Exploring the Differences in SPC Convective Outlook Interpretation Using Categorical and Numeric Information

MAKENZIE J. KROCAK,[a,b,c] JOSEPH T. RIPBERGER,[a] SEAN ERNST,[a,b,c] CAROL L. SILVA,[a] AND HANK C. JENKINS-SMITH[a]

[a] *Institute for Public Policy Research and Analysis, University of Oklahoma, Norman, Oklahoma*
[b] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, National Weather Center, Norman, Oklahoma*
[c] *NOAA/Storm Prediction Center, National Weather Center, Norman, Oklahoma*

(Manuscript received 21 July 2021, in final form 3 December 2021)

ABSTRACT: While previous work has shown that the Storm Prediction Center (SPC) convective outlooks accurately capture meteorological outcomes, evidence suggests stakeholders and the public may misinterpret the categorical words currently used in the product. This work attempts to address this problem by investigating public reactions to alternative information formats that include the following numeric information: 1) numeric risk levels (i.e., "Level 2 of 5") and 2) numeric probabilities (i.e., "a 5% chance"). In addition, it explores how different combinations of the categorical labels with numeric information may impact public reactions to the product. Survey data comes from the 2020 Severe Weather and Society Survey, a nationally representative survey of U.S. adults. Participants were shown varying combinations of the information formats of interest, and then rated their concern about the weather and the likelihood of changing plans in response to the given information. Results indicate that providing numeric information (in the form of levels or probabilities) increases the likelihood of participants correctly interpreting the convective outlook information relative to categorical labels alone. Including the categorical labels increases misinterpretation, regardless of whether numeric information was included alongside the labels. Finally, findings indicate participants' numeracy (or their ability to understand and work with numbers) had an impact on correct interpretation of the order of the outlook labels. Although there are many challenges to correctly interpreting the SPC convective outlook, using only numeric labels instead of the current categorical labels may be a relatively straightforward change that could improve public interpretation of the product.

SIGNIFICANCE STATEMENT: The SPC convective outlook contains vital information that can help people prepare for a severe weather event. The categorical labels in this product are often ordered incorrectly by members of the public. This work shows using numeric levels or probabilities reduces the tendency for people to order the levels incorrectly.

KEYWORDS: Forecasting; Communications/decision making; Decision making; Operational forecasting; Risk assessment; Societal impacts

1. Introduction and background

The convective outlook has been issued by the NOAA/Storm Prediction Center (SPC) since March 1952 (Corfidi 1999). The main goal of the outlook is to communicate the severe weather risk from one to eight days in advance of an event. Since its inception, the structure and underlying forecast information of the convective outlook have undergone a number of changes. In the 1970s, the SPC started including categorical names to differentiate three levels of severe weather threat. These original three levels were Slight, Moderate, and High (Corfidi 1999). In the early 2000s, those categories were aligned with the probabilistic forecast of an event occurring within 25 miles of a point (see Fig. 1 in Ernst et al. 2021). Most recently, two additional categories were added in an attempt to differentiate the lower end of the categorical scale into more distinct categories (Edwards and Ostby 2015; NOAA/Storm Prediction Center 2020, 2021). The Marginal category was added before the Slight category, and the old Slight category was split into Slight and Enhanced Slight. However, given technological constraints, the Enhanced Slight was condensed into just Enhanced. Therefore, the current categorical scale consists of levels labeled Marginal, Slight, Enhanced, Moderate, and High.

Although the convective outlook has been evaluated based on the quality of the forecast (i.e., how close the observed weather matches the forecasted weather, e.g., Hitchens and Brooks 2012, 2017; Herman et al. 2018), fewer studies have assessed how different groups interpret and use the information provided in the convective outlook. Recently, studies have investigated severe weather information use more generally. For example, Ernst et al. (2018) investigated emergency manager use of severe weather information (including the convective outlook) and found their needs change as the event unfolds. As the event becomes closer in time, emergency managers start looking for more detailed information, including the expected likelihood of occurrence. Additional work has found similar results with the weather watch products. In their study of public response to severe weather information, Mason and Senkbeil (2015) found people prefer a product with more information about potential impacts and

Corresponding author: Makenzie J. Krocak, mjkrocak@ou.edu

DOI: 10.1175/WAF-D-21-0123.1

recommended response actions. Specific to the convective outlook, Williams et al. (2020) investigated public use of the outlook information from different sources (i.e., broadcast meteorologists reproducing the outlook information for their audience). They found the public does recognize when there are inconsistencies in the information, and they tend to use the outlook with the highest forecasted category (Williams et al. 2020). Finally, recent work by Ernst et al. (2021) found the general public misorders the categorical labels (most often switching Marginal with Slight and Enhanced with Moderate) and the colors to a lesser extent (most often ranking red as the highest level). While these works are valuable and reveal weaknesses of the current outlook format, they do not attempt to evaluate possible solutions to these problems. This work aims to take these studies one step further and evaluate how the general public interprets alternative types of labeling systems that are consistent with the outlook. We assess how presenting numeric levels (i.e., "Level 2 of 5") or probabilities (i.e., "a 5% chance") impacts an individual's concern and likelihood of response to a tornado threat when compared to presenting only the categorical labels (the current system used at the SPC). In addition, we also explore how different combinations of the current system with additional numeric information impact participant concern and response rankings.

Multiple organizations have called for the use of probabilities in weather forecasts to help improve their interpretation and utility (AMS Council 2008; National Research Council 2006), particularly since studies have shown people infer uncertainty information in forecasts even when it is not explicitly included (e.g., Morss et al. 2008). However, there is still much work to be done to understand the best way to communicate probabilistic forecasts such that people understand and use them appropriately. While some work in the weather domain has shown people interpret a range around single-value forecasts (e.g., a forecast high of 60° is interpreted as a forecast high of 58°–62°; Morss et al. 2008), other work outside of the weather domain shows a range of probability values still implies a point estimate, whether explicitly stated or not (Friedman and Zeckhauser 2014).

Further complicating this challenge is the use of verbal estimates of probabilities. The current SPC categorical scale comprises estimative words (e.g., Slight, Moderate, High), which are highly variable in their numeric interpretations (Wintle et al. 2019). Research dating back to the 1980s has shown even those who create forecasts (in this case, political forecasts) vary in their interpretation of estimative words. Researchers found high variability in verbal probability interpretations, especially when those words are put within additional context (Beyth-Marom 1982). Other work has found not only are interpretations of verbal probability estimates more variable when placed within context, but predictable biases can emerge in people's interpretations. For example, participants in a study of drug-side-effect probabilities estimated the chance of side effects occurring was lower when a leaflet about the drug information was provided to them (Fischer and Jungermann 1996). Finally, within the weather domain, words of estimative probability have been shown to have widely varying numeric interpretations. Some of these words are commonly used in forecasts, like "possible," "expected," and "chance" (Lenhardt et al. 2020).

In addition to the way in which forecasters frame probabilistic information, visualization has also been shown to impact public interpretation. Gerst et al. (2020) investigated long range probabilistic temperature and precipitation forecasts and found altering or removing the white space denoting "equal chance" on the graphics significantly improved user interpretation. Similarly, Ernst et al. (2021) investigated the word and color interpretation within the SPC convective outlook and found individuals incorrectly rank both by risk level. Regarding the word scale, Slight, Enhanced, and Moderate were often misordered. While Marginal was also shown to be switched with Slight, the High category was almost always placed at the top of the scale. The color scale was more often ordered correctly, although the purple or magenta color was sometimes switched with red at the high end of the scale. Beyond general misinterpretations, they also found gender, education, and numeracy had an impact on how people ranked the SPC categories. More numerate women with a higher level of educational attainment had the highest likelihood of correctly ranking the categorical words. These results are important to understand, as they suggest the weather information in the SPC outlook may not be as readily understood by some portions of the population.

While these previous studies highlight the known issues with interpretation of the SPC categorical scale, few provide or test any meaningful solutions to these issues. This work aims to analyze how highlighting information already produced for SPC severe weather forecasts, but not at the forefront of the product's presentation, impacts members of the public's perception of a generic tornado threat. We assess how using the current system of verbal labels, numeric levels, likelihood probabilities, and a combination of verbal and numeric information changes the way members of the public rate their concern about the forecast and their likelihood of taking action in response to the forecast. Finally, we also investigate how several different demographic characteristics influence interpretation, to identify whether these potential communication solutions avoid the interpretation gaps found with the categorical words in Ernst et al. (2021).

2. Data and methods

The data for this study comes from the 2020 Severe Weather and Society Survey (WX20). This annual survey is developed by the University of Oklahoma's Center for Risk and Crisis Management and distributed by Qualtrics. The participants of this survey are a demographically representative sample of 3000 U.S. adults aged 18 and over (see Table 1 and Krocak et al. (2020) for a more detailed description of the survey demographics). There are two types of questions on this survey; recurring questions testing concepts where longitudinal data are important, and one-time questions pertinent to specific problems or research questions (like those explored in this study).

TABLE 1. Demographic representativeness of the WX20 respondents.

	U. S. adult population (%)	Participants (%)
Gender		
Male	51.3	51.3
Female	48.7	48.7
Age		
18–24	12.0	12.0
25–34	18.0	18.0
35–44	16.3	16.3
45–54	16.4	16.4
55–64	16.7	16.7
65 and up	20.6	20.6
Ethnicity		
Hispanic	16.3	16.3
Non-Hispanic	83.7	83.7
Race		
White	77.9	77.7
Black or African American	13.0	13.0
Asian	5.9	5.9
Other race	3.2	3.4
NWS region		
Eastern	31.6	32.0
Southern	27.1	27.1
Central	20.7	20.7
Western	20.6	20.2

Questions for this study were designed to assess the concern and likelihood of taking action given a certain severe weather forecast. Respondents were told to imagine it was 8:00 a.m. on a Saturday morning when they received a tornado forecast indicating a certain risk. We chose to anchor respondents to this time because we wanted them to be thinking of a time frame when they were likely at home with few other obligations to attend to. While some people may work Saturday mornings, fewer people likely do so when compared to a weekday morning. The risk phrases were randomized such that each respondent only saw one forecast. Respondents were then asked to rate their concern about the forecast on a 0–100 scale (where 0 means not at all concerned and 100 means extremely concerned) and their likelihood of taking action (where 0 means not at all likely and 100 means extremely likely).

We test five different information combinations in this work. They include: (i) the current system (just the categorical names), (ii) numerical levels, (iii) probabilities, (iv) the current system and levels, and (v) the current system and probabilities. These combinations are all based on information the SPC currently provides (see Fig. 1 for an example of a convective outlook with this information). We chose to only focus on the middle three (of five) levels because they show the most variation in interpretation (Ernst et al. 2021). Given the five different information combinations and the three different levels, a total of 15 stratifications were tested (see Table 2 for a list). For illustration, respondents in the first stratification would have seen a set of text prompts and questions as follows:

- Forecasters often use a combination of phrases, scales, and probabilities to describe the risk of severe thunderstorms and tornadoes in an area. We want to know how you interpret these forecasts.
- To begin, imagine that it is next Saturday morning at 8:00 AM and you get a tornado forecast indicating that there is A SLIGHT RISK for tornadoes at your location that evening.
- On a scale from 0 to 100, where 0 means not at all concerned and 100 means extremely concerned, how concerned would you be if you were to get this forecast?
- On a scale from 0 to 100, where 0 means not at all likely and 100 means extremely likely, how likely is it that you would change your plans for the day if you were to get this forecast?

We compare median concern and likelihood ratings across the different forecast phrase conditions to assess if and how changing the forecast framing alters public interpretation of the forecast information. In addition to comparing forecast information, we assess differences in interpretation based on demographic differences. Previous work (Ernst et al. 2021) showed gender, education, and numeracy all influenced the likelihood of participants correctly ranking the SPC categorical labels. Therefore, we also assess how concern rating and likelihood of response is influenced by these demographic differences.

We measure gender and education level with multiple-choice questions. The education question asks participants to choose their highest level of completed education. Numeracy is measured using the Berlin Numeracy Test (Cokely et al. 2012) with additional questions adapted from Schwartz et al. (1997) to increase sensitivity to lower levels of numeracy.

3. Results

Our analysis indicates the framing of information in the SPC outlook can change the reported concern and likelihood someone will respond to the forecast. In the following figures, we display the median of the response distribution because the median is less influenced by large outliers than the mean. First, our control data measuring concern ratings for forecasts containing the current system of categorical labels show a misordering of the labels by the general public. The median concern rating for the Slight and Moderate categories are the same at 50 out of 100, while the median concern rating for Enhanced is over 60 (Fig. 2). This is problematic as Moderate is the second highest category (of the full five outlook categories) and often the highest risk category some locations ever see. A similar pattern shows up in the likelihood of response, where the median likelihood for Slight is 50 out of 100, followed by 60 out of 100 for Moderate, and 70 out of 100 for Enhanced (Fig. 3). Again, given a Moderate risk is dictated by a higher forecasted probability of occurrence than an Enhanced risk (NOAA/Storm Prediction Center 2020), it is concerning that respondents report being more likely to respond given an Enhanced risk than a Moderate risk. These results are very similar to Ernst et al. (2021).

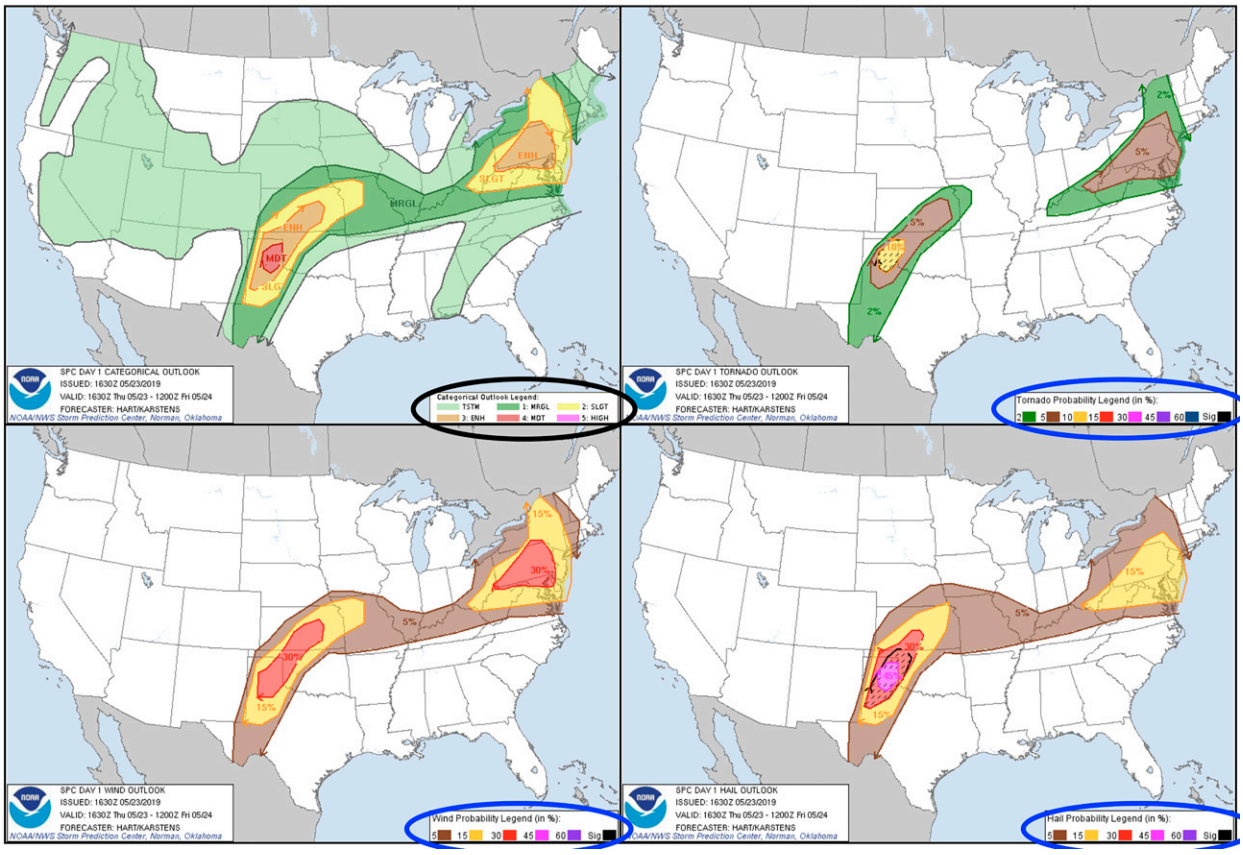

FIG. 1. The 1630 UTC four-panel convective outlook for 23 May 2019. The level and category legend is circled in black, the probability legends are circled in blue.

The numeric level information shows more consistent concern and response ratings when compared to the current system of SPC words. The median concern ratings for levels 2–4 show an increasing rating with each numeric level increase (medians are 50, 60, and 80 out of 100 for levels 2–4, respectively, Fig. 2). The ratings for the likelihood of responding given this numeric level information show identical medians to the concern ratings, again increasing with higher levels (Fig. 3).

The probabilistic forecast information also shows increasing medians with increasing probabilities, although the overall magnitudes of the concern and likelihood of response ratings are lower than the ratings for levels only. For example, the median concern ratings for a 5% forecast, 15% forecast, and 30% forecast are 20, 37.5, and 50 out of 100, respectively (Fig. 2). The likelihood of response ratings is similar at 25, 30, and 50 out of 100 (Fig. 3). Although these ratings are lower than the ratings for the numeric levels, they still increase with increasing probably (and therefore, risk of severe weather occurrence), unlike the current SPC label system.

Given the difference in interpretation between the current system and the numeric labels, we wanted to understand how combining the two types of information influenced concern and likelihood of response ratings. Generally, the median concern ratings for the combined framing (with both categorical and numeric labels) are very similar to the median concern ratings of just the categorical labels (or the current SPC labeling system). The medians for the current system plus levels

TABLE 2. Forecast phrase conditions used in the survey experiment. Each respondent was shown one of the phrases.

Categorical name	A SLIGHT RISK	AN ENHANCED RISK	A MODERATE RISK
Level	A LEVEL 2 of 5 RISK	A LEVEL 3 of 5 RISK	A LEVEL 4 of 5 RISK
Probability	A 5% CHANCE	A 15% CHANCE	A 30% CHANCE
Categorical name and level	A SLIGHT RISK (LEVEL 2 of 5)	AN ENHANCED RISK (LEVEL 3 of 5)	A MODERATE RISK (LEVEL 4 of 5)
Categorical name and probability	A SLIGHT RISK (5% CHANCE)	AN ENHANCED RISK (15% CHANCE)	A MODERATE RISK (30% CHANCE)

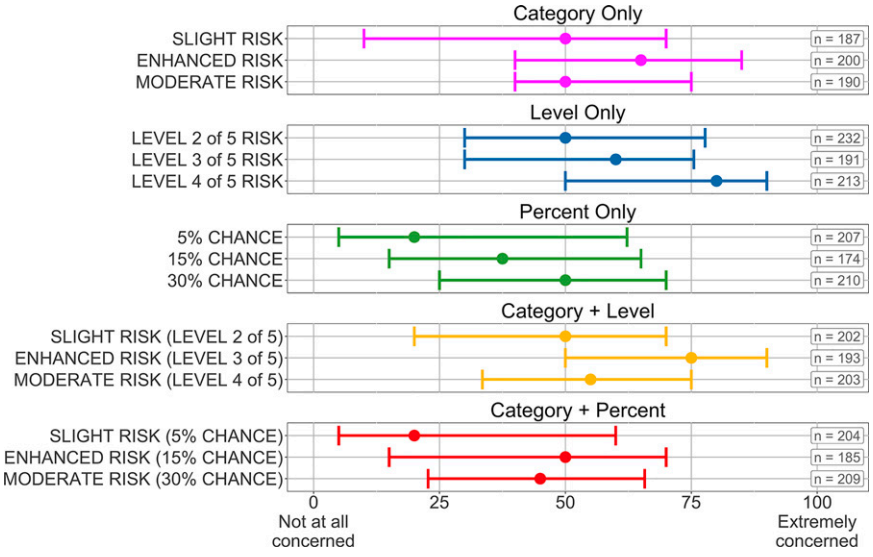

FIG. 2. Concern ratings for WX20 survey respondents with varied convective outlook forecast information. Dots are the medians, and error bars represent the interquartile range. The "Category Only" group is the current system used by the SPC.

format (Slight/level 2, Enhanced/level 3, and Moderate/level 4) are 50, 75, and 55 out of 100, respectively (Fig. 2). In fact, there is a slightly larger difference in median between the Enhanced rating and the Moderate rating (20 points versus 15 points) when you include the level information in the forecast versus the category alone. These results indicate respondents are anchoring to the categorical labels, which means simply adding level numbers to the current outlook system may not help people contextualize the risk any more than using just the current categorical information.

The addition of probabilistic information yields similar results. The median concern ratings are 20, 50, and 45 out of 100 (Fig. 2), which indicates while the Enhanced and Moderate levels are closer to each other in concern ratings than for the words alone, but Enhanced still had a slightly higher median concern rating than Moderate. However, magnitudes of the median values were lower with the probabilities than with the levels, similar to the results found when just the numeric or probabilistic information was presented. Very similar results were found with the likelihood of response ratings,

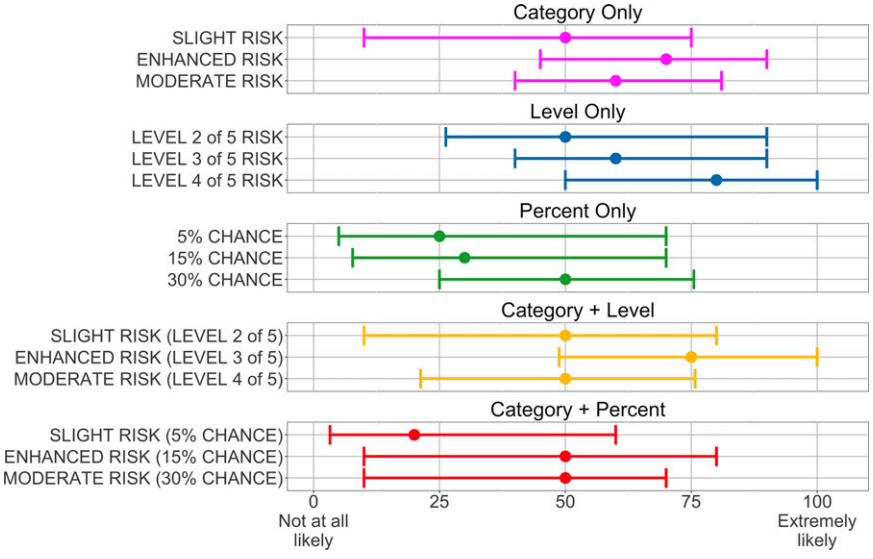

FIG. 3. Likelihood of response for WX20 survey respondents given different severe weather forecast information. Dots are the medians, and error bars represent the interquartile range. The "Category Only" group is the current system used by the SPC.

	Category only			Level Only			Percent Only			Category + Level			Category + Percent		
	SLT	ENH	MOD	2 of 5	3 of 5	4 of 5	5%	15%	30%	SLT 2 of 5	ENH 3 of 5	MOD 4 of 5	SLT 5%	ENH 15%	MOD 30%
All	50	65	50	50	60	80	20	37.5	50	50	75	55	20	50	45
Male	50	70	50	50	50	75	40	40	50	50	75	50	20	50	50
Female	50	55	50	50	70	80	15	25	50	50	72.5	60	10	45	35
Age 18-34	50	50	50	50	50	67.5	45	40	50	50	60	55	22	50	50
Age 35-54	50	72	68.5	50	64	80	47.5	50	50	55	77.5	50	45	50	50
Age 55+	25	55	50	47.5	60	80	10	25	40	50	70	55	10	25	30
Hispanic	50	70	60	50	70	77	20	40	55	55	55	62.5	25	50	50
Non-Hispanic	45	60	50	50	60	80	21.5	30	50	50	75	55	15	42.5	40
White	30	70	50	50	60	80	15	30	50	50	72.5	57	20	39	40
Non-White	60	50	60	50	50	77.5	50	50	46	50	75	50	20	54.5	50
High Numeracy	20	50	50	40	60	80	10	20	50	40	70	55	10	20	30
Low Numeracy	50	65	50	50	62	75	30	40	50	50	75	52.5	20	50	50
Eastern	50	70	60	50	60	80	50	40	50	50	75	55.5	10	70	45
Southern	50	60	60	50	62	80	10	30	50	50	75	50	45	50	37
Central	30	50	50	40	50	75	10	25	30	50	50	55	10	15	42.5
Western	25	70	50	55	63	80	22.5	60	50	45	75	55	15	39	50

FIG. 4. Median concern ratings by demographic group. Green text indicates an increasing rating with each category. Yellow text indicates a tie between two adjacent levels. Red text indicates a misordering of the levels.

as adding the level information to the categorical labels resulted in little change from the stand-alone categorical labels (Fig. 3). Median ratings were 50, 75, and 50 out of 100 for the Slight (level 2), Enhanced (level 3), and Moderate (level 4), respectively. Adding in the probabilistic information resulted in the Enhanced and Moderate levels showing similar ratings (50 out of 100 versus 20 out of 100 for the Slight category), but the overall magnitude of the ratings was again lower (Fig. 3).

In stratifying our data by different individual characteristics, we find numeric information is most often ordered correctly by most groups (Figs. 4 and 5). It is important to note for these results that some of the sample sizes [e.g., like respondents who live in the southern NWS region and who saw "Slight risk (level 2 of 5)"] are small and therefore the results may not be representative of the population. Overall, ethnicity, gender, and age do not show significant differences in interpretation, although younger respondents are less likely to rank the probabilities in ascending order (for both concern and likelihood of response).

Race, numeracy, and NWS region did show some differences in interpretation (Figs. 4 and 5). With regard to the numeric information alone, nonwhite respondents were more likely to rank adjacent levels similarly to each other, particularly the probabilistic values. Furthermore, nonwhite respondents tended to rank their concern ratings higher than white respondents, particularly for the categorical and probabilistic labels. Finally, nonwhite respondents seem to anchor to the words when the categorical labels are presented with the numeric information (as seen by the similar rankings for the categorical information alone and the category + level and category + percent labels). Similar results were found when stratifying by numeracy. Generally, low numeracy respondents had a higher concern rating than high numeracy respondents. This is particularly true for the categories + percent labels. The two groups had much more similar median ratings for the level and level + category labels. Additionally, the higher numerate respondents seemed to anchor more to the numeric information when the category + percent labels were presented together. For the highly numerate responses, the median concern ratings were 10, 19, and 28 (respectively). Conversely, the low numerate responses had median concern ratings of 19, 50, and 50, respectively. While the highly numerate responses increased similarly with each level increase, the

	Category only			Level Only			Percent Only			Category + Level			Category + Percent		
	SLT	ENH	MOD	2 of 5	3 of 5	4 of 5	5%	15%	30%	SLT 2 of 5	ENH 3 of 5	MOD 4 of 5	SLT 5%	ENH 15%	MOD 30%
All	50	70	60	50	60	80	25	30	50	50	75	50	20	50	50
Male	50	75	50	50	60	70	45	50	60	47.5	75	50	25	50	50
Female	50	60	69.5	67.5	62.5	80	20	17.5	50	50	75	55	10	50	40
Age 18-34	50	50	50	50	50	67.5	45	40	50	50	60	55	22	50	50
Age 35-54	50	72	68.5	50	64	80	47.5	50	50	55	77.5	50	45	50	50
Age 55+	25	55	50	47.5	60	80	10	25	40	50	70	55	10	25	30
Hispanic	50	70	60	50	70	77	20	40	55	55	55	62.5	25	50	50
Non-Hispanic	45	60	50	50	60	80	21.5	30	50	50	75	55	15	42.5	40
White	30	70	50	50	60	80	15	30	50	50	72.5	57	20	39	40
Non-White	60	50	60	50	50	77.5	50	50	46	50	75	50	20	54.5	50
High Numeracy	20	50	50	40	60	80	10	20	50	40	70	55	10	20	30
Low Numeracy	50	65	50	50	62	75	30	40	50	50	75	52.5	20	50	50
Eastern	50	70	60	50	60	80	50	40	50	50	75	55.5	10	70	45
Southern	50	60	60	50	62	80	10	30	50	50	75	50	45	50	37
Central	30	50	50	40	50	75	10	25	30	50	50	55	10	15	42.5
Western	25	70	50	55	63	80	22.5	60	50	45	75	55	15	39	50

FIG. 5. Median likelihood of response ratings by demographic group. Green text indicates an increasing rating with each category. Yellow text indicates a tie between two adjacent levels. Red text indicates a misordering of the levels.

low numerate responses had equal median concern ratings for the Enhanced and Moderate levels (Fig. 4). However, this result did not appear when the category + level labels were presented together. High and low numerate respondents had very similar concern ratings for this set of labels (Fig. 4). Very similar patterns are seen in the likelihood of response rankings (Fig. 5). Finally, NWS region did show some differences in interpretation, which is likely due to respondents' experience with severe weather (and subsequently the convective outlook). While the central and southern region respondents ranked the levels and probabilities in ascending order, eastern and western respondents ranked the 15% category higher than the 30% category (Fig. 4). Overall, the levels alone produced the most consistent rankings, although it is important to keep sample size in mind when interpreting the stratified medians.

4. Discussion

Given the increasing use of the SPC convective outlook as a source of information for upcoming severe weather events, it is important to understand how the public interprets the information provided in the product. We investigate how concern about and intent to respond to severe weather varies given different combinations of numeric and verbal forecast information. First and foremost, we find strong support for previous conclusions that public interpretation does not align with the current SPC categorical label scale. The Enhanced label is consistently ranked equal to or above the Moderate label in concern and likelihood of response. This is concerning because Moderate is the second highest risk level, and often the highest severe weather risk many locations will ever see given the rarity of High risks.

Next, we analyze the concern and likelihood of response ratings for different types of numeric information (levels and percentages). We find unlike the current system, the numeric information is much less likely to be misordered. However, the level labels consistently produce higher concern and response ratings compared to the percent labels. This difference in interpretation is important to consider: is it better to have concern ratings aligned more with the actual probability of experiencing the hazard even though this leads to much

lower likelihood of response ratings, or vice versa? These are important questions that will require careful consideration before policy decisions are made.

When the current SPC label system and numeric information are combined, we find respondents generally anchor to the current system. Therefore, the tendency to swap the Enhanced and Moderate levels persisted even when additional numeric information was presented alongside the categorical labels. While it may be tempting to combine the information, this work shows that may not have the intended effect of aiding interpretation to correctly order the categorical labels.

Finally, we also investigate how differences in demographic characteristics influence interpretation of different forecast information. Although we do not find significant differences across gender, age, or ethnicity, we do find some differences across race and numeracy ratings. Most importantly, we find nonwhite and less numerate respondents anchor to the categorical labels when the current SPC labels are presented with the percentages. Essentially, this means adding probabilistic information to the current system serves to help white and more numerate people correctly order the forecast levels, without much impact for nonwhite and less numerate people.

While our results and synthesis herein are tailored to the severe weather domain, we believe they are relevant to other risk communication practices within the weather community. Further work should investigate how these results compare when the full SPC scale is used. Given that we wanted to analyze multiple types of information (e.g., words, numeric levels, percentages, and combinations of the three types), we were unable to assess the full scale due to sample size constraints. It would also be worthwhile to investigate these interpretation challenges and potential solutions with emergency managers and other stakeholders prior to any policy changes. Finally, our analysis was done without the use of accompanying graphics or maps. Future work should investigate if and how the addition of visuals changes the interpretation of severe weather information since the two types of information are often presented together.

Since the initial fielding of this survey experiment, additional work related to the communication of convective outlook information has progressed. For example, researchers have been working with NWS and SPC leadership to construct a literature review of the work that has been done in relation to the outlook to help direct future research efforts (Krocak et al. 2021a). Given the interest in convective outlook communication, future research endeavors should aim to evaluate the full scale of alternative labeling systems and how those systems may serve non-English speaking and other historically underserved populations. In addition to this literature review effort, there is also a new research project beginning spring 2022 aimed at investigating how additional convective outlook information (like intensity information and probabilistic information) may be presented most effectively to the general public (Krocak et al. 2021b). Finally, a recent publication in the tropical cyclone domain (Rosen et al. 2021) found that the inclusion of probabilistic information in tropical cyclone forecasts increased survey respondents' perceived reliability of the forecast, further pointing to the potential value of including numeric information in forecasts.

The challenge of presenting weather forecast information that is easily interpreted and immediately useful is heightened when a large portion of the population has not experienced the hazard being forecasted. Severe weather is rare but highly impactful, increasing the need for useful information well in advance of the event. While the convective outlook contains this much-needed information, numerous studies have uncovered challenges with the framing of the information (e.g., Ernst et al. 2021; Williams et al. 2021). In particular, the current categorical labels are often misinterpreted and misordered. Barring a complete overhaul of the system (which would impact policies in place at TV stations, emergency management jurisdictions, and school districts, just to name a few), there may be some straightforward changes that could improve interpretation by the general public. We have shown that eliminating the verbal categorical labels and providing the numeric levels decreases the tendency for Enhanced and Moderate to be misinterpreted. These findings align well with previous work (e.g., Lenhardt et al. 2020) which highlighted challenges associated with verbal information interpretation. Therefore, one relatively straightforward change might be to only show numeric information on public facing websites and social media. Before any change is made, the policies and procedures for TV stations, emergency manages, school districts, etc., must be considered as they likely require legally mandated alterations. These changes certainly will not solve all interpretation challenges, but could be an important first step to investigate how interpretation changes without completely changing the system currently in place.

Acknowledgments. Data collection for this project was funded by the OU Office of the Vice President for Research. Data analysis was funded by National Oceanic and Atmospheric Administration Project OAR-USWRP-R2O, "FACETs Probability of What? Understanding and Conveying Uncertainty through Probabilistic Hazard Services."

Data availability statement. The survey data are available at https://dataverse.harvard.edu/dataverse/wxsurvey.

REFERENCES

Beyth-Marom, R., 1982: How probable is probable? A numerical translation of verbal probability expressions. *J. Forecasting*, **1**, 257–269, https://doi.org/10.1002/for.3980010305.

Cokely, E. T., M. Galesic, E. Schulz, S. Ghazal, and R. Garcia-Retamero, 2012: Measuring risk literacy: The Berlin Numeracy Test. *Judgement Decis. Making*, **7**, 25–47.

Corfidi, S. F., 1999: The birth and early years of the Storm Prediction Center. *Wea. Forecasting*, **14**, 507–525, https://doi.org/10.1175/1520-0434(1999)014<0507:TBAEYO>2.0.CO;2.

AMS Council, 2008: Enhancing weather information with probability forecasts: An information statement of the American Meteorological Society. *Bull. Amer. Meteor. Soc.*, **89**, 1049–1053, https://doi.org/10.1175/1520-0477-89.7.1041.

Edwards, R., and F. Ostby, 2015: Time line of SELS and SPC. Storm Prediction Center. NOAA, accessed 28 February 2020, https://www.spc.noaa.gov/history/timeline.html.

Ernst, S., D. LaDue, and A. Gerard, 2018: Understanding emergency manager forecast use in severe weather events. *J. Oper. Meteor.*, **6**, 95–105, https://doi.org/10.15191/nwajom.2018.0609.

——, J. T. Ripberger, M. J. Krocak, H. Jenkins-Smith, and C. Silva, 2021: Colorful language: Investigating public interpretation of the Storm Prediction Center convective outlook. *Wea. Forecasting*, **36**, 1785–1797, https://doi.org/10.1175/WAF-D-21-0001.1.

Fischer, K., and H. Jungermann, 1996: Rarely occurring headaches and rarely occurring blindness: Is rarely = rarely? The meaning of verbal frequentistic labels in specific medical contexts. *J. Behav. Decis. 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. *Intell. Natl. Secur.*, **30**, 77–99, https://doi.org/10.1080/02684527.2014.885202.

Gerst, M. D., and Coauthors, 2020: Using visualization science to improve expert and public understanding of probabilistic temperature and precipitation outlooks. *Wea. Climate Soc.*, **12**, 117–133, https://doi.org/10.1175/WCAS-D-18-0094.1.

Herman, G. R., E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184, https://doi.org/10.1175/WAF-D-17-0104.1.

Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's Day 1 convective outlooks. *Wea. Forecasting*, **27**, 1580–1585, https://doi.org/10.1175/WAF-D-12-00061.1.

——, and ——, 2017: Determining criteria for missed events to evaluate significant severe convective outlooks. *Wea. Forecasting*, **32**, 1321–1328, https://doi.org/10.1175/WAF-D-16-0170.1.

Krocak, M. J., J. T. Ripberger, C. Silva, H. Jenkins-Smith, S. Ernst, A. Bell, and J. Allan, 2020: Measuring change: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous United States. Harvard Dataverse, accessed 10 November 2021, https://doi.org/10.7910/DVN/EWOCUA.

——, S. Ernst, J. T. Ripberger, C. Williams, J. Trujillo-Falcón, B. T. Gallo, and P. Marsh, 2021a: The National Weather Service Storm Prediction Center's convective outlook: Conclusions from past research and recommendations for future development. NOAA, 16 pp., https://www.spc.noaa.gov/publications/krocak/otlk-res.pdf.

——, J. T. Ripberger, C. L. Silva, and B. T. Gallo, 2021b: Enhancing the Storm Prediction Center's convective outlook with continuous probabilities and conditional intensity forecasts. NOAA Funding Award Project Narrative Award NA21-OAR4590173, 46 pp.

Lenhardt, E. D., R. N. Cross, M. J. Krocak, J. T. Ripberger, S. R. Ernst, C. L. Silva, and H. C. Jenkins-Smith, 2020: 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. *J. Oper Meteor.*, **8**, 64–78, https://doi.org/10.15191/nwajom.2020.0805.

Mason, J. B., and J. C. Senkbeil, 2015: A tornado watch scale to improve public response. *Wea. Climate Soc.*, **7**, 146–158, https://doi.org/10.1175/WCAS-D-14-00035.1.

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 Research Council, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. National Academies Press, 124 pp.

NOAA/Storm Prediction Center, 2020: SPC products. NOAA/Storm Prediction Center, accessed 1 December 2020, https://www.spc.noaa.gov/misc/about.html.

——, 2021: National Weather Service instruction. Tech. Rep. 10-512, NOAA/Storm Prediction Center, 73 pp., https://www.nws.noaa.gov/directives/sym/pd01005012curr.pdf.

Rosen, Z., M. J. Krocak, J. T. Ripberger, R. Cross, E. Lenhardt, C. L. Silva, and H. C. Jenkins-Smith, 2021: Communicating probability information in hurricane forecasts: Assessing statements that forecasters use on social media and implications for public assessments of reliability. *J. Oper. Meteor.*, **9**, 89–101, https://doi.org/10.15191/nwajom.2021.0907.

Schwartz, L. M., S. Woloshin, W. C. Black, and H. G. Welch, 1997: The role of numeracy in understanding the benefit of screening mammography. *Ann. Intern. Med.*, **127**, 966–972, https://doi.org/10.7326/0003-4819-127-11-199712010-00003.

Williams, C. A., A. J. Grundstein, and J. So, 2020: Should severe weather graphics wear uniforms? Understanding the effects of inconsistent convective outlook graphics on members of the public. *15th Symp. on Societal Applications: Policy, Research, and Practice*, Boston, MA, Amer. Meteor. Soc., 11A.1, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/365011.

——, ——, and ——, 2021: What is being enhanced?: An examination of the Storm Prediction Center's risk category system among members of the public. *16th Symp. on Societal Applications: Policy, Research, and Practice*, virtual, Amer. Meteor. Soc., J4.1, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/380671.

Wintle, B. C., H. Fraser, B. C. Wills, A. E. Nicholson, and F. Fidler, 2019: Verbal probabilities: Very likely to be somewhat more confusing than numbers. *PLOS ONE*, **14**, e0213522, https://doi.org/10.1371/journal.pone.0213522.