I. T. and Stephenson, D. B., 2011, "Forecast Verification: A Practitioner's Guide in Atmospheric Science", John Wiley & Sons, 2nd ed., p292.

Kox, T., Gerhold, L. and Ulbrich, U., 2014, "Perception and use of uncertainty in severe weather warnings by emergency services in Germany", *Atmospheric Research*, 158-159 & 292-301, http://dx.doi.org/10.1016/j.atmosres.2014.02.024.

Moss, R. H. and Schneider, S. H., 2000, "Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting", 33-51, In "Guidance papers on the cross cutting issues of the Third Assessment Report of the IPCC", Pachauri, R.; Taniguchi, T. and Tanaka, K., (eds.) Geneva: World Meteorological Organisation.

Morgan, M. G. and Henrion, M., 1990, "Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis", Cambridge Univ. Press, Cambridge, p346.

Murphy, A. H., 1973, "A new vector partition of the probability score", *Journal of Applied Meteorology and Climatology*, **12**, 595-600.

Murphy, A. H. and Brown, B. G., 1983a, "Interpretation of some terms and phrases in public weather forecasts", *Bulletin of the American Meteorological Society*, **64**, 1283-1289.

Murphy, A. H. and Brown, B. G., 1983b, "Forecast terminology: Composition and interpretation of public weather forecasts", *Bulletin of the American Meteorological Society*, **64**, 13-22.

Murphy, A. H. and Winkler, R. L., 1977, "Reliability of Subjective Probability Forecasts of Precipitation and Temperature", *Applied Statistics*, **26**, 41-46.

Neal, R. A., Boyle, P., Grahame, N., Mylne, K. R. and Sharpe, M., 2014, "Ensemble based first guess support towards a risk-based severe weather warning service", *Meteorological Applications*, **21**, 3 563-577.

National Research Council, 2006, "Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts", National Academies Press, Washington, D.C., p112.

NOAA NWS (National Oceanic and Atmospheric Administration, National Weather Service), 1984, "Operational Manual, Chapter 11: Zone and Local Forecasts, Section 8: Probability of Precipitation Forecasts (POP)", http://www.nws.noaa.gov/wsom/manual/archives/NC118411.HTML (retrieved August 10th, 2014).

Patt, A. G. and Schrag, D. P., 2003, "Using specific language to describe risk and probability", *Climatic Change*, **61**, 1, 17-30.

Regan, H.M., Colyvan, M., Burgman, M.A., 2002, "A Taxonomy and Treatment of Uncertainty for Ecology and Conservation Biology", Ecological Applications, **12**, 2, 618-628.

Rogell, R. H., 1972, "Weather terminology and the general public", *Weatherwise*, **25**, 126-132.

Sink, S. A., 1995, "Determining the public's understanding of precipitation forecasts: Results of a survey", *National Weather Digest*, **19**, 3, 9-15.

Spiegelhalter, D., Pearson, M. and Short, I., 2011, "Visualizing Uncertainty About the Future", *Science*, **333**, 6048, 1393-1400.

Visschers, V. H. M., Meertens, R. M., Passchier, W. W. F. and De Vries, N. N. K., 2009, "Probability Information in Risk Communication: A Review of the Research Literature", *Risk Analysis*, **29**, 2, 267-287.

Wallsten, T. S., Fillenbaum, S. and Cox, J. A., 1986, "Base Rate Effects on the Interpretation of Probability and Frequency Expressions", *Journal of Memory and Language*, **25**, 5, 571-587.

Weber, E. U. and Hilton, D. J., 1990, "Contextual effects in the interpretations of probability words: Perceived base rate and severity of events", *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 781-789.

Weissmann, M., Göber, M., Hohenegger, C., Janjic, T., Keller, J., Ohlwein, C., Seifert, A., Trömel, S., Ulbrich, T., Wapler, K., Bollmeyer, C. and Deneke, H., 2014, "The Hans-Ertel Centre for Weather Research – Research objectives and highlights from its first three years", *Meteorologische Zeitschrift*, in press.

Wilks, D. S., 2006, "Statistical Methods in Atmospheric Sciences", Academic Press, 3rd ed., p704.

Wilson, L. J. and Giles, A., 2013, "A new index for the verification of accuracy and timeliness of weather warnings", *Meteorological Applications*, **20**, 206-216.