How Likely is That Chance of Thunderstorms? A Study of How National Weather Service Forecast Offices Use Words of Estimative Probability and What They Mean to the Public

EMILY D. LENHARDT
University of Oklahoma Center for Risk and Crisis Management, Norman, OK

RACHAEL N. CROSS
University of Oklahoma Center for Risk and Crisis Management, Norman, OK

MAKENZIE J. KROCAK
University of Oklahoma Center for Risk and Crisis Management, Norman, OK

JOSEPH T. RIPBERGER
University of Oklahoma Center for Risk and Crisis Management, Norman, OK

SEAN R. ERNST
University of Oklahoma Center for Risk and Crisis Management, Norman, OK

CAROL L. SILVA
University of Oklahoma Center for Risk and Crisis Management, Norman, OK

HANK C. JENKINS-SMITH
University of Oklahoma Center for Risk and Crisis Management, Norman, OK

(Manuscript received Day Month Year; review completed Day Month Year)

ABSTRACT

One of the most challenging aspects of weather forecasting is effectively communicating forecast information to the public. No forecast is ever completely certain as no meteorological phenomenon is guaranteed to occur. As such, the uncertainty in forecast information should be communicated in a way that makes sense to end users. Previous studies of the communication of probabilistic information suggest that, while the general public are more apt to communicate uncertainty with Words of Estimative Probability (WEPs), they prefer to receive that information numerically. Other work has suggested that a combination of numbers and WEPs is the best method for communicating probability, but little has been done to assess the communication and interpretation of probabilistic, meteorological information. In this study, we code 8900 tweets from the National Weather Service (NWS) Weather Forecast Offices (WFOs) and analyze them to find how forecasters communicate probabilistic forecast information to the public via Twitter. This analysis reveals that WFO messaging is dominated by the use of WEPs, with few numerical descriptions of probability. These WEPs are generally vague, unqualified notions of probability that may impede the public's ability to interpret the information that forecasters are trying to communicate. Based on this analysis, two publically fielded surveys are also analyzed in order to understand how participants tend to interpret the most common qualified and unqualified WEPs that WFOs used on Twitter. Findings suggest that people tend to interpret qualified WEPs more concisely than the unqualified words. However, both categories experience a wide range of interpretations, implying that WEPs communicate relatively vague notions of probability that mean very different things to different people.

Corresponding author address: Emily D. Lenhardt, 5 Partners Place, 201 Stephenson Parkway, Suite 2300, Norman, OK 73019
E-mail: emily.lenhardt@ou.edu

1. Introduction

Following the recommendations by the National Research Council (NRC 2006), the National Institute of Standards and Technology (NIST 2013), and academic experts in weather risk communication (e.g., AMS Council 2008; Morss et. al 2008; Joslyn and Savelli 2010), the National Weather Service (NWS) is exploring and expanding the use of probabilistic information to convey weather forecast uncertainty. For example, the National Severe Storms Laboratory (NSSL) is exploring the development of a new forecast and warning paradigm that supplements current watch, warning, and advisory products with geographically specific probabilistic information that constantly evolves throughout the life of a threat (Rothfusz et al. 2018). As the NWS moves in this direction, questions about how to best communicate uncertainty are becoming increasingly urgent. This study contributes to this area of research by (1) systematically documenting the words and phrases that NWS forecast offices currently use to communicate probabilistic information in severe thunderstorm and tornado forecasts; and (2) examining how members of the public interpret these words and phrases. We begin by briefly summarizing previous research. We then introduce the data and methods we use for the analysis and highlight the study results. We conclude by discussing the implications of the study and suggesting directions for future research.

2. Summary of Previous Research

Words of Estimative Probability (WEPs) are a common method of communicating probabilistic information to the general public. An example of the use of a WEP is “severe thunderstorms have a low chance of occurring this evening,” where “low chance” is used to communicate the likelihood of severe thunderstorm occurrence. Unfortunately, these words can be vague in nature and their meaning can be difficult for users to interpret. Despite this confusion, a majority of studies show that people prefer to express probabilistic information using words and receive such information numerically (Fischer and

Jungermann 1996, Willems et al. 2019, MacLeod and Pietravalle 2017, Wintle et al. 2019, Morss et al. 2008). Despite past work that indicates people inherently understand words better than numbers (Wallsten et al. 1986a), recent findings suggest that numbers communicate probabilities more effectively (Willems et al. 2019, Wintle et al. 2019) and potentially carry less ambiguity (Friedman and Zeckhauser 2014, Fischer and Jungermann 1996). In fact, as early as the 1980s, researchers have argued that expressing probability using words is a poor way to convey confidence (Beyth-Marom 1982), and therefore forecasting organizations should use numbers instead.

Although numbers are potentially more useful in communicating probability, there are different ways numbers can be used. Ranges of probabilities are one of various numerical methods used to express uncertainty about the occurrence of an event, though there is some disagreement in the literature about the effectiveness of ranges in communicating probabilistic information. Friedman and Zeckhauser (2014), for example, argue that there is little logical difference between ranges and single point estimates because most readers condense ranges into single numbers when making decisions. They therefore recommend point estimates as they communicate less ambiguity. Other studies, Budescu et al. (2014) and Wallsten et al. (1986a), in particular, assert that point estimates might be too precise to express an uncertain opinion, meaning uncertainty intervals are therefore preferable. While results are mixed about how numerical values should be expressed, many agree that WEPs most effectively convey the intentions of the communicator when they are combined with numerical values (Budescu et al. 2014, Wintle et al. 2019).

While combining words and numerical estimates can lead to more consistent public interpretations that better represent expert understanding of the situation, there are other factors that can influence an individual’s interpretation of probabilistic information. These factors include the interaction between base rate (an event’s frequency of occurrence) and expected outcome severity (the perceived severity of the future event). Previous work has shown that when base rate is kept consistent and the severity of an expected outcome changes, people tend to associate higher probabilities with higher severity outcomes (Harris and Corner 2011). For example, when a narrative described a chance of global sea levels rising, participants considering an island barely above sea level interpreted the probability of sea rise as higher than those considering an island protected from the sea. Thus, people's interpretation of probabilistic information is influenced by a "severity bias" which can lead them to associate higher probabilities with events that they deem as having more severe consequences (Harris and Corner 2011).

Situational context can also influence how people interpret probabilistic information, as participants in a medical study assigned lower probability values to outcomes on drug labels when no context was given (Fischer and Jungermann 1996). Additionally, the consideration of base rate was inversely related to frequency, meaning base rates played a larger role when an event was thought to occur less often (Fischer and Jungermann 1996). Finally, numerical estimates of probability can also lead people to draw more deliberate, logic-based conclusions, while WEPs can cause people to think more intuitively and express more uncertainty than is the case when numbers are used to communicate such information (Windschitl and Wells 1996). However, regardless of the method of communicating probability, there are many factors at play that need to be considered.

In summary, previous work outlines three main ways of communicating probabilistic information: through numbers, WEPs, or a combination thereof. Recent studies suggest that probability is best expressed using a combination of WEPs and numbers, though recommendations for the specific method of expressing the numeric aspect of this combination still varies across studies. Figure 1 outlines a schematic of these various methods of expressing probability as suggested by past research and applies them to a meteorological context. We focus on a weather forecast context because, while some studies have addressed how uncertainty in forecast information is communicated (e.g. Morss et al. 2008), little work has been done to evaluate how the public interprets WEPs in a meteorological context (for an exception see Wallsten et al. 1986b). Therefore, while Figure 1 focuses on the previously mentioned methods of expressing probability, it also includes a distinction between messages that do and do not communicate forecast information in order to focus on probability information inherent in forecasts.

3. Data and Methods

a. NWS Probabilistic Communication

Twitter messages ("tweets") from National Weather Service Weather Forecast Office (WFO) accounts were used to examine how forecasters communicate probabilistic information about severe thunderstorms and tornadoes. The tweets were obtained from the University of Oklahoma (OU) Center for Risk and Crisis Management (CRCM) Severe Weather and Social Media data collection program. This data collection program connects with Twitter's Streaming API to automatically and to continuously collect and archive tweets that include keywords and phrases that relate to extreme weather. In addition to the text of each tweet, the project archives the set of metadata provided by Twitter about the tweet itself and the user that created it (see Ripberger et al. 2014 for more information on the project). This analysis focuses on tweets from WFO accounts in 2018 that (1) included one or more of the following keywords/phrases: *severe weather*, *hail*, *tornado*, *severe storm thunderstorm* but (2) did *not* include one or more of the following words: *outlook*, *warning*, *watch*, *advisory*. Outlook, warning, watch, and advisory tweets were excluded because they typically relay information about a product (e.g., "Tornado Warning including Miltonvale KS") rather than a forecast. Given this study's focus on the communication of probabilistic forecasts, such tweets were excluded from data collection for the sake of getting relevant data, rather than simply "cut and paste" notifications about newly issued products. For consistency, the sample was limited to WFO accounts that posted more than 100 tweets that met these criteria. There were 89 WFO accounts that did so. This analysis focused on a random sample of 100 tweets from each of these 89 accounts (n = 8900) throughout 2018.

Following dataset construction, the tweets were initially coded based on the schematic provided in Figure 1. After looking at the data, however, it became apparent that the schematic required revision. While past research distinguished between various types of numeric expressions of probability, the NWS WFO tweets contained two different types of WEPs. In order to account for this, a new schematic was developed (Fig. 2) and the tweets were coded based on whether they included forecast information, were probabilistic or deterministic, and if they included WEPs or numerically expressed probabilities. Examples of tweets that fell into each category are provided in Table 1.

After distinguishing between tweets with numeric probabilities and WEPs, those with WEPs were broken down by whether the WEPs were qualified or unqualified. The two categories were differentiated by whether or not the WEP had a qualifier attached to the phrase (see examples in Table 2). If a qualifier was present, then the WEP was determined to be "qualified" and, if a qualifier was absent, then the WEP was determined to be "unqualified." Phrases with qualifiers expressing a change in a probability, but not what the original probability was, were added to the unqualified WEP category because they are relative and not easily quantifiable. Examples of such phrases include "increased chance," "decreased potential," and "diminishing threat." Similarly, WEPs such as "greatest potential," "highest threat," and "lesser chance" were categorized as unqualified. While these phrases do contain a word that acts like a qualifier, they only make sense in relation to an initial condition or in an external context which is not necessarily given or known by the reader.

The words "risk" and "threat" were included in the WEP category if the context in which they were used seemed to imply a probability of occurrence rather than the severity of the event. When the word "risk" was explicitly referencing the Storm Prediction Center's (SPC's) Convective Outlook, the word was not added to the WEP category. This is because the use of phrases like "moderate risk" in SPC's Convective

Outlooks corresponds to certain numerical thresholds and definitions rather than a probability of occurrence that NWS forecasters assign to an event. Also, when the word “threat” was used to describe the types of threats associated with an event, it was not added to the WEP category as it was not used to describe a probability. Finally, some WEPs described probabilities of spatial coverage and time, rather than the likelihood of an event’s occurrence. These were included in the dataset, though they were categorized as unqualified because they were vague and, if interpreted one way, could imply a probability of occurrence. An example of one of these vague phrases is “occasional chance.”

The above criteria in addition to the schematic in Figure 2 were used as the guidelines for coding the tweets. Any tweets with questionable interpretations were discussed between two coders before the final classification in order to minimize discrepancies in the data. As such, while messages from the NWS about current weather observations can be interpreted differently by different people (e.g. some might consider a probabilistic forecast to be deterministic based on wording), the guidelines each coder followed ensured the data was coded consistently. In general, discrepancies in interpretations were approached by considering how it would likely be understood by the general public, rather than from a more meteorological and analytical point of view. The minimal subjectivity that remains within the coding scheme should not discredit the methods employed, but rather speak to the inherent subjectivity that lies in forecast interpretations by anyone receiving the information.

To statistically examine the subjectivity of the coding methodology, the two lead researchers separately coded 130 tweets that were randomly selected by a third researcher from the original dataset. The results of this analysis are shown in Table 3. Intercoder agreement ranged from 91% to 96%, which indicates very minimal subjectivity in the coding process.

b. 2018 Survey

Public interpretations of WEPs that forecasters commonly used were assessed using data from the Severe Weather and Society Survey (WX), a yearly survey of U.S. adults that is administered by the Center for Risk and Crisis Management at the University of Oklahoma (see Silva et al. 2017, Silva et al. 2018 for more information on WX surveys). This analysis uses data from WX18 and WX19. WX18 was fielded in June 2018 as an online questionnaire completed by 3,000 U.S. adults (age 18+) across the CONUS that match the demographic characteristics of the U.S. population. The sample of participants was provided by Qualtrics, a company that maintains a diverse panel of internet users in the U.S. who agree to complete online surveys. Qualtrics uses a quota system to produce representative samples, a practice used by many other internet sampling companies. For WX18, the quotas capture a diverse sample of survey participants that generally represents the geographic and demographic attributes of the target population.

Data from WX18 allow for examination of how members of the public interpret a variety of *qualified* WEPs that WFOs used on Twitter. Interpretations were elicited using the following question:

People use different phrases to explain the possibility that a [severe thunderstorm | tornado] will happen. When you see the following phrases, what percent chance comes to mind? Please indicate the chance as a percent that ranges from 0 to 100, where 0 means no chance and 100 means that it is certain.

- ***Very low** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Extremely low** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Pretty low** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Small** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Low** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Slight** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Moderate** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Good** [chance/probability] of a [severe thunderstorm/tornado]*
- ***Significant** [chance/probability] of a [severe thunderstorm/tornado]*

Each respondent was given an opportunity to evaluate a random sample of five phrases. The words "chance" and "probability" were randomized in the experiment to identify possible differences between the two. While people associated the word "chance" with slightly lower percentages, the variation was not statistically significant (see Appendix A). The event was also randomized to identify the possibility that event severity may influence perceptions of probability. For both qualified and unqualified WEPs, people associated nominally lower probabilities with severe thunderstorms than with tornadoes, but the differences were negligible (see Appendix B).

c. 2019 Survey

WX19 was fielded in June/July 2019 using the same procedure and methodology as WX18 (Silva et al. 2018). Data from WX19 allow for an examination of how members of the public interpret a variety of *unqualified* WEPs that WFOs used on Twitter. Interpretations were elicited using the following survey experiment:

Forecasters use different phrases to explain the possibility that a [severe thunderstorm | tornado] will happen. When you see the following phrases, what percent chance comes to mind? Please indicate the chance as a percent that ranges from 0 to 100, where 0 means no chance and 100 means that it is certain.

- *There is a* ***chance*** *of [severe thunderstorms/tornadoes] this afternoon and evening*
- *[Severe thunderstorms/Tornadoes] are* ***possible*** *this afternoon and evening*
- *[Severe thunderstorms/Tornadoes]* ***may*** *occur this afternoon and evening*
- *[Severe thunderstorms/Tornadoes] are* ***expected*** *this afternoon and evening*

Each respondent was given an opportunity to evaluate all of the phrases. As with WX18, the event was randomized to identify the possibility that event severity may influence perceptions of probability; no significant differences were identified (see Appendix A).

4. Results

a. NWS Probabilistic Communication

Within the database of 8900 severe weather related tweets from 89 WFOs, 67.1% of the tweets relayed forecast information to the public. Of these tweets, 65.3% communicated probabilistic forecast information, with the rest being deterministic-type forecasts. This resulted in 3902 out of 8900 tweets that expressed probability either numerically, with words, or using a combination of both. 99.9% of these probabilistic tweets used some kind of WEP while only 0.07% contained exclusively numeric expressions of probability. Only 0.08% of the tweets used both numbers and words; these tweets were placed into the WEP category due to this paper's focus on WEPs.

When communicating a severe weather forecast on Twitter, NWS offices primarily used unqualified WEPs to do so. In fact, unqualified phrases are used about 48 times more often than qualified words (Fig. 3). In the rare case that a qualified phrase was used, it was more common for it to be combined with another occurrence of an unqualified word in the same message. The top ten most commonly used phrases from each category, the percentage of times the phrase was used, and an example tweet containing each phrase, are shown in Tables 4 and 5. For example, out of all the times the NWS used a qualified word, 14.4% of the time the phrase "low threat" was used. While there was not as much variance in the frequency of the qualified WEPs (Table 4), there is a large difference between the most used unqualified WEP and the rest of the phrases. The most commonly used unqualified WEP was "possible" as it was used 39.4% of the time an unqualified WEP was used (Table 5). The next highest unqualified WEP was "expected" which was used 8.81% of the time (Table 5). Given that unqualified words were used more frequently by the NWS than qualified WEPs, "possible" was the most commonly used WEP by the NWS. There was also a larger variety of unqualified WEPs used, with 61 different qualified words and phrases used and 79 different unqualified words.

b. 2018 survey results

In the 2018 Severe Weather and Society Survey, participants were asked to numerically interpret qualified WEPs by providing their assessment of the implied probabilities for each one. The distributions of the values participants associated with each WEP show a general trend of smaller probabilities being associated with WEPs such as “very low” and “small,” and higher probabilities associated with terms such as “significant” and “good” (Fig. 4a). To quantify this trend, mean values for all analyzed WEPs are given in Table 6. The WEP with the lowest mean value as interpreted by the public is “very low” with an average of 14.0%. “Very small” had a similar interpretation with an average of 14.4%. Typically, WEPs such as “low” and “slight” were thought of as representing a higher probability than their “very low” and “very small” counterparts. What might be considered surprising is that “good” was on average interpreted as a higher value than “moderate” by more than 10%. Finally, “significant” was interpreted as representing the highest probability with an average of 70.7%.

At first glance, the public interpreted WEPs such as “very low” and “small” more consistently than they did WEPs such as “significant” and “moderate” (Fig. 4a). Standard deviations from the mean range from 14.7% for “small” to 26.9% for “significant” (Table 6). However, the average standard deviation for this set of unqualified WEPs is 17.9%. Therefore, the value interpretations by the public vary drastically for each WEP (Fig. 4a). The next highest values of standard deviation after “small” are for “very low,” “very small,” “slight,” and “low” (Table 6). On the other hand, higher standard deviation values are calculated for “moderate,” “good,” and “significant” (Table 6), which could indicate that it is more difficult for participants to interpret these higher probability WEPs than it is the lower ones.

The standard deviation values for each qualified WEP implies that people’s interpretation of words that indicate a low probability tend to be within the same, relatively restricted range. Another way to illustrate this is with the lower and upper bounds of the interquartile range (IQR) of the data (Table 6). The WEPs “very low” and “very small” both range from 5% to 20%, with 5% being the 25th quantile and 20% being the 75th quantile. “Small” and “low” range from 10% to 25%. For the higher probability words, “moderate” ranges from 25% to 50%, “significant” from 60% to 90%, and “good” has bounds of 30% to 60%. Overall, we see that the lower probability WEPs have IQRs of about 15%, while these higher probability WEPs have IQRs of about 30%.

c. 2019 survey results

In the 2019 Severe Weather and Society Survey, participants were asked to numerically interpret unqualified WEPs. Some of the most used unqualified WEPs from the Twitter results (Table 5) were evaluated in this portion of the survey. For “chance,” “may,” and “possible,” the distributions were very similar with people associating values near 50%, 25%, and 10% most frequently (Fig. 4b). Conversely, the distribution for “expected” had much more variance and peaked near 80% (Fig. 4b). These results likely reflect the fact that these terms are relatively vague, and thus respondents are anchoring to common probabilities like 25% and 50%.

The unqualified WEP with the lowest mean value was “chance” with an average of 30.9% (Table 6). Meanwhile, “expected” had the highest average of 59.4%. As also seen in Table 7, standard deviations for the unqualified words tested in WX19 range from 21.3% for “chance” to 27.8% for “expected.” The average standard deviation was 23.3%. This value is higher than the average standard deviation for qualified WEPs (17.9%), suggesting that unqualified WEPs, while extremely common in NWS tweets, likely communicate less information to the public, since they are not as consistently understood by readers. Supporting this claim, “may” and “possible” have a 25th quantile of 20% and a 75th quantile of 50%. “Chance” ranges from 15% to 50% and “expected” goes from 45% to 80% (Table 7). Therefore, unqualified WEPs result in less agreement in interpretation by participants than do the qualified WEPs shown in Table 6. This poses a significant concern, considering a large majority of WEPs used by NWS WFOs when communicating forecast uncertainty.

5. Conclusions and Discussion

Previous studies within non-meteorological contexts have suggested that the communication of probabilistic information often differs in terms of how the communicator wants to portray probability and how the audience would prefer to receive such information. Generally, people tend to communicate probabilistic information using words, but want to receive the same type of information numerically (Fischer and Jungermann 1996, Willems et al. 2019, MacLeod and Pietravalle 2017, Wintle et al. 2019). Other work suggests that a combination of words and numbers may be ideal for increasing understanding of probability (Budescu et al. 2014, Wintle et al. 2019). The main intent of this study was to understand which tactics NWS WFOs use to communicate probability in severe thunderstorm and tornado forecasts, and to combine this information with how members of the public understand the types of communication that are used.

Initially, we expected that the NWS would communicate probabilities using qualified WEPs or numbers, with few instances where both were used simultaneously. While coding the data, it became apparent that the NWS primarily communicated probabilistic information using WEPs. These WEPs were either unqualified (WEPs without a qualifier) or qualified (WEPs with an associated, qualifying word), with the NWS primarily using unqualified WEPs. Again, previous studies have shown that communicating probability with words tends to be less effective than with numbers (Willems et al. 2019, Wintle et al. 2019). Therefore, these results raise questions about how effective unqualified WEPs are at communicating probability versus qualified WEPs, and whether WFOs are using unqualified WEPs to communicate general forecast uncertainty instead of a probability forecast.

When looking at how the public interprets both qualified and unqualified WEPs within a meteorological context, the qualified WEPs that represented lower probabilities had the lowest standard deviations. These lower deviations from the mean may indicate that people interpreted the WEPs similarly and that there is less confusion over what terms such as “very low” and “small” mean as compared to words such as “good” and “expected.” This could further imply that people are better at interpreting words that represent lower probabilities as compared to their less used, higher probability counterparts.

Overall, variability was higher for numerical interpretations of unqualified words than for the qualified examples tested in WX18. The only qualified WEPs that had a higher standard deviation than some of the unqualified words were “good” and “significant.” When comparing the average of the standard deviations for the qualified and unqualified WEPs, the unqualified WEPs had a higher mean deviation. This implies that people are less certain about what unqualified WEPs mean, and may indicate that the terms tested in WX19 are more vague than those tested in WX18. Given that the NWS primarily communicates probabilities using these unqualified terms, the message they are communicating is likely being interpreted differently by different people. For example, if the NWS sends out a message saying “storms are expected,” people could interpret that to mean anything between a 45% to 80% chance of storms occurring. This implies that in general, using Words of Estimative Probability to communicate severe weather forecasts may be ineffective if the goal is to provide consistent and easy to interpret information about forecast uncertainty that people can use to make protective action decisions.

Based on the study results, we recommend increasing the use of qualifiers when WEPs are used. As seen from the survey comparison, people tend to interpret qualified WEPs more consistently than their unqualified counterparts, giving people a better understanding of forecast information. If the intent behind the forecast information is to raise a general awareness of the potential occurrence of a severe weather event, the use of unqualified WEPs might be sufficient. It should be noted, however, that the level of public awareness will vary greatly with such unqualified WEPs and, if a more concise interpretation is desired, adding a qualifier will lessen the variability in how people interpret forecast information.

While these findings provide important information about how WFOs communicate probabilistic information and how members of the public interpret this information, we note that this study is relatively narrow in its focus on severe thunderstorms and tornadoes. It is quite possible that WFOs communicate probabilistic information in different ways for different hazards. For example, we hypothesize that forecasters will use more precise WEPs (and numerical expressions of probability) in domains where they receive relatively specific probabilistic guidance (such as the probability of precipitation) and less precise WEPs in domains where they receive less specific guidance. Likewise, there may be differences in communication for relatively small-scale discrete events (like thunderstorms and tornadoes) and large-scale events like winter storms and hurricanes. Given these possibilities, more research is necessary before we can generalize the results of this study to other forecast contexts.

In addition, we note that this study focuses exclusively on WFO text messages on Twitter, one of many communication outlets that WFOs use during severe weather events. It is likely that forecasters express probabilities in different ways on different mediums; for example, they might use text on Twitter to note basic uncertainties and graphics to provide more specific probabilistic information to people who have more time to engage with the information. Likewise, they may supplement relatively short tweets with more in-depth Facebook messages and/or YouTube videos that use more precise expressions of probability to convey forecast information. Again, more research is necessary before we can generalize the results of this study to other methods of communication.

Finally, we note that the survey results utilized in this study focus on consistency of interpretation across segments of the US adult population. We do not examine consistency with forecaster intentions, nor do we examine the implications of inconsistent interpretation. We therefore urge future research on what forecasters mean when they use different WEPs and, perhaps more importantly, what happens when people misinterpret these meanings. Do people draw incorrect inferences about risk and therefore make poor decisions when they misjudge the probability of an event because they misinterpret a vague statement like "severe storms are possible tomorrow"? What about the opposite? What are the costs of overly precise forecast statements that do not pan out? These are open questions that require more attention as researchers continue to address questions about how to best communicate probabilistic information in weather forecasts.

Acknowledgments. The data for this project were collected with support from the University of Oklahoma. The analysis was conducted with support from NOAA's Office of Weather and Air Quality through the U.S. Weather Research Program.

APPENDIX A

WX18 Question Variations

Qualified WEPs	Chance Mean (%)	Chance Standard Deviation (%)	Probability Mean (%)	Probability Standard Deviation	Mean Difference (%)	Standard Deviation Difference

				(%)		(%)
Very low	13.6	14.5	14.4	15.1	-0.87	-0.60
Very small	13.7	14.2	15.0	16.1	-1.26	-1.94
Pretty low	16.8	15.9	17.6	16.7	-0.81	-0.80
Small	17.7	14.4	19.7	15.0	-2.05	-0.65
Low	19.9	15.1	21.5	16.6	-1.57	-1.54
Slight	22.0	16.4	23.4	15.2	-1.34	-1.13
Moderate	39.4	20.0	39.9	18.9	-0.48	1.11
Good	47.8	21.5	49.8	22.9	-2.04	-1.36
Significant	70.6	25.8	71.0	28.1	-0.40	-2.24

APPENDIX B

WX19 Question Variations

Qualified WEPs	**T-Storm Mean (%)**	**T-Storm Standard Deviation (%)**	**Tornado Mean (%)**	**Tornado Standard Deviation (%)**	**Mean Difference (%)**	**Standard Deviation Difference (%)**
Very low	14.1	13.9	14.0	15.8	0.09	-1.88
Very small	14.3	12.9	14.4	17.1	-0.12	-4.16
Pretty low	17.2	15.4	17.1	17.2	0.16	-1.86
Small	19.8	14.6	17.6	14.7	2.18	-0.11
Low	21.4	14.7	20.1	16.8	1.22	-2.09
Slight	23.2	16.2	22.1	15.4	1.11	0.80
Moderate	40.8	18.8	38.4	20.2	2.43	-1.41
Good	49.2	21.1	48.3	23.1	0.84	-1.92
Significant	72.0	26.2	69.6	27.5	2.44	-1.22
Unqualified WEPs	**T-Storm Mean (%)**	**T-Storm Standard Deviation (%)**	**Tornado Mean (%)**	**Tornado Standard Deviation (%)**	**Mean Difference (%)**	**Standard Deviation Difference (%)**
Chance	33.3	21.5	28.5	20.9	4.83	0.51
May	37.4	21.8	30.4	21.7	7.06	0.08
Possible	38.7	21.8	33.7	22.2	4.99	-0.36
Expected	62.0	26.2	56.8	29.1	5.16	-2.86

REFERENCES

AMS Council. (2008). Enhancing weather information with probability forecasts. Bulletin of the American Meteorological Society, 89, 1049-1053.

Beyth-Marom, R., 1982: How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. Journal of Forecasting, **1**, 257-269, 10.1002/for.3980010305.

Budescu, D., S. Broomell, and H. Por, 2009: Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change. *Psychological Science*, **20**, 299–308, https://doi.org/10.1111/j.1467-9280.2009.02284.

Budescu, D., H. Por, S. Broomell, 2012: Effective communication of uncertainty in the IPCC reports. *Climate Change*, **113**, 181-200, 10.1007/s10584-011-0330-3

Budescu, D., H. Por, S. Broomell, and M. Smithson, 2014: The Interpretation of IPCC Probabilistic Statements Around the World. *Nature Climate Change*, **4**, 508-512, 10.1038/nclimate2194.

Fischer, K., and H. Jungermann, 1996: Rarely Occurring Headaches and Rarely Occurring Blindness: Is Rarely = Rarely?. *Journal of Behavioral Decision Making*, **9**, 153-172, https://doi.org/10.1002/(SICI)1099-0771(199609)9:3<153::AID-BDM222>3.0.CO;2-W.

Friedman, J. A., and R. Zeckhauser, 2014: Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden. *Intelligence and National Security*, **30**, 77-99, 10.1080/02684527.2014.885202.

Harris, A. J. L., and A. Corner, 2011: Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37**, 1571-1578, http://dx.doi.org/10.1037/a0024195.

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. Meteorological Applications, 17(2), 180-195.

MacLeod, A., and S. Pietravalle, 2017: Communicating risk: variability of interpreting qualitative terms. *EPPO Bulletin*, **47**, 57-68, https://doi.org/10.1111/epp.12367.

Morss, R.E., J.L. Demuth, and J.K. Lazo, 2008: Communicating Uncertainty in Weather Forecasts: A Survey of the U.S. Public. *Wea. Forecasting*, **23**, 974–991, https://doi.org/10.1175/2008WAF2007088.1.

National Institute of Standards and Technology. (2013). Technical investigation of the May 22, 2011, tornado in Joplin, Missouri. Final Report.

National Research Council. (2006). Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts. National Academies Press.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A proposed next-generation paradigm for high-impact weather forecasting. Bull. Amer. Meteor. Soc., **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Silva, C. L., Ripberger, J. T., Jenkins-Smith, H. C., Krocak, M, 2017: Establishing a Baseline: Public Reception, Understanding, and Responses to Severe Weather Forecasts and Warnings in the Contiguous United States. University of Oklahoma Center for Risk and Crisis Management, http://risk.ou.edu/downloads/news/WX17-Reference-Report.pdf

Silva, C. L., Ripberger, J. T., Jenkins-Smith, H. C., Krocak, M, and Wehde, W. W., 2018: Refining the Baseline: Public Reception, Understanding, and Responses to Severe Weather Forecasts and Warnings in the Contiguous United States. University of Oklahoma Center for Risk and Crisis Management, http://risk.ou.edu/downloads/news/WX18-Reference-Report.pdf

Wallsten, T. S., D. V. Budescu, A. Rapoport, R. Zwick, B. Forsyth, 1986a: Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology:General*, **115**, 348-365, http://dx.doi.org/10.1037/0096-3445.115.4.348.

Wallsten, T. S., S. Fillenbaum, and J. A. Cox, 1986b: Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, **25**, 571-587, https://doi.org/10.1016/0749-596X(86)90012-4.

Willems, S. J. W., C.J. Albers, I. Smeets, 2019: Variability in the interpretation of Dutch probability phrases - a risk for miscommunication. eprint arXiv:1901.09686.

Windschitl, P. D., and G. L. Wells, 1996: Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, **2**, 343 - 364, 10.1037//1076-898X.2.4.343.

Wintle, B.C., H. Fraser, B.D. Wills, A.E. Nicholson, F. Fidler, 2019: Verbal probabilities: *Very likely* to be *somewhat* more confusing than numbers. *PLoS One*, **14**, 1-18, https://doi.org/10.1371/journal.pone.0213522.

Zwick, R., T. Wallsten and D. Budescu, 1993: Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. *Management Science*, **39**, 176-190, 10.1287/mnsc.39.2.176

TABLES AND FIGURES

Table 1. Tweets were categorized based on a specific set of criteria representative of the most common ways of expressing uncertainty that were mentioned in the literature. This table shows an example of a tweet from each of the first three categories. The "Forecast" and "Not a Forecast" categories were not literature-based, rather they were added later so that the focus of this study would remain on communication of forecast information specifically.

Forecast	**Not a Forecast**
Heavy Rain, Large Hail And Damaging Winds Possible Today And Tonight. #sdwx #mnwx (NWS Aberdeen)	The image below shows the number of hail reports equal to or greater than 3 inches that occurred in El Paso county, CO from 1950 to 2017 (68 yrs). In 2018, we have had at least 2 days were hail &gt;= 3', with one of these events occurring in the middle of the night. #cowx (NWS Pueblo)
Probabilistic Forecast	**Deterministic Forecast**
Scattered thunderstorms will continue this evening. While widespread severe weather is not expected, an isolated strong to severe storm is possible. #iawx (NWS Des Moines)	Some scattered thunderstorms are ongoing across the northwest portion of our area tonight. Expect heavy downpours and occasional small hail with these storms. (NWS Wilmington)
Words of Estimative Probability	**Numeric Probabilty**
Expect above normal warmth &amp; rain &amp; storm chances today ahead of a developing cold front &amp; strong storm system. Expect the best chances of storms &amp; severe weather by late afternoon/early evening in the Delta, but some conditional threat is possible by late afternoon elsewhere. (NWS Jackson MS)	815am: Thunderstorm activity flaring up with Caribbean disturbance. 40% chance of development within the next 5 days as it moves into the eastern or central Gulf. Heavy rain remains a threat for South FL late this weekend into Memorial Day weekend #flwx https://t.co/JfcD43cTBy (NWS Miami)

Table 2. List of accepted quantifiers used in tweets that were coded as being probabilistic and qualified.

Low	Slight	Small
Good	Limited	Isolated
Low-end	Some	Marginal
Minimal	Very	Very low
Very small	Big	Bit
High-end	High	Highly conditional
Little	Minimum	Moderate
Not much	Outside	Pretty good
Pretty	Pretty low	Significant
Slim	Some conditional	Very good
Very limited	Very minimal	

Table 3. After coding the 8900 tweets used in the dataset, coders were given a random set of 130 tweets out of the dataset to recode in order to test intercoder reliability. High percentage values of agreement indicate that the subjectivity inherent in the coding scheme is minimal.

	Forecast vs. Not Forecast	**Probabilistic vs. Not Probabilistic**	**WEP vs. Numeric**	**Classification of Qualified vs. Unqualified**
Agreements (#)	125	118	122	121
Disagreements (#)	5	12	8	9
Accuracy (%)	96	91	94	93%

Table 4. For each of the top ten qualified WEPs observed in the database of severe weather tweets, an example is given for how that phrase is used by WFOs. The specific WFO for which the example is from is shown in parentheses. The percent that each word was used out of every occurrence of a qualified WEP is shown on the right.

Qualified WEPs	Example	Percent (%)
Low Threat	A few isolated thunderstorms moving into the I-69 corridor this evening. **Low threat** for severe wind, hail. #miwx (NWS Detroit)	14.4
Low Chance	920am - Severe storms possible this afternoon w/damaging wind the greatest threat. **Low chance** for a tornado north of I-94. #swiwx #wiwx (NWS Milwaukee)	12.3
Slight Chance	There is a **slight chance** for #thunderstorms Sunday afternoon for W WA. Small hail &amp; brief downpours are possible as well. If you are outdoors and hear thunder, head indoors! #WAwx (NWS Seattle)	11.2
Low Risk	Strong to severe thunderstorms possible Friday afternoon and evening, mainly across the interior. Damaging winds possible, along with localized flash flooding and the **low risk** of an isolated tornado (NWS Boston)	5.88
Small Chance	Pleasant conditions this weekend. There is a **small chance** for an afternoon thunderstorm on Saturday. #swiwx (NWS Milwaukee)	5.35
Good Chance	@DPA_Insight Mountain areas will have a **good chance** at some convection and thunderstorm activity this afternoon. Not the coast, though… (NWS San Diego)	4.28
Limited Threat	Another round of storms is possible this afternoon for eastern #OKwx and western #ARwx, possibly severe. Main threats are damaging winds with **limited threat** for tornadoes. Stay weather aware today. (NWS Tulsa)	3.21
Low Confidence	@IowaStormChasr Yes, severe weather is likely to occur over the next 2 weeks. It's June in Iowa, which is the peak severe weather month for this state. That being said, **low confidence** with any significant severe weather event attm. (NWS Des Moines)	3.21
Isolated Chance	RT @NWSNorthPlatte: Thunderstorm chances continue Friday and Saturday with an **isolated chance** of severe weather both days. #NEwx (NWS Cheyanne)	2.14
Low Potential	**Low potential** for severe storms this afternoon. 60 mph winds and 1 inch hail are the main threat. #iawx #ilwx #mowx (NWS Quad Cities)	2.14

Table 5. For each of the top ten unqualified WEPs observed in the database of severe weather tweets, an example is given for how that phrase is used by WFOs. The specific WFO for which the example is from is shown in parentheses. The percent that each word was used out of every occurrence of an unqualified WEP is shown on the right.

Unqualified WEPs	Example	Percent (%)
Possible	Strong to severe storms are **possible** this afternoon through 9 pm. The main threat will be damaging wind gusts of 40-60 mph...with heavy rain possible. Keep an eye on the weather today. #alwx (NWS Birmingham)	39.4
Expected	Strong to severe thunderstorms are **expected** late this afternoon through this evening. Stay weather aware and be ready to seek shelter if necessary. (NWS DC/Baltimore)	8.81
May	A few severe storms **may** develop, with damaging wind as the main threat. Large hail up to 1 inch &amp; heavy rain also possible #kswx #newx #cowx (NWS Goodland)	8.60
Could	5:54 pm Radar Update: Some scattered storms ongoing, generally along and north of Hwy 36 moving to the SE. A few storms **could** become strong with hail and gusty winds possible. Overall severe threat remains low with these storms. (NWS Kansas City)	7.67
Chance	The week will start with a chilly wet Monday morning lasting into the afternoon near the Lake. The pick day of the week will be Tuesday. The end of the week features a warming trend with daily shower and thunderstorm **chances** (NWS Milwaukee)	5.64
Potential	A cold front will make its way into Idaho on Wednesday, setting the stage for severe thunderstorm **potential** for East Idaho. Large hail and damaging winds possible. (NWS Pocatello)	4.53
Likely	Thunderstorms are **likely** tonight. Some of the storms could produce large hail and heavy rainfall rates over 1 inch per hour.#iawx #ilwx #mow (NWS Quad Cities)	4.00
Threat	Severe weather **threat** today. Tornado **threat** is low but not zero. More of a large	3.11

	hail/damaging winds threat. Few storms by 5 PM towards #Ozona, Sterling City, and #Sweetwater with the storms expanding as they move towards #SanAngelo/#Abilene/#Brownwwood and #Brady. #sjtwx #txwx (NWS San Angelo)	
Not Expected	In addition to the heavy rains, isolated damaging wind potential will exist with overnight thunderstorms across parts of far SE #OKwx. Widespread severe weather is **not expected**, with the highest severe weather chances well south of the Red River. #arwx (NWS Tulsa)	2.59
Can’t Be Ruled Out	A few showers possible across our northern areas this afternoon. A thunderstorm **can't be ruled out**, with lightning the main threat along with and brief, heavy rain. #nywx (NWS Binghamton)	2.48

Table 6. Out of all WEPs included in WX18, those that were also found in the severe weather tweets were analyzed. Respondents were asked to give each word a numerical value. The mean percentage given for each WEP is given, as well as standard deviation from the mean and the 25th and 75th quantiles.

Qualified WEPs	**Mean (%)**	**Standard Deviation (%)**	**25th Quantile (%)**	**75th Quantile (%)**
Very low	14.0	14.8	5	20
Very small	14.4	15.2	5	20
Pretty low	17.2	16.3	10	20
Small	18.7	14.7	10	25
Low	20.7	15.9	10	25
Slight	22.7	15.8	10	30
Moderate	39.6	19.5	25	50
Good	48.8	22.2	30	60
Significant	70.7	26.9	60	90
AVERAGE		17.9		

Table 7. The top four unqualified WEPs found in the severe weather tweets were included in WX19. Respondents were asked to give each word a numerical value. The mean percentage given for each WEP is given, as well as standard deviation from the mean and the 25th and 75th quantiles.

Unqualified WEPs	**Mean (%)**	**Standard Deviation (%)**	**25th Quantile (%)**	**75th Quantile (%)**
Chance	30.9	21.3	15	50
May	33.9	22.0	20	50
Possible	36.2	22.1	20	50
Expected	59.4	27.8	45	80
AVERAGE		23.3		

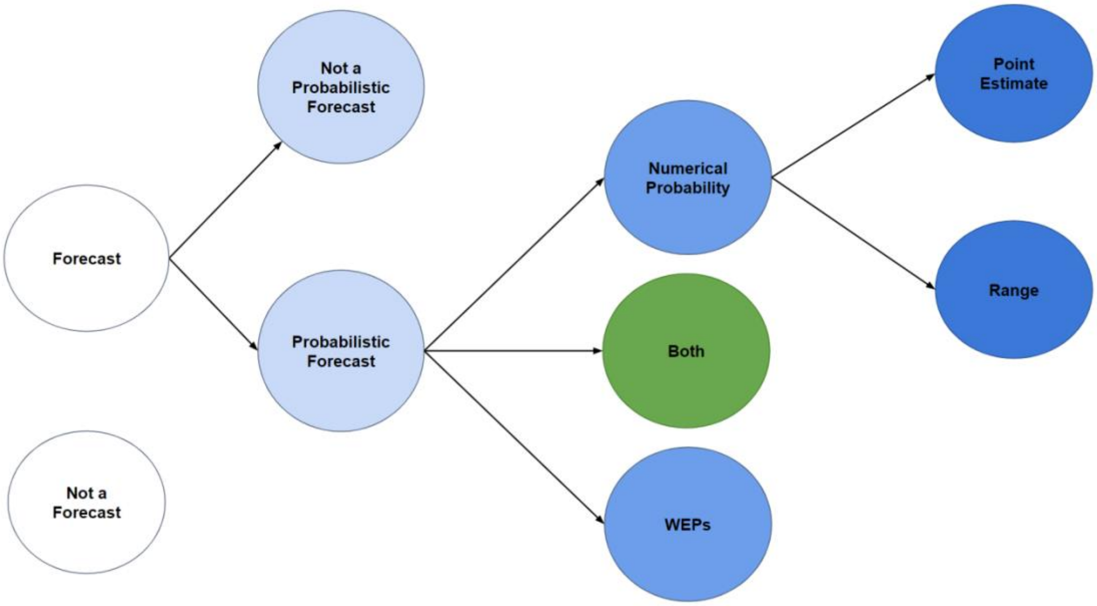

Figure 1. This schematic was developed from past work on the communication of probabilistic information and applied to a meteorological context. It shows the three main methods of expressing probability as suggested by past research in addition to differentiating between whether or not something is a forecast and whether or not a forecast is probabilistic in nature. The green circle represents the method of expressing probability that research suggests is best for communicating probabilistic information.

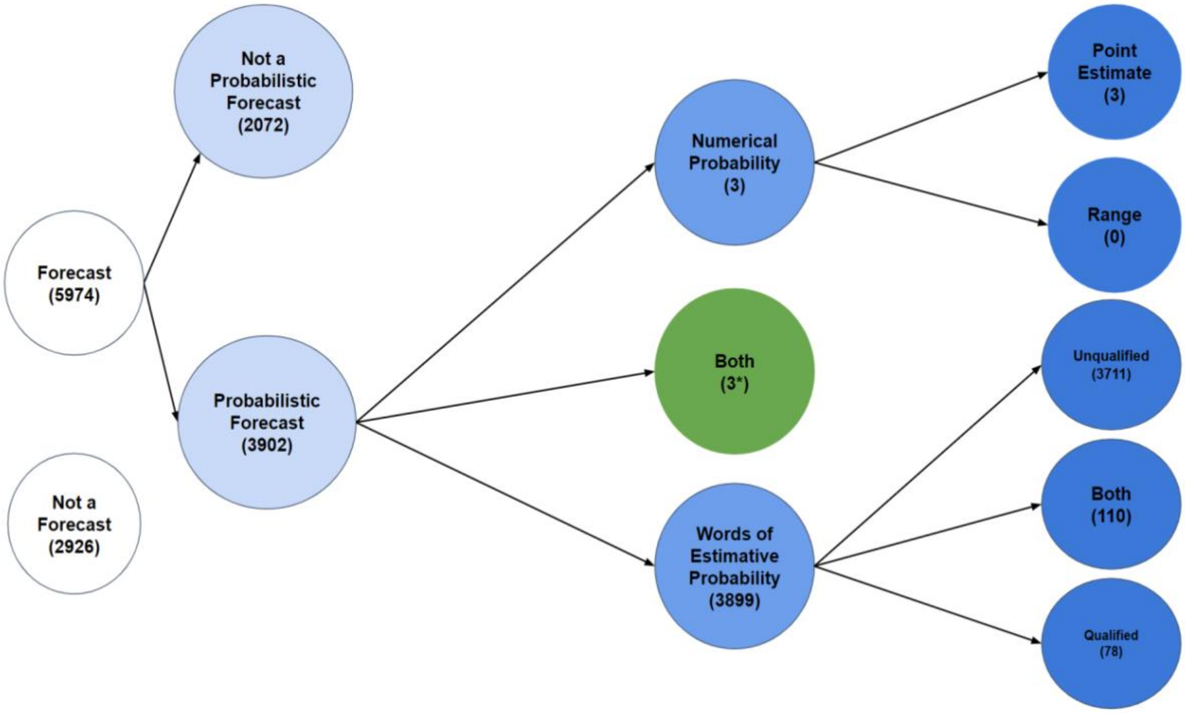
Not a Probabilistic Forecast (2072)
Forecast (5974)
Probabilistic Forecast (3902)
Not a Forecast (2926)
Numerical Probability (3)
Both (3*)
Words of Estimative Probability (3899)
Point Estimate (3)
Range (0)
Unqualified (3711)
Both (110)
Qualified (78)

Figure 2. All tweets were coded based on this schematic and the numbers in parentheses represent the number of tweets from each category. For example, if a tweet relayed forecast information but in a non-probabilistic way, then that tweet would be included in the "Not a Probabilistic Forecast" category, and would not be coded beyond that. This schematic was built off of Fig. 1, with variations of WEPs added. *The value here was added to the WEP category in order to focus more on the various types of WEPs used by the NWS. The number shown for the WEP category includes this value.

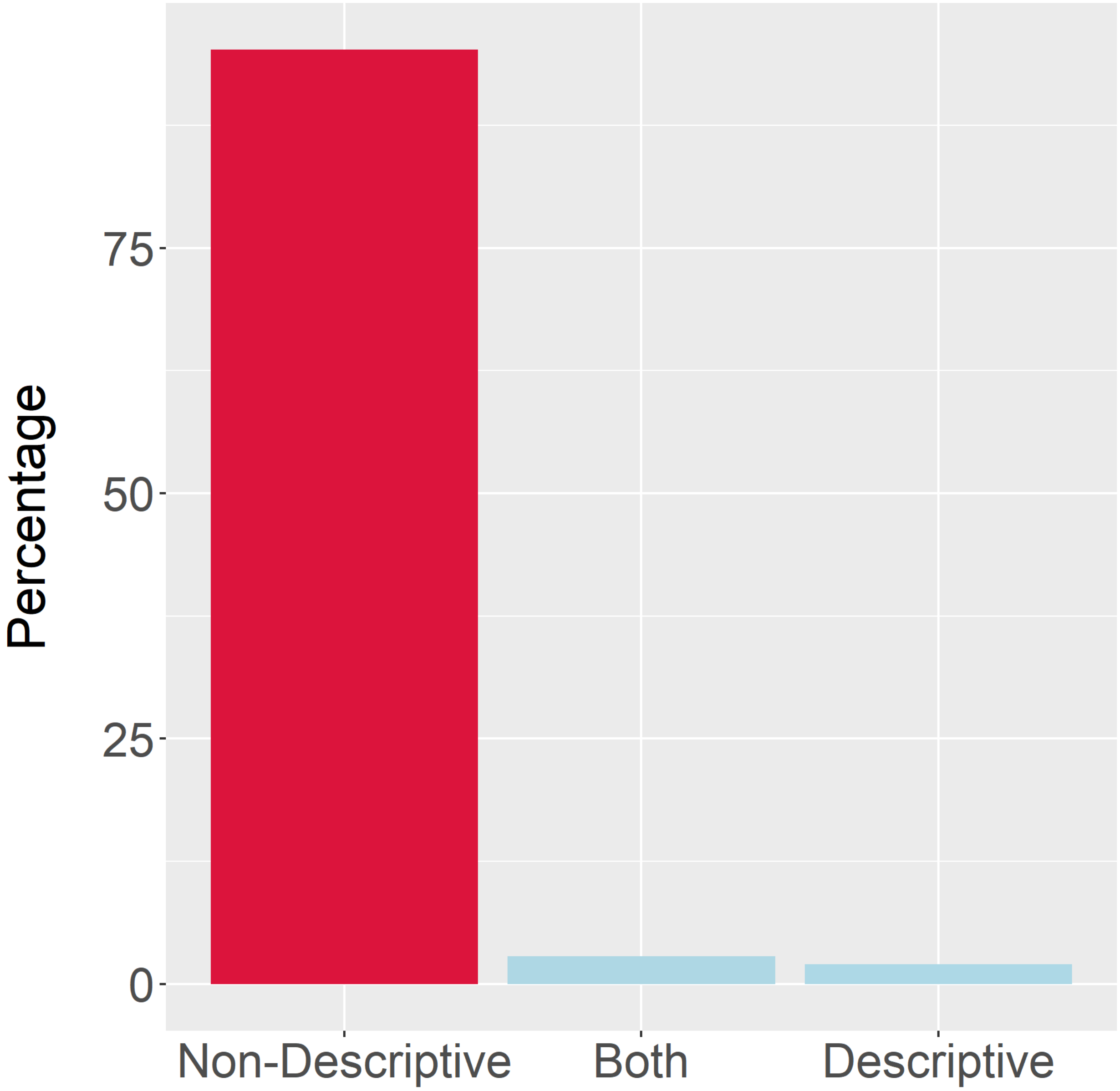

Figure 3. Percentage of tweets that contained unqualified and qualified WEPs along with tweets that contained both. The percentages are of all tweets coded as containing words of estimative probability and not out of the total number of tweets.

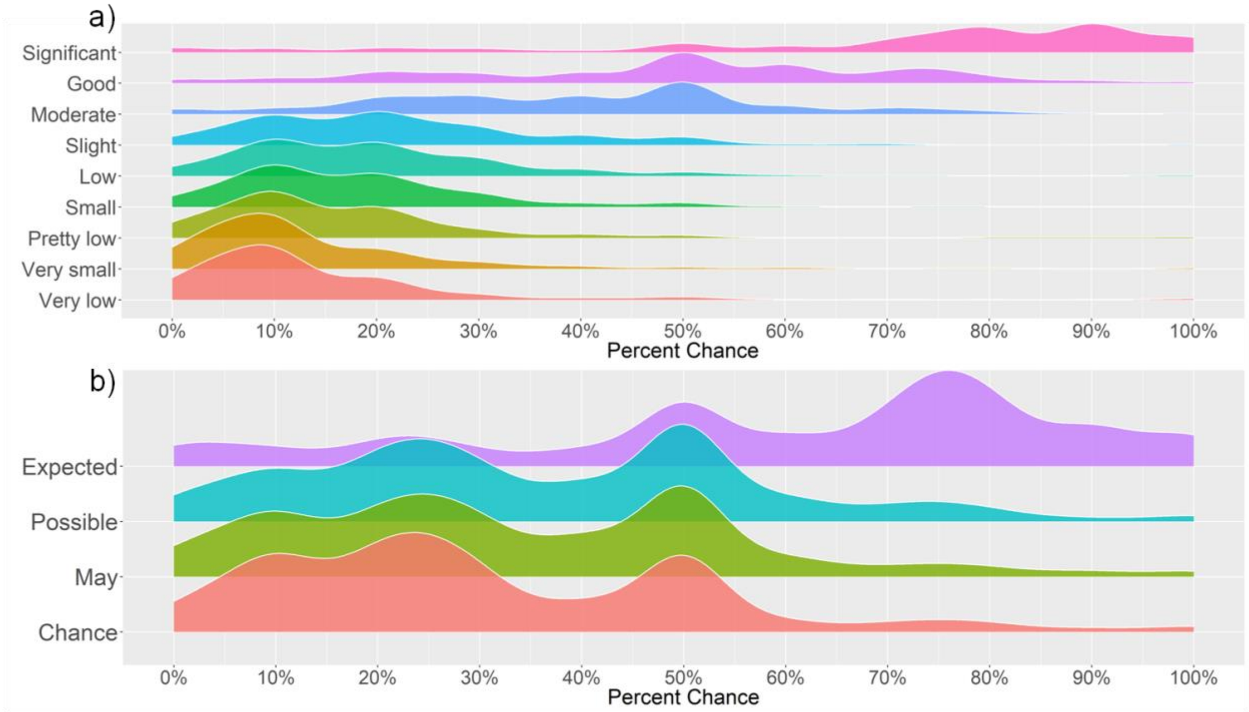

Figure 4. Density curves show the distribution of responses when survey respondents were asked to assign a percentage to various (a) qualified and (b) unqualified WEPs.