Meteorological Applications

RESEARCH ARTICLE OPEN ACCESS

Developing Experimental Probabilistic Intensity Forecast Products for Landfalling Tropical Cyclones

Robert Eicher[1] | Daniel J. Halperin[1] | Benjamin C. Trabing[2,3] | Derek Lane[4] | Deanna Sellnow[5] | Timothy Sellnow[5] | Madison Croker[1]

[1]Embry-Riddle Aeronautical University (ERAU), Daytona Beach, Florida, USA | [2]University Corporation for Atmospheric Research, Cooperative Programs for the Advancement of Earth System Science, Boulder, Colorado, USA | [3]National Hurricane Center, Miami, Florida, USA | [4]University of Kentucky (UKY), Lexington, Kentucky, USA | [5]Clemson University, Clemson, South Carolina, USA

Correspondence: Benjamin C. Trabing (ben.trabing@noaa.gov)

Received: 4 April 2025 | **Revised:** 8 July 2025 | **Accepted:** 5 August 2025

Funding: This work was supported by University Corporation for Atmospheric Research Subaward No. SUBAWD002943, an NWS Partners Project. Funding for the qualitative analysis that followed and the survey in Study 2 was provided by NOAA Award NA22OAR4590185. Development of the LDP was funded by HFIP Award NA19OAR4320073.

Keywords: probabilstic forecasting | risk communication | tropical cyclones

ABSTRACT

An increasing body of evidence indicates that publics want more probabilistic information included in their weather forecasts. However, more guidance on incorporating probability information into weather risk communication is needed. The National Hurricane Center (NHC) recently developed prototype forecast graphics that include probabilistic values of intensity at landfall when landfall is possible. The goal of this research was to develop those prototypes into a forecast product that expresses technical uncertainty in an intensity forecast in a manner that is understandable and effective to various publics. In Study 1, an online survey among Florida residents was conducted. Quantitative analysis of the survey data showed few significant differences between the prototypes and the currently operational forecast track graphic, commonly referred to as the cone of uncertainty (COU). Analysis of the responses to open-ended questions in the survey and feedback from focus group participants consisting of NHC partners working in hurricane-prone areas guided revisions to improve the prototypes. In Study 2, the modified prototypes produced an improvement in understanding of certain aspects of the intensity forecast. Promisingly, most people surveyed preferred the additional probabilistic information in the prototypes to the status quo COU message. In fact, nearly 90% of respondents indicated that they preferred at least some percentage values in their weather forecasts as opposed to forecasts with words only. This suggests that further development of a probabilistic landfall intensity product might be warranted.

1 | Introduction

The National Hurricane Center's (NHC's) tropical cyclone (TC) forecast track graphic, commonly referred to as the cone of uncertainty (COU), was introduced to the NHC website in 2002 (ERG 2019). It has since become the most viewed graphic on the NHC website and is widely distributed through both traditional news and social media (Millet et al. 2020). The cone represents the probable track of the center of a TC, and the size of the cone is calculated from official forecast position errors at each forecast time over the previous 5 years (National Hurricane Center 2024). Put simply, the graphic illustrates what Gustafson and Rice (2020) refer to as *technical* uncertainty in the track forecast. Technical uncertainty is defined as a prediction limited by observation error and modeling assumptions (Gustafson and Rice 2020).

Although the NHC has made substantial improvements in TC track forecasting during the past half century (Cangialosi et al. 2020), accurately predicting TC intensity[1] continues to be a challenge (Gall et al. 2013), especially when rapid intensification[2] occurs (Trabing and Bell 2020). The NHC's TC intensity forecast errors were fairly consistent from the 1970s to the early 2000s, averaging around 15 knots of error for a 48-h forecast, but have recently shown some improvement (Cangialosi et al. 2020; Trabing and Bell 2020). Despite the known uncertainty in intensity forecasts, no operational product exists that quantifies the technical uncertainty in TC intensity forecasts (Bhatia and Nolan 2015). The NHC's hurricane specialists can only express confidence in an official forecast to an inexact degree, for example "there is some chance," through the publicly available forecast discussion and key messages products (National Hurricane Center 2024). However, the forecast discussion is meteorological in nature and interpreting such information requires preexisting scientific knowledge that end-users may not possess (Drake 2012). Additionally, research has shown that vague expressions of uncertainty can be interpreted variably and may lead to misunderstandings (Wallsten et al. 1986; Windschitl and Weber 1999).

Sellnow and Sellnow (2019) argue that effective risk communication includes admitting both what is known and unknown. Conveying uncertainty information in weather forecasts can be beneficial. Sellnow et al. (2002) found that "inappropriately unequivocal predictions" during the 1997 Red River Valley Flood in Minnesota "ultimately diminished the effectiveness of the region's crisis communication and planning" (Sellnow et al. 2002). Joslyn and LeClerc (2012) discovered that "uncertainty information improved decision quality overall and increases trust in the forecast" (p. 126). That study focused specifically on road maintenance in icy conditions, but the authors believe it has implications on severe weather warnings and "other domains" (Joslyn and LeClerc 2012). Bica et al. (2019) focused specifically on how to communicate uncertainty regarding hurricanes. They discovered critical "opportunities for the innovation of new information products to support risk communication" and "risk representations should convey uncertainty as appropriate in understandable, meaningful ways so that people can make best use of the information in interpreting risk" (Bica et al. 2019). Although this study focused specifically on Twitter (as it was known then), it points to the need for additional work in this area.

A growing body of social science research indicates that publics want more numerical information in their weather forecasts. In a review of uncertainty literature related to climate change and COVID-19, Halvey (2020) concluded, "[q]uantitative expressions of uncertainty, such as a numerical range surrounding an estimate or a percent likelihood, maximize the clarity of uncertainty expression and either maintain or increase an audience's level of trust in the data's source." Ripberger et al. (2022) conducted a systematic review of relevant research regarding the effective communication of probabilistic information in weather forecasts, motivated in part by their notion that probabilities are "notoriously difficult to communicate effectively to lay audiences." They discovered that although verbal expressions are commonly used to express uncertainty in the weather enterprise (e.g., "unlikely" or "good chance"), there is strong evidence to support the inclusion of a numeric "translation" of uncertainty that is intelligibly presented. Similarly, Rosen et al. (2021) found that 75.4% of people in hurricane prone areas prefer both words and numbers in hurricane forecasts. Moreover, Rosen et al. (2021) found that people perceive forecast messages using only words as less reliable than messages that use only numbers or both words and numbers.

This research suggests that publics might be better served if the NHC were to communicate technical uncertainty in TC forecasts using numerical probabilities. However, Durbach and Stewart (2011) found that messages with probability distributions tend to "overload" subjects with information, leading to poor decisions. Messages with three-point approximations and quantiles resulted in better decision making (Durbach and Stewart 2011). They concluded that concise messages and formats are easier to use. Essentially, Durbach and Stewart (2011) emphasize the importance of concise presentation of uncertainty information as it leads to better decision making.

The goal of this research is to develop a TC intensity forecast product that expresses the technical uncertainty in a manner that is both understandable and effective. The NHC has recently developed a few prototype forecast graphics designed to convey intensity forecast uncertainty using probabilistic information when landfall is possible. The broad question being, will the prototypes from the NHC effectively communicate the forecast uncertainty? This research consists of two studies using an iterative process of online surveys and focus groups with multiple versions of prototype graphics to answer the following research questions (RQs):

RQ1. How, if at all, do the prototypes differ from the status quo in terms of message comprehension?

RQ2. How, if at all, do the prototypes differ from the status quo in terms of perceived message effectiveness?

RQ3. How, if at all, do the prototypes differ from the status quo in terms of risk perception?

RQ4. What observations will various publics make from the prototypes and what information is most salient?

Prior experience with TCs has been shown to have some impact on risk perception and risk salience (Bostrom et al. 2018) and on TC forecast comprehension (Eicher et al. 2023). It is also well documented that residents outside of the center of the COU do not internalize the risk (e.g., Millet et al. 2020). It is therefore reasonable to ask:

RQ5. Which demographic differences (i.e., prior experience, length of residency in a hurricane prone state, or region within the state) significantly impact the results?

This research is a starting point for formatting and designing a product that details the uncertainty associated with the forecast intensity of landfalling TCs. Since no product like this currently exists, we chose to use the operational COU graphic

as a baseline, or status quo. Prior studies have evaluated how the COU is (mis)interpreted (e.g., Broad et al. 2007; Bostrom et al. 2018) and have tested potential modifications to the forecast product (e.g., Ruginski et al. 2016; Millet et al. 2024). These studies focused on communicating the uncertainty in forecast track information whereas this study focuses on communicating intensity information specifically at landfall. Therefore, the goal of this study is not to suggest modifications to or a replacement for the COU. Instead, here we evaluate prototypes that could provide information that adds to the existing suite of operational forecast products. While the COU's main purpose is to provide the most likely track of a TC's center, the COU graphic also includes current and categorical forecast intensity information (i.e., tropical depression, tropical storm, hurricane, major hurricane) at each forecast point along the forecast track in the form of symbols such as "S" for tropical storm and "H" for hurricane (National Hurricane Center 2024). This categorical, deterministic format is much different than the probabilistic information included in the prototypes used in this study, but the COU does provide a well-known, long-existing operational forecast product that can be used as a baseline or reference point at which to compare the prototypes.

2 | Study 1: Initial Prototypes

2.1 | Methods

Huang et al. (2016) completed a meta-analysis of 38 studies involving actual responses to hurricane warnings and 11 studies involving expected responses to hypothetical hurricane scenarios conducted since 1991. They found "the effect sizes from actual hurricane evacuation studies are similar to those from studies of hypothetical hurricane scenarios for 10 of 17 variables that were examined," which suggests "laboratory and internet experiments could be used to examine people's cognitive processing of different types of hurricane warning messages" (Huang et al. 2016). However, methodological triangulation, the use of more than one approach to a research question, provides a more comprehensive picture of the results than any single method could do alone (Heale and Forbes 2013). NOAA recently began employing methodological triangulation to improve other NHC products (Eosco and Williamsberg 2023), and the present study uses a combination of quantitative and qualitative data to assess the comprehension, effectiveness, and risk perception associated with the prototype forecast graphics.

To test the research questions, we began with an online survey developed using Qualtrics survey management software. Participants were recruited through a Qualtrics proprietary panel. Once participants accessed the online survey through Qualtrics and provided consent to participate, the survey system randomly assigned individuals to one of the prototype forecast graphics or the status quo COU graphic. After viewing the graphic, participants were asked a series of questions measuring comprehension, effectiveness, and risk perception. Participants were able to review the graphic as many times as they desired while answering the questions. Each variable's measuring instrument (i.e., comprehension, effectiveness, risk perception, and risk salience) mostly contained closed-ended questions but also included a few open-ended questions to elicit more in-depth responses from participants about the graphic. The final questions focused on risk salience and demographics.

2.2 | Survey Participants

Respondents had to be at least 18 years of age and residing in the state of Florida to participate in the survey. Florida was chosen as it is the state most prone to TCs (Jarell et al. 2021). A total of 1161 Florida residents participated in this study. Extreme values in response time, the top and bottom 5%, were removed from the dataset, leaving only participants who completed the survey in a reasonable amount of time between 3 and 30 min. The filtered total yielded 1058 participants, 67.9% of whom identified as female and 31.0% as male, ranging in age from 18 to 90 years ($M = 46.32$, SD = 18.97). The participants identified themselves as 66% Caucasian, 13.8% African American, 11.2% Hispanic or Latinx, 2.9% mixed race, and 1.2% Asian. Chi-square tests indicate no significant difference in any of the demographics across the conditions. Perhaps most importantly, given prior research related to TC experience, the length of time participants reported living in Florida was nearly equally distributed across all conditions (Table 1).

TABLE 1 | Crosstabulation of condition by duration of residence in Florida among survey participant (χ^2 (5, 1058) = 16.88, $p = 0.661$).

	Condition					
How long have you lived in Florida?	**Status quo**	**Prototype M1T1**	**Prototype M2T2**	**Prototype M1T2**	**Prototype M2T1**	**Total**
0–1 year	4	6	13	8	8	39
2–5 years	22	20	18	26	21	107
6–10 years	21	25	21	23	29	119
11–15 years	25	16	24	11	18	94
16–20 years	19	21	25	25	20	110
> 20 years	120	110	122	115	122	589
Total	211	198	223	208	218	1058

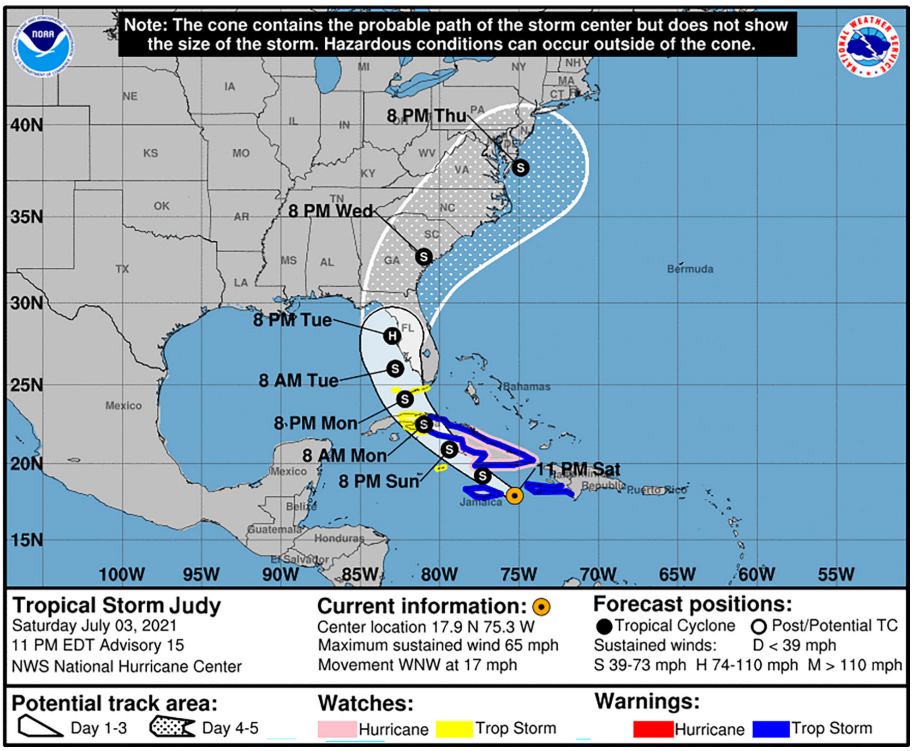


FIGURE 1 | The status quo COU graphic.

2.3 | Stimuli

For this study, four prototype forecast graphics were created based on NHC advisory 15 for TC Elsa from July 3, 2021, at 11 PM EDT. At that point in the storm's development, the actual COU for Elsa covered almost all of Florida, making it relevant to all Florida residents and therefore relevant to the survey participants (Figure 1). In all graphics, the name of the TC was changed to "Judy" to avoid confusion with the real system. Participants were given the following introduction just before viewing the graphic:

> You are about to see a forecast graphic for a hypothetical tropical cyclone named "Judy." Judy is not a real tropical cyclone but is based on a real scenario. In this scenario, Judy is currently a tropical storm located about 140 miles east of Kingston, Jamaica, with maximum sustained winds of 70 mph. Judy is currently moving toward the west-northwest at 23 mph but is expected to turn north in the direction of Florida.

That storm information closely matches the real Elsa scenario on July 3, 2021.

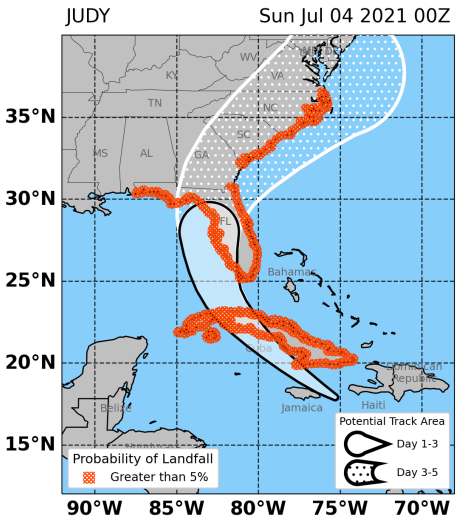
Intensity probabilities for the prototype graphics were generated by the landfall distribution product (LDP) using the case of TC Elsa (Trabing et al. 2023). The LDP uses the Monte Carlo Wind Speed Probability Model (WSP; DeMaria et al. 2013) which is a statistical ensemble based on the error characteristics of NHC forecasts and the spread of several track forecast models. Generating TC intensity probabilities is challenging as the track, intensity, and wind structure of a TC are not independent, especially near land (DeMaria et al. 2009). A slight shift in the forecast track could cause dramatic changes in the intensity and structure of a TC if the center of the storm happens to move over land. The LDP outputs probabilistic intensity estimates as well as estimates of the most likely and strongest reasonable intensity, defined as the 10% exceedance value, at landfall (Trabing et al. 2023).

Each prototype graphic included a map showing the forecast cone along with the probability of landfall at various regions with the maps varying by color scheme based on that probability. "Map 1" (M1) outlined any coastline that had a greater than 5% chance of experiencing landfall in red (Figures 2 and 3). The 5% value is used as a lower bound in other NHC products and was chosen for consistency. On "Map 2" (M2), the landfall probabilities were color coded as yellow, orange, and red for categories of low, medium, and high, respectively (Figures 4 and 5). Each prototype graphic also included a table that highlighted the strongest reasonable intensity that could be expected at landfall, but the tables varied by how much additional information was included. "Table 1" (T1) contained a single additional column of landfall probabilities (Figures 2 and 5). "Table 2" (T2) contained all the information in T1 plus an additional four columns of probabilities of various intensities described by the Saffir Simpson Hurricane Wind Speed Scale (Schott 2012; Figures 3 and 4). The prototypes are named according to their combination of map and table (e.g., "M1T2" is Map 1 combined with Table 2, Figure 3). The control condition was the status quo COU from TC Elsa advisory 15 (Figure 1).

2.4 | Survey Measures

2.4.1 | Comprehension (RQ1)

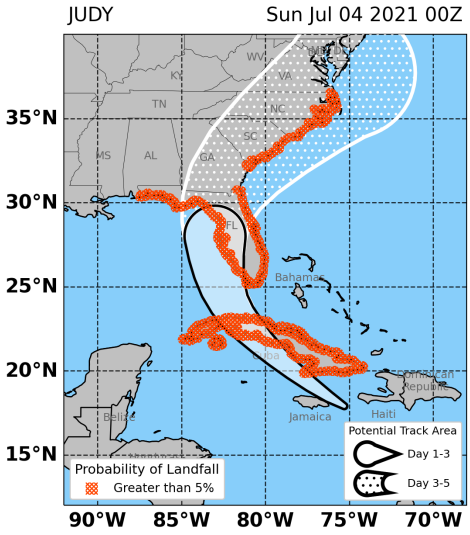
Comprehension was measured based on what the participants reported understanding (perceived comprehension) and what they actually understood (actual comprehension). These measures



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Strongest Reasonable Landfall Intensity
CUBA	96%	Cat 1 (80 mph)
FL Pan	41%	Cat 2 (100 mph)
West FL	60%	Cat 2 (100 mph)
East FL	9%	Cat 1 (90 mph)
SC	12%	Cat 1 (75 mph)
NC	8%	Cat 1 (75 mph)

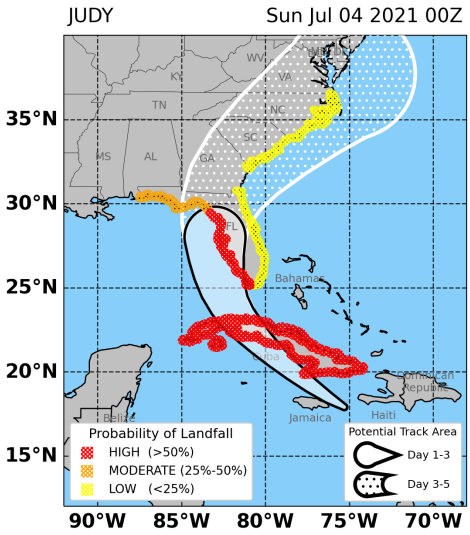
FIGURE 2 | Prototype M1T1.



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Tropical Depression or Weaker (<39 mph)	Tropical Storm (39-73 mph)	Hurricane Cat 1-5 (>74 mph)	Major Hurricane Cat 3-5 (>110 mph)	Strongest Reasonable Landfall Intensity
CUBA	96%	7%	72%	21%	<2%	Cat 1 (80 mph)
FL Pan	41%	20%	36%	44%	2%	Cat 2 (100 mph)
West FL	60%	21%	38%	41%	2%	Cat 2 (100 mph)
East FL	9%	25%	37%	38%	2%	Cat 1 (90 mph)
SC	12%	28%	64%	8%	<2%	Cat 1 (75 mph)
NC	8%	32%	60%	8%	<2%	Cat 1 (75 mph)

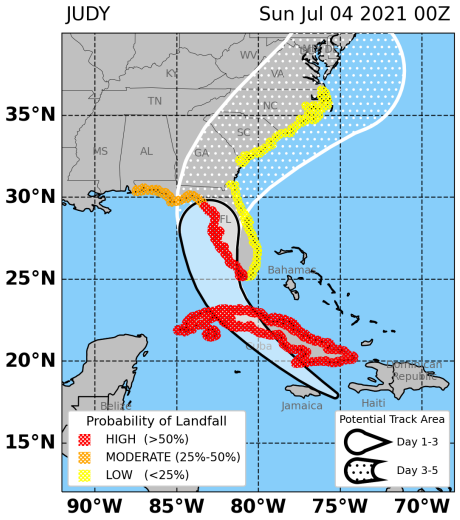
FIGURE 3 | Prototype M1T2.



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Tropical Depression or Weaker (<39 mph)	Tropical Storm (39-73 mph)	Hurricane Cat 1-5 (>74 mph)	Major Hurricane Cat 3-5 (>110 mph)	Strongest Reasonable Landfall Intensity
CUBA	96%	7%	72%	21%	<2%	Cat 1 (80 mph)
FL Pan	41%	20%	36%	44%	2%	Cat 2 (100 mph)
West FL	60%	21%	38%	41%	2%	Cat 2 (100 mph)
East FL	9%	25%	37%	38%	2%	Cat 1 (90 mph)
SC	12%	28%	64%	8%	<2%	Cat 1 (75 mph)
NC	8%	32%	60%	8%	<2%	Cat 1 (75 mph)

FIGURE 4 | Prototype M2T2.



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Strongest Reasonable Landfall Intensity
CUBA	96%	Cat 1 (80 mph)
FL Pan	41%	Cat 2 (100 mph)
West FL	60%	Cat 2 (100 mph)
East FL	9%	Cat 1 (90 mph)
SC	12%	Cat 1 (75 mph)
NC	8%	Cat 1 (75 mph)

FIGURE 5 | Prototype M2T1.

TABLE 2 | Perceived comprehension (PC) scores.

	Status quo	Prototype M1T1	Prototype M2T2	Prototype M1T2	Prototype M2T1
M	3.98_{a}	$3.81_{a,b}$	3.60_{b}	3.61_{b}	3.68_{b}
SD	0.8	0.92	0.98	0.9	0.9

Note: Items ranged from 1 = strongly disagree to 5 = strongly agree. Means with different subscripts differ at $p < 0.005$ using Tukey HSD post hoc comparisons.

were developed following several studies that measured both perceived and actual understanding of TC terminology (e.g., Whitmer and Sims 2021; Lindner et al. 2019; Eicher et al. 2023). Perceived comprehension (PC) included seven items on a five-point scale ranging from strongly disagree (1) to strongly agree (5). Sample items included "I understood most of the content in the forecast graphic" and "I had a hard time figuring out what the weather report was communicating" (reverse coded). A PC index was created by averaging the scores (Cronbach's $\alpha = 0.90$, $M = 3.73$, SD = 0.91).

Actual comprehension (AC) was measured via a four-question multiple choice test. The questions were worded such that the correct answers could be deduced from the graphic assuming it was properly understood and interpreted. Sample test questions included "Within the United States, Judy is most likely to make landfall in," and "If Judy were to make landfall in West Florida (e.g., Tampa Bay area), the strongest intensity that could reasonably be expected at landfall is?" Participants were awarded 1 point for each correct answer and an AC index was created by summing the scores.

2.4.2 | Effectiveness (RQ2)

Perceived message effectiveness (PME) was measured using a 9-item scale very similar to that used by Sellnow et al. (2015) which was designed based on the recommendations proposed by Noar et al. (2010). Each PME item was measured on a five-point scale ranging from strongly disagree (1) to strongly agree (5). Sample items included "This graphic would catch my attention," "The graphic is relevant to me," and "This graphic would make me less likely to check for updated forecasts" (reverse coded). A PME index was created by averaging the scores (Cronbach's $\alpha = 0.86$, $M = 3.98$, SD = 0.71).

2.4.3 | Risk Perception (RQ3)

Risk perception (RP) was measured using a modified version of the scale employed by Demuth et al. (2016) for measuring TC risk perception. The measurement included seven items on a five-point scale ranging from strongly disagree (1) to strongly agree (5). Sample items included "I think the potential impact from Judy is significant" and "I do not currently think Judy presents any threat to Florida" (reverse coded). An RP index was created by averaging the scores (Cronbach's $\alpha = 0.80$, $M = 3.81$, SD = 0.70).

2.4.4 | Risk Salience and Demographics (RQ4 and RQ5)

Following the questions that were specific to the forecast graphic were a series of more generalized questions related to risk salience. The questions were designed to gauge what forecast information would be most important to Florida residents in this scenario (Table 8). Respondents were asked to rate the importance of various components of the TC forecast using a five-point scale from not at all important (1) to extremely important (5). To determine if there were any demographic differences in the results, we also asked the participants how many hurricanes they have experienced, how long they have resided in Florida, and the zip code of their residence.

2.4.5 | Qualitative Data

The survey included open-ended prompts that encouraged participants to comment on their previous responses (e.g., "Please provide any comments you have regarding your understanding of the forecast graphic"). The comments were separated by condition and then assigned to a trained undergraduate research assistant who was masked to the research conditions. The undergraduate research assistant conducted a thematic analysis as outlined by Braun and Clarke (2006).

2.5 | Results

2.5.1 | Comprehension

A one-way ANOVA with Tukey HSD post hoc comparisons showed a significant difference in Perceived Comprehension (PC). Specifically, this analysis showed that people who viewed the status quo cone graphic reported significantly higher PC scores than those who viewed three of the four prototypes, $F(4, 1057) = 6.65$; $p < 0.001$; $\eta^2 = 0.025$ (Table 2). It is not surprising that people perceive the graphic that they have become familiar with over the last two decades as the most understandable. However, Prototype M1T1 did not score statistically lower.

There was no significant difference in Actual Comprehension among any of the forecast graphics, $F(4, 1057) = 0.68$, $p = 0.607$ (Table 3). Overall, the average score across all conditions was 1.94, indicating that participants were able to answer approximately two of the four questions correctly. This finding suggests the prototypes were no more or less understandable than the status quo cone overall.

Subsequently, a Kruskal–Wallis test using Bonferroni correction was performed to determine if any items on the AC scale stood out from the overall results. Indeed, a significantly higher number of participants that viewed the prototypes were able to correctly answer the question "If Judy were to make landfall in West Florida (e.g., Tampa Bay area), the strongest intensity that could reasonably be expected at landfall is" than those that viewed the status quo COU, χ^2 (4, 1055) = 25.10, $p < 0.001$. Specifically, more participants that viewed Prototype M2T1 correctly answered the question about strongest reasonable intensity than any other condition (Figure 6). A significantly higher number of participants that viewed the status quo COU were able to correctly answer the question "If Judy were to make landfall in West Florida, which of the following statements best describes the expected intensity" χ^2 (4, 1050) = 12.20, $p = 0.016$ (Figure 7). This result was somewhat counterintuitive because the COU does not explicitly provide information regarding landfall intensity.

2.5.2 | Effectiveness

There was no statistically significant difference in PME among any of the forecast graphics, $F(4, 1057) = 1.98$; $p = 0.095$. Overall, the average score across all conditions was 3.97, indicating that in general the participants agreed all graphics were effective (Table 4).

A series of one-way ANOVAs with Tukey HSD post hoc comparisons showed that people who viewed Prototype M1T2 were less

TABLE 3 | Actual comprehension (AC) scores.

	Status quo	Prototype M1T1	Prototype M2T2	Prototype M1T2	Prototype M2T1
M	1.91	1.88	1.89	2.00	2.00
SD	1.01	0.98	1.07	1.11	1.03

Note: Scale ranged from 0 = none correct to 4 = all correct.

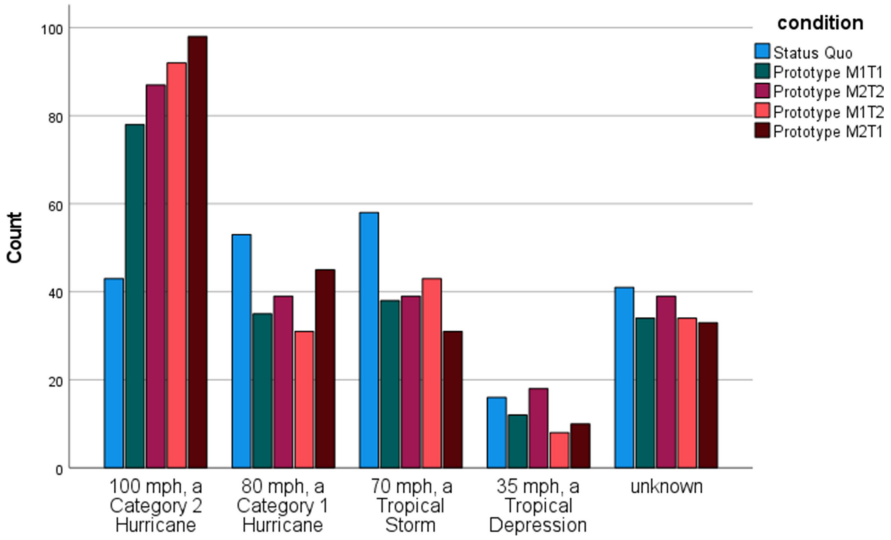


FIGURE 6 | Responses to the question "If Judy were to make landfall in West Florida (e.g., Tampa Bay area), the strongest intensity that could reasonably be expected at landfall is?" by condition. The correct answer is the leftmost response in the figure.

inclined to agree with the statement "this graphic would catch my attention" overall. This result is significantly lower than those who viewed the status quo COU, $F(4, 1056) = 2.63$, $p = 0.033$, $\eta^2 = 0.01$ (Table 5). Similarly, participants who viewed Prototype M2T2 were less likely to agree with the statement "the graphic is relevant to me" overall and that result is significantly lower than those who viewed the status quo COU and significantly lower than those who viewed Prototype M1T1, $F(4, 1055) = 4.48$, $p = 0.001$, $\eta^2 = 0.017$ (Table 5). It is important to note that all graphics showed the same forecast track over Florida, only the table of probabilities varied, yet that portion of the forecast message was interpreted differently by Florida residents. Those results suggest that the expanded table can reduce certain aspects of message effectiveness.

2.5.3 | Risk Perception

There was no statistically significant difference in risk perception (RP) among any of the forecast graphics ($F(4, 1057) = 0.25$, $p = 0.908$). Overall, the average score across all conditions was 3.81, indicating that in general the participants agreed all graphics depicted some risk (Table 6).

A series of one-way ANOVAs with Tukey HSD post hoc comparisons showed that people who viewed Prototype M2T2 were less inclined to agree with the statement "it is likely Judy will hit Florida" overall and that result is significantly lower than those who viewed the status quo COU, $F(4, 1056) = 2.96$, $p = 0.019$, $\eta^2 = 0.011$ (Table 7). On the other hand, those who viewed Prototype M2T1 were more likely to agree with the statement "I believe Judy is more of a threat to Florida than a typical Tropical Storm would be" overall and that result is significantly higher than those who viewed the status quo COU, $F(4, 1056) = 2.60$, $p = 0.035$, $\eta^2 = 0.01$ (Table 7). This result suggests that Prototype M2T1 viewers believed the TC posed a higher risk to Florida than a typical tropical storm. Considering that Judy was a tropical storm that was predicted to strengthen in the scenario posed to the participants, this result is noteworthy.

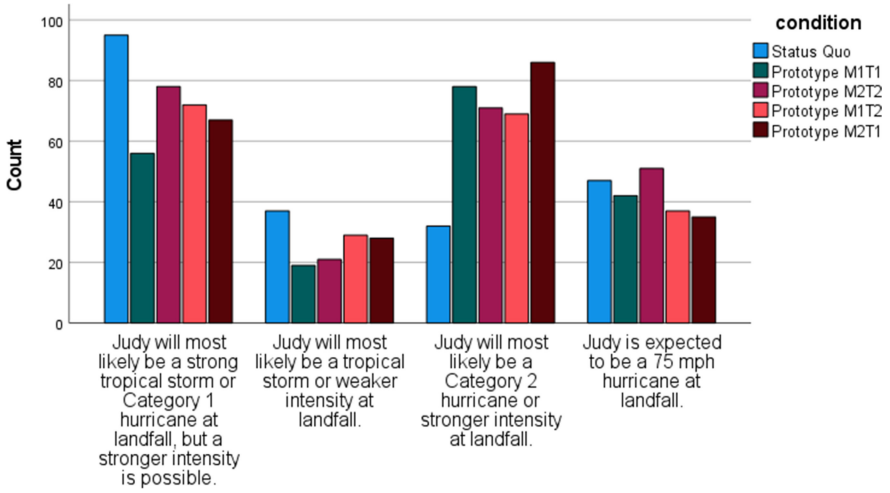


FIGURE 7 | Responses to the question "If Judy were to make landfall in West Florida, which of the following statements best describes the expected intensity?" by condition. The correct answer is the leftmost response in the figure.

TABLE 4 | Perceived message effectiveness (PME).

	Status quo	Prototype M1T1	Prototype M2T2	Prototype M1T2	Prototype M2T1
M	4.07	3.99	3.9	3.91	3.98
SD	0.68	0.68	0.71	0.76	0.73

Note: Items ranged from 1 = strongly disagree to 5 = strongly agree.

TABLE 5 | Responses to specific PME items by condition.

		Status quo	Prototype M1T1	Prototype M2T2	Prototype M1T2	Prototype M2T1
The graphic would catch my attention.	*M*	4.09_{a}	$3.9_{a,b}$	$3.83_{a,b}$	3.77_{b}	$3.94_{a,b}$
	SD	0.97	1.08	1.1	1.18	1.11
The graphic is relevant to me.	*M*	4.20_{a}	4.16_{a}	3.84_{b}	$4.00_{a,b}$	$3.96_{a,b}$
	SD	1	0.88	1.1	1.07	1.07

Note: Item ranged from 1 = strongly disagree to 5 = strongly agree. Means with different subscripts differ at $p < 0.05$ using Tukey HSD post hoc comparisons.

TABLE 6 | Risk perception (RP).

	Status quo	Prototype M1T1	Prototype M2T2	Prototype M1T2	Prototype M2T1
M	3.8	3.8	3.77	3.82	3.84
SD	0.66	0.72	0.7	0.71	0.72

Note: Items ranged from 1 = strongly disagree to 5 = strongly agree.

TABLE 7 | Responses to specific RP items by condition.

		Status quo	Prototype M1T1	Prototype M2T2	Prototype M1T2	Prototype M2T1
It is likely Judy will hit Florida.	*M*	4.28_{a}	$4.13_{a,b}$	3.99_{b}	$4.16_{a,b}$	$4.17_{a,b}$
	SD	0.82	0.85	0.99	0.91	0.87
I believe Judy is more of a threat to Florida than a typical Tropical Storm would be.	*M*	3.39_{b}	$3.53_{a,b}$	$3.57_{a,b}$	$3.6_{a,b}$	3.71_{a}
	SD	1.00	1.06	1.04	1.05	1.05

Note: Item ranged from 1 = strongly disagree to 5 = strongly agree. Means with different subscripts differ at $p < 0.05$ using Tukey HSD post hoc comparisons.

TABLE 8 | Mean rating of importance on risk salience items by certain demographics.

	Overall ($n = 1058$)	In FL over 20 years ($n = 589$)	Experienced + hurricanes ($n = 692$)
How important is it for you to be given the probabilities of Judy reaching certain categories or intensities?	4.17	4.3	4.25
How important is it for you to learn that Judy is a threat to your state?	4.16	4.29	4.23
How important is it for you to be given the strongest reasonable intensity at landfall?	4.12	4.28	4.20
How important is it for you to be given the probabilities of intensity at landfall?	4.12	4.26	4.20
How important is it for you to learn where Judy could make landfall?	4.08	4.20	4.15
How important is it for you to see all of this information on a map?	4.00	4.15	4.09
How important is it for you to be given the probabilities of different landfall locations?	3.97	4.06	4.03
How important is it for you to be given the weakest reasonable intensity at landfall?	3.48	3.52	3.50

Note: Items ranged from 1 = not at all important to 5 = extremely important. There was not a significant variation in the means by condition.

2.5.4 | Risk Salience

Ordering the mean responses illustrated that the probability of the TC strengthening to a given category or intensity was the most important message characteristic (Table 8). There was no significant variation in the mean ratings by condition/graphic. The second most important component was simply knowing that a TC is a threat to the state, followed by knowing the strongest reasonable intensity at landfall. The least important of the eight items was knowing the weakest reasonable intensity at landfall.

2.5.5 | Demographics

Participants were asked to provide the zip code of their current residence. The data were then filtered according to coastal versus inland zip codes and then further separated by coastal regions of Florida (i.e., West, East, and Panhandle). Analysis showed those living along the west coast of Florida ($n = 196$), roughly the area outlined in red in Prototypes M2T2 and M2T1 and most directly in the path of "Judy," were significantly more likely to agree with the statement "this graphic is relevant to me" regardless of which graphic they viewed, $F(3, 1052) = 5.59$,

$p = 0.001$, $\eta^2 = 0.016$. The same group also perceived a significantly higher risk (RP), $F(3, 1054) = 2.69$, $p = 0.045$, $\eta^2 = 0.008$, perceived the message to be significantly more effective (PME), $F(3, 1054) = 4.56$, $p = 0.004$, $\eta^2 = 0.013$, and scored significantly higher on AC, $F(3, 1054) = 4.99$, $p = 0.002$, $\eta^2 = 0.014$. Note, apart from AC, the scores of those on the west coast of Florida were similar to those who reported living on the coast in the Florida panhandle (roughly the area outlined in orange in Prototypes M2T2 and M2T1), but due to the relatively small sample from that specific region ($n = 35$) the means were not significantly different from those living elsewhere in Florida. Also, a factorial ANOVA revealed a nonsignificant condition by region interaction across the measures.

Given that prior experience with TCs has been shown to have some impact on risk perception and risk salience (Bostrom et al. 2018) and on TC forecast comprehension (Eicher et al. 2023), the data were filtered according to those that have lived in Florida longer than 20 years and those that have experienced at least 3 TCs following Eicher et al. (2023). The overall results did not change significantly. Interestingly, the order of important information on the risk salience list did not change at all (Table 8).

2.5.6 | Qualitative Data

Approximately two-thirds of the survey participants responded to the open-ended questions. In many cases, the response offered no additional information (e.g., "no further comment"). However, certain replies provided valuable insight that clarifies the quantitative results.

Analysis of the comments regarding the status quo COU revealed a general lack of understanding that the "H" and "S" implied the TC intensity and that participants wanted to know more about the expected intensity. "It was not obvious what the circled H and S meant," "I do not know what the Hs and Ss stood for," "I really do not see the category levels on the graphic," "don't know the wind speed," and "I need more info on how the [sic] storm builds in wind speeds" are just some examples of comments by participants. Similar themes were noted when participants were asked what information would be most important to them in this scenario. "Where and how intense landfall will be is the most important," "what category it will be when makes landfall," "the landfall areas with proximity and the level of forecasted intensity," and "landfall and category are the most important" are examples of replies to the question about important information. The responses seem to confirm the need for a forecast product that provides the expected landfall intensity.

Analysis of Prototype M1T1 revealed there was room for improvement in the monochromatic color scheme. "I think if it had more colors it would be more attention grabbing" and "colors would make it easier to understand" were some of the suggestions. Interestingly, there were far more mentions of wanting additional updates with this graphic than with the status quo cone. "The graphic is a probability [sic] forecast and needs to be monitored by people for changes," "effective but circumstances change with time and updates would be something I would seek," "No doubt the graphic catches eyes but how often is it updated," "I would like more news on it," and "I would like updates as the hurricane center updates their forecast" are some examples of comments that mentioned a desire for further updates. When asked about important information, timing was mentioned a few times in this condition. For example, "how rapidly it's approaching if it could speed up and come make landfall earlier" and "the category and when it will make landfall is important."

At least one participant noticed the additional colors in Prototype M2T2, replying, "very appealing graphic and colorful." Moreover, at least one participant found the additional columns of probabilities in Prototype M2T2 to be helpful, stating, "really like the chart of probability." However, many participants found it to include too much information. "Too many variables listed," "it was a lot of information at once seemed cluttered," and "I don't understand much of this" are example responses. It is also worth noting that some reference to maps was made in response to the question about important information in this condition. For example, "a map provides a good visual to see the potential impact," "map graphics are very helpful," and "all graphs and maps give a clearer picture to the public." This seems to suggest that the probability information might be better represented on a map rather than in a table.

Similar to Prototype M2T2, many participants noted that the expanded table in Prototype M1T2 was overwhelming. "There is a lot of info at once," "rather busy too much info to process easily," "too much information to digest about different intensities and percentages," and "too many numbers and columns to read" are examples from participants. Like Prototype M1T1, timing was mentioned a few times in response to the question about important information. For example, "timing of landfall" and "the exact day the storm will hit and the exact area even the time of day and strength of winds" were some of the responses to that question. Interestingly, this was the only condition in which color was not mentioned.

Prototype M2T1 included the color-coded landfall probability map and, based on the responses from some participants, that seemed to be an improvement. "Color coded works great as a visual," "I like the colors they really make a difference," and "it's clear and the use of colors is important" are examples of responses from different participants. Similar to Prototype M1T1, there was also a general theme of wanting more updates. "You would want to watch for updates on this storm," "need more updates on a regular basis," and "keeping updated on these storms are [sic] a must" are just some of the comments suggesting that the TC needed to be further monitored. It should be noted that not all the comments were positive. "Simple graphic yet not so simple to understand" wrote one participant. Also, once again, timing was mentioned. For example, "where is it going to land and how much time we have" and "when it will hit my location and the speed or category" were some of the responses to the question about important information. This seems to suggest that the prototypes lack a valuable piece of information contained in the status quo COU (i.e., days and times).

2.6 | Conclusions From Study 1

Considering that the COU graphic has been in use for over 20 years and is now the most viewed graphic on the NHC

website (Millet et al. 2020), it is not surprising that participants thought they understood the status quo COU more than most of the prototypes. In fact, Perceived Comprehension of the status quo COU increased ($M = 4.09$) when the results are filtered to only include participants that have lived in Florida longer than 20 years.

The lack of significant differences in Actual Comprehension between the prototypes and the COU could be viewed as a positive result. Given the increased complexity of the prototypes, it implies that the additional information did not cause additional confusion. Furthermore, within some of the items on the AC index, there were hints that the prototypes were at least a step in the right direction. A significantly higher number of people who viewed the prototypes, especially Prototype M2T1, were able to correctly answer the question about the strongest reasonable intensity at landfall compared with the status quo COU. The strongest reasonable intensity is information that is not currently included in the status quo COU. It is, however, information that is highly sought after, as it ranked third most important in the survey. That suggests that by conveying the information that the TC could be stronger at landfall than currently predicted, the prototypes represent an improvement over existing products that provide only a deterministic intensity forecast.

The fact that viewers of the status quo COU were more likely overall to correctly answer the AC question about the expected intensity at landfall is a bit more difficult to interpret. That result begins to make sense when one considers that the expected intensity at landfall was not specifically stated in any of the graphics. That information can be deduced by correctly interpreting the probability information in either Prototype M2T2 or Prototype M1T2. In fact, the next highest number of correct answers to that question came from Prototype M2T2 and Prototype M1T2 viewers, respectively. However, the status quo COU graphic included an "H" symbol, indicating hurricane strength, on the forecast track near the coastline. That symbol is not intended to represent the expected landfall intensity, but the expected intensity at that specific forecast time. Nonetheless, based on survey comments revealing some misunderstanding of that symbol, participants may have interpreted it as such. In fact, prior research by Drake (2012) found that users often misinterpret the TC intensity forecast as the probable storm intensity at landfall.

While there was not a significant difference in the PME index overall or the Risk Perception Index overall, there are perhaps some meaningful results in the individual questions. Specifically, Prototype M1T2 was deemed least likely to catch the attention of the participants, and Prototype M2T2 was deemed least relevant and least likely to hit Florida. Considering all graphics depicted the exact same forecast track, the fact that the prototypes with additional numbers were also the ones rated as least relevant, least attention grabbing, and least of a threat might indicate that the survey participants were overwhelmed or distracted by that much probability information. That idea was suggested in several survey comments. It is also possible that the additional information in the tables reduced the amount of overlap with the map and therefore decreased the relevance and attraction.

Also included in the Risk Perception index was the statement "I believe Judy is more of a threat to Florida than a typical Tropical Storm would be." Viewers of the prototypes were overall more likely to agree with that statement than viewers of the status quo cone and in a statistically higher amount with Prototype M2T1. Recall that in the scenario posed to the participants, Judy was a tropical storm that was predicted to intensify into a hurricane. Prior research has shown that lower category TCs are often dismissed by publics as familiar, that is, "*just* a tropical storm," and therefore non-threatening (e.g., Ruin et al. 2008). This suggests that viewers of the prototypes, especially Prototype M2T1, were less likely to dismiss the situation. Indeed, there was a general theme of wanting further updates in the survey comments from those who viewed Prototypes M1T1 and M2T1. Encouraging situational awareness is a potential positive outcome of those prototypes.

Regarding the demographic questions, the answers are consistent with previous research. Prior studies have shown that the center of the cone is the most relevant part to most people (e.g., Millet et al. 2020); indeed, we noted some regional differences in the results. Our survey also demonstrated that the probability of stronger intensities was information that everyone, regardless of demographics, found valuable and confirms prior suggestions that the public wants numerical probability information (e.g., Rosen et al. 2021; Ripberger et al. 2022).

Overall, the quantitative analysis combined with comments from survey respondents suggests some of the prototypes are a step in the right direction. The results also suggest the prototypes might fill a gap in the current TC forecast product suite by offering probability information that end users find valuable. This first study provides motivation and guidance for the next phase of the research.

3 | Study 2: Updated Prototypes

3.1 | Formative Research

As a follow-up to Study 1, the research team assembled NHC partners (i.e., emergency managers, broadcast meteorologists, National Weather Service personnel) and representative experts (i.e., risk communication and graphic design experts) using a Communities of Practice approach (Wenger 1998). The goal was to garner more constructive feedback on the prototypes' design, which, combined with the results from Study 1, could be used to improve the prototypes. Participants had to be actively contributing to the weather enterprise and competent in its shared repertoire (Wenger 1998) and working in a TC-prone area of the United States. The Communities of Practice framework has been extended in ways that inform effective communication for problem solving in the contexts of natural disaster warnings (Sellnow et al. 2017). Additionally, input from stakeholders in this fashion proved to be valuable to the NHC in the development of the now operational storm surge products (Morrow et al. 2015).

The larger group of NHC partners and experts was split into focus groups of 6–10 participants that met online via Zoom. The Zoom platform allowed for participation all around the US coastline from New Jersey to Texas and provided recordings for later analysis. This section outlines the feedback from focus

group participants, along with a subsequent literature review that informed the modification of the prototypes.

3.1.1 | First Modification—Color Coding

Feedback from focus group participants echoed that from Study 1 participants to some extent. Focus group members almost unanimously agreed that a single-color category for any probability of landfall over 5% was not useful. For example, Participant #3-3 stated:

> I came away with the same impression as [Participant #3-1]. Everything looks to be the same. I can't really distinguish, outside of the cone being there, which areas are potentially more at risk than others. My eyes are just immediately drawn to the red being everywhere.

There was general agreement that color-coding the probability of landfall into categories, as was done in Prototypes M2T2 and M2T1, was at least a step in the right direction. There was some discussion about what colors should be implemented and how to accommodate color blind users. Even still, the focus group participants largely preferred those prototypes. Participant #3-4 summarized:

> I really like that there is more detail. I think we're going in the right direction here. But we're mixing and matching. We're trying to apply a different color scale to every product ... I don't know if I'd be interpreting this properly.

In related literature, Gerst et al. (2020) used visualization science to improve the representation of probability information in long-range temperature and precipitation outlooks from the Climate Prediction Center. The authors noted that "a vast majority of respondents used color as their primary cue for outlook interpretation," which is precisely what the visualization literature would predict (Gerst et al. 2020). In a review of research-backed guidelines, Franconeri et al. (2021) noted that practitioner guides recommend color-coding accompanying text to match what is in the graphic to ensure that the viewer will match the pattern in the data to the relevant reference in the visualization. Similarly, in a controlled study, Lin et al. (2013) demonstrated that "semantically-resonant" colors (i.e., colors that evoke a certain concept) improve speed on chart reading tasks.

Based on those suggestions along with the numerous comments related to color in Study 1, original prototypes that portrayed landfall probability monochromatically (i.e., Map 1 in Study 1) were dropped from consideration in this round. Additionally, tables on the prototypes were modified with additional color coding for both landfall probabilities and intensity probabilities. The hope was that color coding the table would increase the overlap with the map and perhaps make it more relevant and attention-grabbing.

3.1.2 | Second Modification—Adding Important Probabilistic Information

The focus group participants indicated the smaller table in Prototype M2T1, including the probability of landfall and the strongest reasonable landfall intensity, provided simple, easy-to-read, useful information that provided some measure of forecast uncertainty. For example, Participant #4-5 stated:

> I still like prototype [M2T1]. It's a little cleaner ... And again, I still like that strongest, reasonable landfall intensity, information, and that's in there without a lot of other numbers to go with it.

Participant #4-9 agreed:

> I think [M2T1] is, I think, perhaps the best because it is cleaner. It's simpler. And like I said, that's less information that everyone viewing, regardless of background, has to try to make sense of.

Many suggested that it should also include a column providing the *most likely* intensity at landfall. In other words, two columns of numerical information was good, but perhaps not quite enough. For example, Participant #4-6 asked, "so I'm wondering, could you just add one column to prototype [M2T1] like next to the strongest reasonable, you had a most likely intensity?"

Several focus group members indicated that Prototype M2T2, with the additional four columns of probabilities of various intensities, could be helpful to some users. While referring to the larger table, Participant #3-4 explained:

> I did want to point out also that the percentages, the actual percentages for specific locations, I think that is an excellent tool for some of our more sophisticated partners. I think that's really, really good information that we can't really get anywhere else.

The general consensus was that there may not be a single version of the table that satisfies the needs of every possible end user.

Concerning what probabilistic information should be incorporated, note that all the original prototypes included the strongest reasonable intensity at landfall in the table, defined as the 10% exceedance value (i.e., $<10\%$ of the forecasts fall above this threshold; Trabing et al. 2023), but none of them specified the most likely intensity at landfall. That information could be deduced, albeit with a broad range of possible categories (e.g., category 1-5 or 3-5), by correctly interpreting the probability distributions in the original prototypes with the larger table. However, as Durbach and Stewart (2011) discovered, probability distributions can be difficult for some people to correctly interpret. Given the feedback from focus group members and the apparent confusion regarding the question about expected landfall intensity in Study 1, most likely landfall intensity seemed like an especially important addition.

It was tempting to also include the weakest reasonable intensity to mimic the three-point (minimum–median–maximum) approximations suggested by Durbach and Stewart (2011). However, results from Study 1 showed that information consistently ranked last on the list of important information to Florida residents. Therefore, all prototypes were modified to specify only the most likely intensity and the strongest reasonable intensity at landfall.

3.1.3 | Third Modification—With or Without the COU

Each focus group session included at least some debate about whether the actual COU should be included as part of the prototype graphics. Since the COU is designed such that the entire track of the TC can be expected to remain within the COU roughly two-thirds of the time (National Hurricane Center 2024), roughly one-third of the time the TC tracks outside of the COU. As the prototype graphics included probability of landfall in addition to probability of intensity, they essentially highlighted this fact, which was the source of the debate. Some focus group members felt it was too confusing to show some probability of landfall outside of the COU, while others argued it was important to draw attention to that possibility.

Based on previous research, both sides of the debate had good arguments. The National Oceanic and Atmospheric Administration (NOAA), contracted ERG to conduct a thorough review of both government and academic research findings concerning the COU. ERG stated that the COU may be "the most misinterpreted product within the tropical cyclone product suite" (ERG 2019, 1). Specifically, they found that viewers of the COU tended to focus too much on the center of graphic and therefore fail to recognize that landfall was possible elsewhere. Similarly, a survey of Florida residents conducted by Evans et al. (2022) found that 48% of respondents incorrectly believed that the COU shows all possible paths of the center of the COU. Such research could be used to argue that including the COU would only cause additional confusion. However, the same research could be used to argue that the addition of the COU to these prototypes is necessary to dispel common misconceptions. Considering this duality, the research team opted to test new prototypes both with and without the COU included.

3.1.4 | Fourth Modification—A New Version Without a Table

Another common source of debate in the focus group sessions concerned the inclusion of a table of probabilities in the prototypes. Opinions varied but focus groups participants thought the table could lead to confusion, simply was not helpful, or might lead to "information overload." It was also suggested by several focus group members that the probability information in the tables might be better displayed graphically. Participant #4-8 said, "a graphical representation of those numbers would be helpful," and that was echoed by Participant #4-6, "I would use some of those numbers and create a new graphic," and Participant #4-9, "we need to be able to give them the information that they need visually, and like I said, I understand the numbers, but not everyone is going to read those numbers." Following those recommendations, another major change was the creation of a prototype that did not include any table of probabilities. Instead, some of that probability information was displayed in the form of color-coded maps.

There are several reasons to believe that a visual representation might be more effective than a table. First, as Knaflic (2015) explains, tables require longer processing time than graphics, meaning "a well-designed graph will typically get the information across more quickly than a well-designed table." Second, while the research on precisely what combination of audio, visual, and textual information promotes learning is mixed, after a review of the empirical research, Trypke et al. (2023) recommend removing either the visualization or the written text when the material is complex. Third, while conducting research on earthquake early warning messages, Sellnow et al. (2019) discovered visual intensity displays were more effective than numerical displays. More specific to this research, in a small experiment with undergraduate psychology students, Liu et al. (2019) found some success conveying both TC track and intensity uncertainty using only maps. Finally, the Durbach and Stewart (2011) study cautioned against providing too much probability information and advocated for more concise formats.

3.1.5 | Final Modifications—Visual Improvements

In addition to the previously mentioned suggestions, the focus group participants provided suggestions for visually improving the prototypes, such as repositioning the legend to make it more prominent and reducing the size of the latitude/longitude lines to improve visibility of other map features.

Franconeri et al. (2021) point out that when multiple pieces of information compete for attention, that competition tends to be won by information that is different or brighter in color, largest in size, or presented at the top or left of a display. The legend was modified accordingly to make it more prominent and easier to read in all the new prototypes.

3.2 | Testing the Modified Prototypes

While the modifications made to the initial prototypes were based on expert opinions and grounded in prior research, the research team nonetheless wanted to see if those modifications would lead to different answers to the research questions. To test that, we launched a second survey.

3.2.1 | Survey Participants

Participants in this online experiment were once again recruited from Qualtrics' Florida population panel ($N = 616$). This sample was 70.7% female and ranged in age from 18 to 85 years ($M = 46.9$, SD = 16.9). The participants identified themselves as 63.3% Caucasian, 16.1% African American, 10.2% Hispanic or Latinx, and 2.0% Asian. Chi-square tests indicated there was no

significant difference in any of the demographics across the conditions, once more confirming the participants were sufficiently randomized.

3.2.2 | Stimuli

For this study, three prototype forecast graphics were created based on National Hurricane Center advisory 4A for TC Ian from September 24, 2022, at 2 AM EDT (Figure 8). Two versions of each prototype were created, each with and without the COU on the map. The first set of prototypes was a modified version of M2T2 from Study 1 and included a longer table of intensity probabilities along with a color-coded map of landfall probabilities (Figures 9 and 10). Although the long table did not test well in Study 1, the research team wanted to see if it could become part of a viable prototype with the previously mentioned modifications. The second set was a modified version of M2T1; a shorter table of intensity probabilities accompanied the color-coded landfall probability map (Figures 11 and 12). The final set did not include any table of probabilities. Instead, the probabilistic information detailed in the short table was portrayed as a series of color-coded maps (Figures 13 and 14). Just as before, intensity probabilities for the prototype graphics were generated by the landfall distribution Product (LDP) using the case of TC Ian (Trabing et al. 2023).

Much like the TC forecast in Study 1 based on Elsa, the actual cone of uncertainty for Ian covered almost the entire state of Florida, making it relevant to all Florida residents and therefore relevant to the panel of survey participants. Ian, which developed after Study 1 was conducted, was chosen because of its greater potential impact and the higher levels of uncertainty in the forecast, exactly what these prototypes are meant to address. In other words, the graphics based on Ian included more meaningful probabilities in all levels of the intensity scale compared with Elsa, while the general forecast tracks remained similar.

Consistent with Study 1, the name of the TC was changed to "Judy" in all graphics to avoid confusion with the real system. Participants were given the following instructions just before viewing the graphic:

> You are about to see a forecast graphic for a hypothetical tropical cyclone named "Judy." Judy is not a real tropical cyclone but is based on a real scenario. In this scenario, Judy is currently a tropical storm located in the Central Caribbean Sea, about 350 miles southeast of Kingston, Jamaica with maximum sustained winds of 40 mph. Judy is currently moving toward the west at 13 mph but is expected to turn north in the direction of Florida while strengthening into a hurricane.

That introduction closely matches the real Ian scenario on September 24, 2022. The biggest changes were made to the prototypes themselves, as described previously based on study 1.

3.2.3 | Survey Measures

To enable direct comparisons between the revised prototypes and the original ones, the same survey instrument

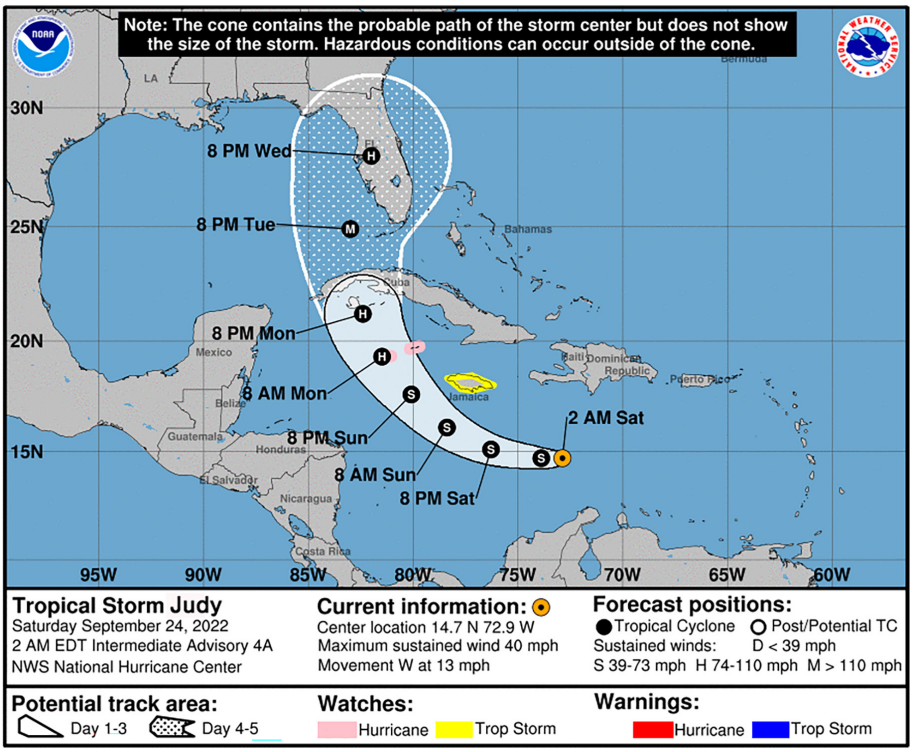
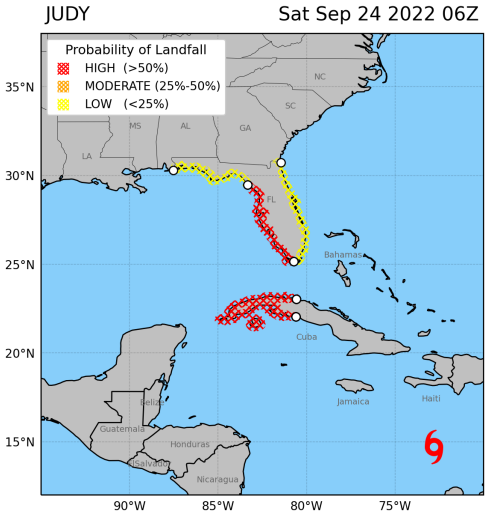


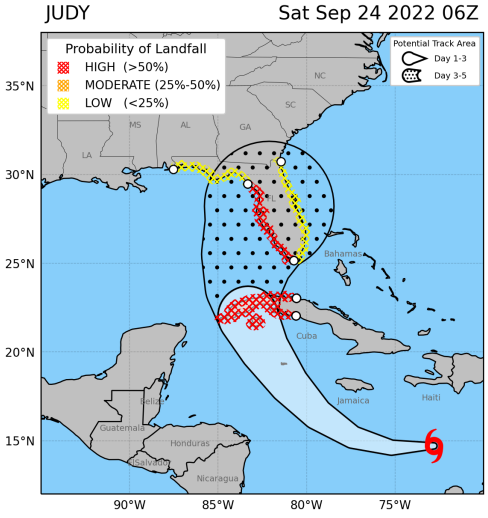
FIGURE 8 | Condition 1, status quo.



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Tropical Depression or Weaker (<39 mph)	Tropical Storm (39-73 mph)	Hurricane Cat 1-5 (>74 mph)	Major Hurricane Cat 3-5 (>110 mph)	Strongest Reasonable Landfall Intensity
W CUBA	77%	2%	8%	90%	44%	Cat 4 (140 mph)
FL Pan	15%	<2%	6%	93%	60%	Cat 4 (140 mph)
West FL	59%	<2%	7%	94%	59%	Cat 4 (145 mph)
East FL	10%	<2%	11%	89%	59%	Cat 4 (140 mph)

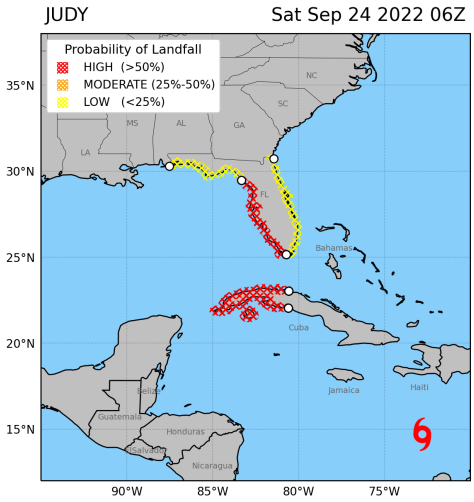
FIGURE 9 | Condition 2, long table, no cone.



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Tropical Depression or Weaker (<39 mph)	Tropical Storm (39-73 mph)	Hurricane Cat 1-5 (>74 mph)	Major Hurricane Cat 3-5 (>110 mph)	Strongest Reasonable Landfall Intensity
W CUBA	77%	2%	8%	90%	44%	Cat 4 (140 mph)
FL Pan	15%	<2%	6%	93%	60%	Cat 4 (140 mph)
West FL	59%	<2%	7%	94%	59%	Cat 4 (145 mph)
East FL	10%	<2%	11%	89%	59%	Cat 4 (140 mph)

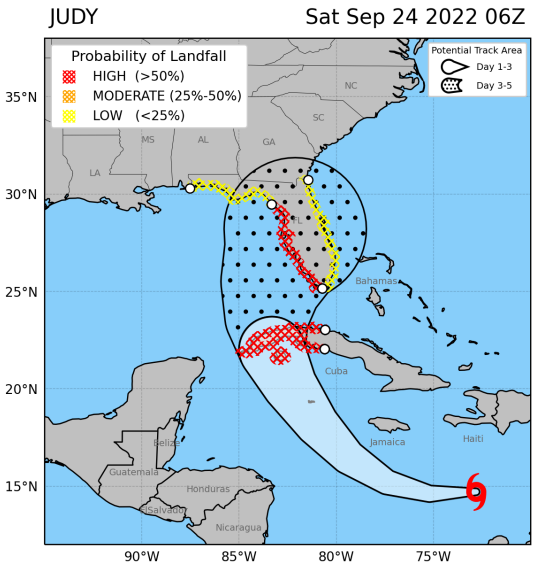
FIGURE 10 | Condition 5, long table, with cone.



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Most Likely Landfall Intensity	Strongest Reasonable Landfall Intensity
W CUBA	77%	Cat 2 (110 mph)	Cat 4 (140 mph)
FL Pan	15%	Cat 3 (115 mph)	Cat 4 (140 mph)
West FL	59%	Cat 3 (115 mph)	Cat 4 (145 mph)
East FL	10%	Cat 3 (115 mph)	Cat 4 (140 mph)

FIGURE 11 | Condition 3, short table, no cone.



Tropical Cyclone Intensity Probabilities at Landfall

Region	Probability of Landfall	Most Likely Landfall Intensity	Strongest Reasonable Landfall Intensity
W CUBA	77%	Cat 2 (110 mph)	Cat 4 (140 mph)
FL Pan	15%	Cat 3 (115 mph)	Cat 4 (140 mph)
West FL	59%	Cat 3 (115 mph)	Cat 4 (145 mph)
East FL	10%	Cat 3 (115 mph)	Cat 4 (140 mph)

FIGURE 12 | Condition 6, short table, with cone.

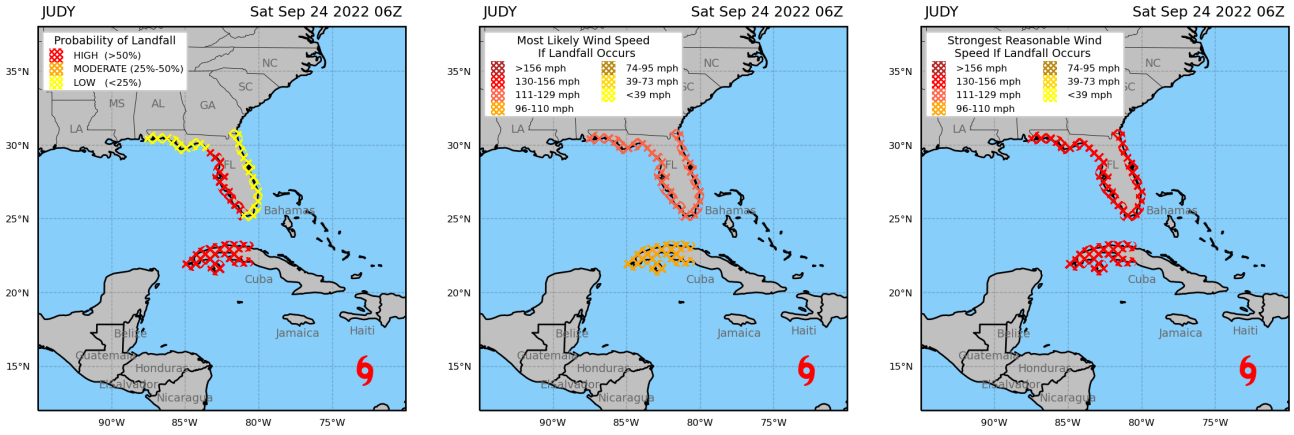


FIGURE 13 | Condition 4, color-coded map, no cone.

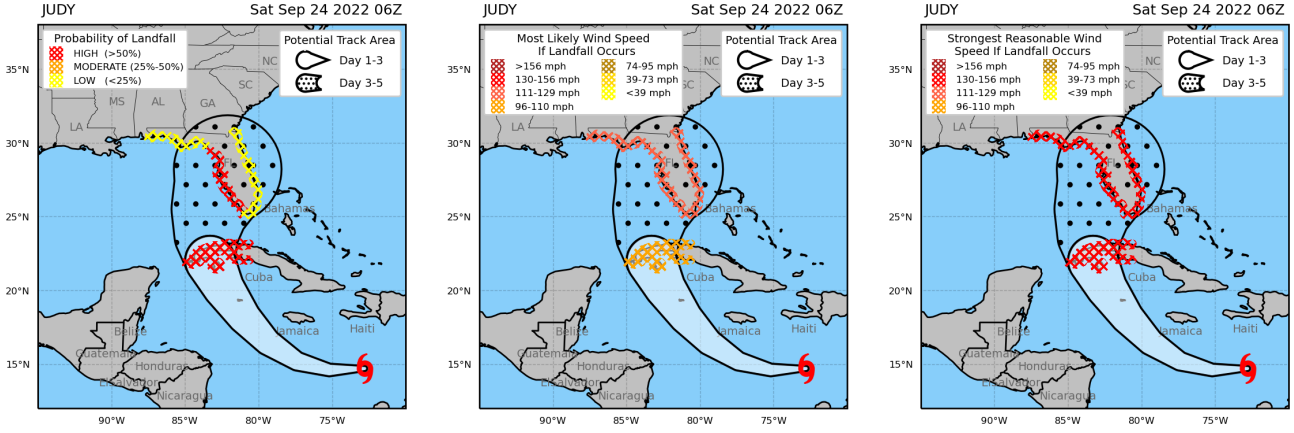


FIGURE 14 | Condition 7, color-coded map, with cone.

was employed with only minor changes. Participants were randomly assigned to view one of the 7 conditions. The status quo COU graphic once again served as the baseline. Participants were able to review the graphic for as long as they wanted and were encouraged to leave the graphic open in another browser window to have available while answering the questions. Following the graphic were the same series of questions designed to measure variables of comprehension,

effectiveness, and risk perception. Reliability measurements were nearly identical to the original survey for the Perceived Comprehension (PC) index (Cronbach's $\alpha = 0.90$, $M = 3.76$, SD = 0.88) and the PME index (Cronbach's $\alpha = 0.86$, $M = 3.98$, SD = 0.72); improved slightly for the Risk Perception (RP) index (Cronbach's $\alpha = 0.83$, $M = 3.92$, SD = 0.75).

Numeracy was added for Study 2 based on feedback from focus group participants suggesting that non-expert users might have difficulty interpreting the table of probabilities. Additionally, prior research has shown a relationship between numeracy and risk perception (e.g., Kahan et al. 2012; Peters et al. 2006), and a relationship between numeracy and comprehension of visual uncertainty representations (Toet et al. 2019; Tak et al. 2015). Tak et al. (2015) specifically examined visual representations of uncertainty in temperature forecasts, while Toet et al. (2019) adopted a case of ensemble forecasting from life sciences. Toet et al. (2019) found that participants with lower numeracy provided more inaccurate responses. However, numeracy only weakly correlated with education level, suggesting that numeracy is a stronger predictor of biases in graphical uncertainty interpretations than education.

We employed the Subjective Numeracy Scale (SNS) as originally developed by Fagerlin et al. (2007). The SNS has proven to be both reliable and highly correlated with objective numeracy measures and it has been validated in risk communication (Zikmund-Fisher et al. 2007). The SNS was also the numeracy scale employed in Tak et al. (2015). Perhaps most importantly, the scale included questions that are directly relevant to this study. Specifically, we were interested in the question, "when you hear a weather forecast, do you prefer predictions using percentages or predictions using only words," which is measured on a scale ranging from "always prefer percentages" (1) to "always prefer words" (6) and reversely coded (Fagerlin et al. 2007). To keep the total length of the survey roughly the same in both studies, questions related to risk salience were dropped from the survey in Study 2 as those responses seemed clear and unlikely to change.

In total, eight items are designed to be averaged together to form the SNS (Fagerlin et al. 2007). The SNS demonstrated good reliability in this study (Cronbach's $\alpha = 0.83$, $M = 4.07$, SD = 1.06). Once compiled, participants were split into three tercile groups based on their SNS score; those with the lowest numeracy scores were compared against those with the highest numeracy scores.

3.3 | Results

3.3.1 | Comprehension

As with Study 1, it is expected that the COU would exhibit greater Perceived Comprehension (PC) since it has been in use for over two decades, whereas the prototypes have not previously been viewed by the respondents. The status quo COU scored the same on both surveys despite the fact that one was based on Elsa and the other was based on Ian. However, a notable difference with Study 2 is that only one of the prototypes scored significantly lower than the status quo COU graphic based on a one-way ANOVA with Tukey HSD post hoc comparisons. In Study 1, all but one prototype scored significantly lower than the status quo COU graphic. Specifically, the Study 2 analysis showed that people who viewed the status quo COU graphic (Condition 1) reported significantly higher PC than those who viewed the long table prototype with the cone (Condition 5), $F(6, 615) = 2.16$; $p = 0.046$; $\eta^2 = 0.021$ (Table 9).

There was not a significant difference in Actual Comprehension (AC) among any of the prototypes in Study 1. In contrast here, the short table, both with and without the cone (conditions 3 and 6), received significantly higher AC scores than the status quo cone, $F(6, 615) = 2.85$, $p = 0.010$, $\eta^2 = 0.027$ (Table 10). The average score across all conditions was 1.94 in Study 1, indicating that the average participant was able to answer approximately two of the four questions correctly. That improved to an average of 2.30 with this set of prototypes. This suggests an overall improvement to the understandability of the prototypes.

As before, the researchers were curious if any of the prototypes stood out from the status quo COU as designed—for conveying the strongest reasonable intensity. A Kruskal–Wallis test using Bonferroni correction revealed that a significantly higher number of participants that viewed the prototypes were able to correctly answer the AC question "If Judy were to make landfall in West Florida (e.g., Tampa Bay area), the strongest intensity that could reasonably be expected at landfall is" than those that viewed the status quo cone, χ^2 (6, 616) = 30.47, $p < 0.001$. Specifically, more participants that viewed the long table prototype without the cone (Condition 2) correctly answered the question about strongest reasonable intensity than any other condition (Figure 15). Adding the most likely intensity information was perhaps constructive as the responses to that specific question were not significantly different in Study 2.

TABLE 9 | Perceived comprehension scores.

	Condition						
	1—Status quo cone	2—Long table, no cone	3—Short table, no cone	4—Map, no cone	5—Long table with cone	6—Short table with cone	7—Map with cone
M	3.98_{a}	$3.62_{a,b}$	$3.88_{a,b}$	$3.73_{a,b}$	3.59_{b}	$3.76_{a,b}$	$3.72_{a,b}$
SD	0.89	0.97	0.94	0.8	0.85	0.86	0.8

Note: Items ranged from 1 = strongly disagree to 5 = strongly agree. Means with different subscripts differ at $p < 0.05$ using Tukey HSD post hoc comparison.

3.3.2 | Effectiveness

In Study 1, no significant difference in PME was found among any of the forecast graphics. Here, the long table with the cone (Condition 5, Figure 11) scored significantly lower in PME than most other conditions; $F(6, 614) = 2.13$, $p = 0.049$, $\eta^2 = 0.021$ (Table 11).

3.3.3 | Risk Perception

There was no significant difference in risk perception among any of the forecast graphics; $F(6, 614) = 0.64$, $p = 0.695$. Overall, the average score across all conditions was 3.92, which was comparable to the average of 3.81 in Study 1 and indicates that, in general, the participants agreed all graphics depicted some risk (Table 12).

3.3.4 | Numeracy

Numeracy skills were equally distributed across all conditions (i.e., a Chi-square test was not significant) indicating that participants were properly randomized (Table 13). When the participants were filtered according to scores on the subjective numeracy scale (SNS, low vs. high), some significant differences were discovered. Specifically, in terms of Actual Comprehension (AC), those that scored highest on the SNS (top third) found the long table without the cone (Condition 2, $M = 3.09$) to be

TABLE 10 | Actual comprehension scores.

	Condition						
	1—Status quo cone	2—Long table, no cone	3—Short table, no cone	4—Map, no cone	5—Long table with cone	6—Short table with cone	7—Map with cone
M	1.88_{b}	$2.40_{a,b}$	2.51_{a}	$2.26_{a,b}$	$2.13_{a,b}$	2.48_{a}	$2.43_{a,b}$
SD	1.11	1.40	1.33	1.28	1.32	1.19	1.17

Note: Scale ranged from 0 = none correct to 4 = all correct. Means with different subscripts differ at $p < 0.05$ using Tukey HSD post hoc comparisons.

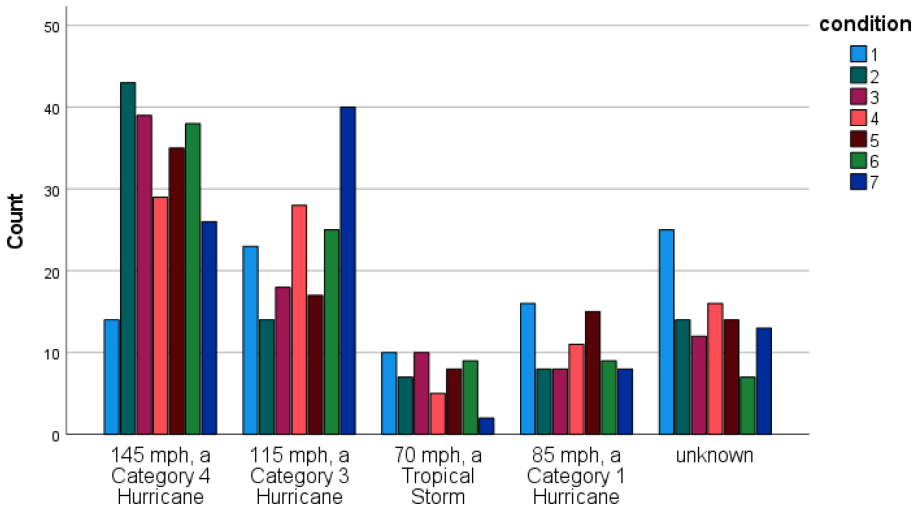


FIGURE 15 | Responses to the question "If Judy were to make landfall in West Florida (e.g., Tampa Bay area), the strongest intensity that could reasonably be expected at landfall is?" by condition. The correct answer is the leftmost response in the figure. Conditions 2, 3, and 6 prototypes differ from the status quo at $p < 0.01$ using Bonferroni correction.

TABLE 11 | Perceived message effectiveness (PME).

	Condition						
	1—Status quo cone	2—Long table, no cone	3—Short table, no cone	4—Map, no cone	5—Long table with cone	6—Short table with cone	7—Map with cone
M	4.07_{a}	$3.89_{a,b}$	4.02_{a}	4.03_{a}	3.79_{b}	$3.91_{a,b}$	4.09_{a}
SD	0.62	0.83	0.78	0.63	0.81	0.72	0.6

Note: Items ranged from 1 = strongly disagree to 5 = strongly agree. Means with different subscripts differ at $p < 0.05$ using LSD post hoc comparisons.

TABLE 12 | Risk perception.

	Condition						
	1—Status quo cone	2—Long table, no cone	3—Short table, no cone	4—Map, no cone	5—Long table with cone	6—Short table with cone	7—Map with cone
M	3.91	3.87	3.93	3.90	3.84	3.96	4.03
SD	0.75	0.81	0.78	0.72	0.72	0.72	0.76

Note: Items ranged from 1 = strongly disagree to 5 = strongly agree.

TABLE 13 | Numeracy × condition cross-tabulation.

	Condition							
Numeracy	1—Status quo cone	2—Long table, no cone	3—Short table, no cone	4—Map, no cone	5—Long table with cone	6—Short table with cone	7—Map with cone	Total
Low (< 3.65)	27	27	30	28	33	35	30	210
Med (3.65–4.63)	30	27	28	36	29	28	32	210
High (> 4.63)	31	32	29	25	27	25	27	196
Total	88	86	87	89	89	88	89	616

significantly more understandable than the status quo cone (Condition 1, $M = 2.00$); $F(6, 189) = 2.95$, $p = 0.009$, $\eta^2 = 0.086$.

On the other hand, those that scored lowest on the SNS (lowest third) perceived the color-coded map without a table but with a cone (Condition 7, $M = 3.70$) to be the most understandable (PC) and the long table without the cone (Condition 2) to be the least understandable ($M = 3.02$); $F(6, 203) = 2.03$, $p = 0.036$, $\eta^2 = 0.063$. The results for AC were not quite significant ($p = 0.054$) but similar to PC in that Condition 7 once again scored highest for those on the lowest end of the SNS. Interestingly, those that scored lowest on the SNS also found the color-coded map without a table but with a cone (Condition 7, $M = 4.01$) to be significantly more effective than the long table without the cone (Condition 2, $M = 3.42$); $F(6, 202) = 2.58$, $p = 0.020$, $\eta^2 = 0.071$. In short, these results indicate that those who prefer numbers found the additional numbers helpful, while those who do not prefer numbers found the lack of numbers to be helpful.

3.3.5 | Comparisons to Original Prototypes

To determine if the modifications made to the prototypes resulted in any actual improvements, the results of both surveys were combined, and a series of one-way ANOVAs was conducted. In terms of Perceived Comprehension, the status quo COU graphics scored highest overall ($M = 3.98$ in both surveys). However, the new version of the short table without a cone (Condition 3, $M = 3.88$) was the next highest and scored significantly higher than all versions of the long table, both old and new. Apart from the new version of the long table graphic with the cone (Condition 5), all the new prototypes scored significantly higher on AC than all conditions from Study 1, $F(11, 1662) = 5.75$; $p < 0.001$, $\eta^2 = 0.037$. The newly created color-coded map without a table but with a cone (Condition 7) scored highest on the risk perception index overall ($M = 4.03$); however, the results were not significant. In terms of message effectiveness, the new version of the long table graphic with the cone scored lowest overall (Condition 5, $M = 3.79$) and significantly lower than most of the other prototypes, both old and new, $F(11, 1661) = 1.89$; $p = 0.036$, $\eta^2 = 0.012$. To summarize, despite efforts to improve the long table, the participants in the second survey did not perceive it to be any more understandable or effective.

4 | Summary, Conclusions, and Future Research

RQ1 asked if the prototypes differ from the status quo in terms of message comprehension. As for Perceived Comprehension (PC), it is encouraging that all but one of the modified prototypes did not score significantly lower than the status quo. In other words, participants thought they understood most of the prototype forecast graphics in Study 2 almost as much as a forecast graphic that has been widely used and distributed for over two decades. That is an improvement over the original prototypes since only one of those was in the same league as the status quo COU in terms of PC. Considering that the COU is well known for being misinterpreted (ERG 2019), we cannot consider a statistical tie to be a win. However, there was a step in the right direction with respect to prototype design from Study 1 to Study 2 when compared with the status quo COU graphic.

Regarding Actual Comprehension (AC), the modifications made in Study 2 were successful in communicating the strongest reasonable intensity. Participants that viewed the modified short table, both with and without the cone (Conditions 3 and 6 in Study 2), scored significantly higher on AC than the status quo COU graphic. The AC index included a specific question about

the strongest reasonable intensity, and that question was far more likely to be answered correctly by someone who viewed one of the prototypes. That represents a marked improvement over the original prototypes that scored noticeably lower on that same set of questions. Given that AC of the strongest reasonable intensity is more likely to inform actual protective actions than PC, the improvement in AC of the prototypes between Study 1 and Study 2 is a very encouraging result.

RQ2 wondered if the prototypes differ from the status quo in terms of perceived message effectiveness. The only significant differences in the Perceived Message Effectiveness (PME) went in the undesired direction. The long table with the cone (Condition 5 in Study 2) scored significantly lower in PME than most other conditions. In the first study, there was not a significant difference in PME among any of the forecast graphics. In other words, the efforts to make the long table version of the graphic more effective only made it less effective, while the other graphics were not noticeably different.

RQ3 questioned if the prototypes differ from the status quo in terms of risk perception. Risk perception did not change from one round to the next but remained consistently high. That might represent a fault in the research design more so than a fault in the prototype graphics. The goal of introducing these prototype graphics is not to increase risk perception so much as it is to increase awareness. It is well known that people become "anchored" (Strack and Mussweiler 1997; Tversky and Kahneman 1974) to early TC forecasts (e.g., Drake 2012) and fail to realize that the situation may have changed. That is a common concern among NHC hurricane specialists that they are hoping to address (Berg et al. 2019; Eosco and Sprague-Hilderbrand 2020). It is also well documented that publics often dismiss a less intense TC as "*just* a tropical storm" or "*just* a category 1 hurricane" and do not recognize its potential significance (e.g., Ruin et al. 2008). The NHC is hoping that these prototypes will increase awareness that there is uncertainty in the intensity forecast and encourage their audience to stay engaged and informed. The average Risk Perception score across all conditions was 3.92, which indicates that the participants agreed all graphics depicted some risk. However, what we really want to know is do the participants realize that the level of risk can change. Therefore, the Risk Perception Index may not be asking the right questions. More meaningful results might be obtained from some of the engagement measures employed by Shulman et al. (2020) or other information-seeking scales in future studies.

RQ4 wondered what observations publics will make from the prototypes and what information is most salient. In Study 1, we learned that the probability of "Judy" reaching certain categories or intensities was ranked as the most important piece of information to Florida residents, followed closely by knowing that the TC was a threat to the state and knowing the strongest reasonable intensity at landfall. The cone of uncertainty is designed to highlight areas that might be at risk, but NHC does not currently have a forecast product that addresses the other top three most important pieces of information to Floridians. Our goal with this research was, in part, to test a product to fill that void. Through a combination of both studies, we learned the importance of also including the most likely intensity at landfall.

RQ5 asked if there were any demographic differences in the results; in fact, there were some. As should be expected based on previous literature, those directly in the path of "Judy" found it to be more relevant and more of a risk than those residing in other regions. Possibly because they found it more relevant and were therefore paying more attention, those directly in the path were also better able to answer questions about the forecast in Study 1. Interestingly, there were not any significant differences based on prior experience with hurricanes.

The research team did not expect the demographic results to change significantly based on the modifications that were made to the prototypes; so we did not repeat those tests in Study 2. Instead, the focus shifted to numeracy as the primary contributing factor to RQ5 in Study 2. As would be expected based on previous research (e.g., Toet et al. 2019; Tak et al. 2015), there were differences according to self-reported numeracy skills. In short, those who prefer numbers found additional numbers helpful in comprehending the forecast, while those who do not prefer numbers found the graphics without any numbers to be more effective. This result seems to confirm the suggestion from focus group participants that there may not be a one-size-fits-all solution to more effectively communicating the technical uncertainty in TC intensity forecasts.

Perhaps the most motivating result comes from the responses to a single question on the Subjective Numeracy Scale—"when you hear a weather forecast, do you prefer predictions using percentages (e.g., 'there will be a 20% chance of rain today') or predictions using only words (e.g., 'there is a small chance of rain today')." The results suggested an overwhelming preference for numbers among the Florida residents surveyed. Specifically, 34.2% of the participants responded, "always prefer percentages," 55% indicated they preferred some combination of words and percentages, and only 10.8% answered "always prefer words." This confirms prior studies (e.g., Rosen et al. 2021) showing that most people want some amount of numerical probability in their weather forecasts and provides a reason for further development in the area of probabilistic TC intensity forecast products.

Regarding future research, the original prototypes that included a longer table did not test well in Study 1. As mentioned previously, despite efforts to improve the long table, the participants in Study 2 did not perceive it to be any more understandable or effective. It would therefore be tempting to remove that version of the graphic from consideration entirely. However, the long table without the cone (Condition 2) scored significantly higher in terms of Actual Comprehension among those that scored highest on the Subjective Numeracy Scale. While the long table version of the graphic appears to be information overload for most people, no matter how it is presented, it might be exactly the right amount of information for certain people. Although the long table version of the graphic will therefore not be completely removed from consideration, future research will focus on making improvements to other versions of the graphic.

The research team has already assembled expert focus groups to solicit feedback; additional modifications are currently being made based on their recommendations. An obvious limitation of this current research is that the prototypes were tested solely in Florida and in English. A future study will require tests

of prototypes in hurricane-prone areas beyond Florida and in both English and Spanish. Additionally, we are optimistic about future tests regarding practical applications of the product in mock social media posts and television weather broadcasts. Toward the goal of more effectively communicating the technical uncertainty in TC intensity forecasts, there remains work to be done.

Author Contributions

Robert Eicher: writing – original draft, investigation, formal analysis, data curation, conceptualization. **Daniel J. Halperin:** conceptualization, project administration, writing – review and editing, software. **Benjamin C. Trabing:** software, writing – review and editing. **Derek Lane:** formal analysis, writing – review and editing, methodology. **Deanna Sellnow:** writing – review and editing, methodology. **Timothy Sellnow:** writing – review and editing, methodology. **Madison Croker:** investigation.

Acknowledgments

Funding for the survey data in Study 1 was provided by UCAR Subaward No. SUBAWD002943, an NWS Partners Project. Funding for the qualitative analysis that followed and the survey in Study 2 was provided by NOAA award NA22OAR4590185. Development of the LDP was funded by HFIP Award NA19OAR4320073.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Full survey instruments are provided in the Supporting Information. Per Institutional Review Board guidelines, only anonymized, aggregated data can be shared and will be made available upon request to the corresponding author.

Endnotes

[1] The standard measure of a tropical cyclone's intensity used by the National Hurricane Center is the maximum sustained surface wind, or highest one-minute average wind at an elevation of 10 m (see https://www.nhc.noaa.gov/aboutgloss.shtml).

[2] Rapid intensification is defined as an increase in the maximum sustained winds of a tropical cyclone of at least 30 kt in a 24-h period (see https://www.nhc.noaa.gov/aboutgloss.shtml).

References

Berg, R., J. Sprague-Hilderbrand, G. Eosco, and J. Schauer. 2019. "Tackling Some Smoldering Challenges of Communicating Hurricane Risks via the Social, Behavioral, and Economic Sciences." In *5th Conference on Weather Warnings and Communications*. AMS.

Bhatia, K. T., and D. S. Nolan. 2015. "Prediction of Intensity Model Error (PRIME) for Atlantic Basin Tropical Cyclones." *Weather and Forecasting* 30: 1845–1865.

Bica, M., J. L. Demuth, J. E. Dykes, and L. Palen. 2019. "Communicating Hurricane Risks: Multi-Method Examination of Risk Imagery Diffusion." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. Association for Computing Machinery.

Bostrom, A., R. Morss, J. K. Lazo, J. Demuth, and H. Lazrus. 2018. "Eyeing the Storm: How Residents of Coastal Florida See Hurricane Forecasts and Warnings." *International Journal of Disaster Risk Reduction* 30: 105–119.

Braun, V., and V. Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3: 77–101.

Broad, K., T. Leiserowitz, J. Weinkle, and M. Steketee. 2007. "Misinterpretations of the 'Cone of Uncertainty' in Florida During the 2004 Hurricane Season." *Bulletin of the American Meteorological Society* 88: 651–667.

Cangialosi, J. P., E. Blake, M. DeMaria, et al. 2020. "Recent Progress in Tropical Cyclone Intensity Forecasting at the National Hurricane Center." *Weather and Forecasting* 35: 1913–1922.

DeMaria, M., J. A. Knaff, M. J. Brennan, et al. 2013. "Improvements to the Operational Tropical Cyclone Wind Speed Probability Model." *Weather and Forecasting* 28: 586–602.

DeMaria, M., J. A. Knaff, R. Knabb, C. Lauer, C. R. Sampson, and R. T. DeMaria. 2009. "A New Method for Estimating Tropical Cyclone Wind Speed Probabilities." *Weather and Forecasting* 24: 1573–1591.

Demuth, J. L., R. E. Morss, J. K. Lazo, and C. Trumbo. 2016. "The Effects of Past Hurricane Experiences on Evacuation Intentions Through Risk Perception and Efficacy Beliefs: A Mediation Analysis." *Weather, Climate, and Society* 8: 327–344.

Drake, L. 2012. "Scientific Prerequisites to Comprehension of the Tropical Cyclone Forecast: Intensity, Track, and Size." *Weather and Forecasting* 27: 462–472.

Durbach, I. N., and T. J. Stewart. 2011. "An Experimental Study of the Effect of Uncertainty Representation on Decision Making." *European Journal of Operational Research* 214: 380–392.

Eicher, R., L. Taylor, and T. Brown. 2023. "Human or Machine? How Much Difference in Understanding and Trust Does a Human Element Make in Storm Forecasts?" *Electronic News* 17: 203–222. https://doi.org/10.1177/19312431231158120.

Eosco, G., and J. Sprague-Hilderbrand. 2020. "Accelerate Effective Risk Communication of Warnings." In *2020 Tropical Cyclone Operations and Research Forum (TCORF)/74th Interdepartmental Hurricane Conference, Lakeland, FL*. Harvard University.

Eosco, G., and C. Williamsberg. 2023. "Connecting the Dots Between NOAA's Hurricane Social Science Efforts: Using Hurricane Supplemental Triangulated Findings to Guide Research, Development, and Operations." In *2023 Tropical Cyclone Operations and Research Forum (TCORF)/77th Interdepartmental Hurricane Conference*. Harvard University.

ERG. 2019. *Cone of Uncertainty Social and Behavioral Science Research. NOAA BPA #EA-133C-14-BA-0040*. ERG.

Evans, S. D., K. Broad, A. Cairo, et al. 2022. "An Interdisciplinary Approach to Evaluate Public Comprehension of the 'Cone of Uncertainty' Graphic." *Bulletin of the American Meteorological Society* 103: E2214–E2221.

Fagerlin, A., B. J. Zikmund-Fisher, P. A. Ubel, A. Jankovic, H. A. Derry, and D. M. Smith. 2007. "Measuring Numeracy Without a Math Test: Development of the Subjective Numeracy Scale." *Medical Decision Making* 27: 672–680.

Franconeri, S. L., L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman. 2021. "The Science of Visual Data Communication: What Works." *Psychological Science in the Public Interest* 22: 110–161.

Gall, R., J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer. 2013. "The Hurricane Forecast Improvement Project." *Bulletin of the American Meteorological Society* 94: 329–343.

Gerst, M. D., M. A. Kenney, A. E. Baer, et al. 2020. "Using Visualization Science to Improve Expert and Public Understanding of Probabilistic Temperature and Precipitation Outlooks." *Weather, Climate, and Society* 12: 117–133.

Gustafson, A., and R. E. Rice. 2020. "A Review of the Effects of Uncertainty in Public Science Communication." *Public Understanding of Science* 29: 614–633.

Halvey, A. K. 2020. *Engaging With Uncertainty: Best Practices for Science Communication During the Climate Crisis and COVID-19*. University of Michigan. https://stpp.fordschool.umich.edu/research/policy-brief/engaging-uncertainty-best-practices-science-communication-during-climate.

Heale, R., and D. Forbes. 2013. "Understanding Triangulation in Research." *Evidence-Based Nursing* 16: 98.

Huang, S., M. K. Lindell, and C. S. Prater. 2016. "Who Leaves and Who Stays? A Review and Statistical Meta-Analysis of Hurricane Evacuation Studies." *Environment and Behavior* 48: 991–1029.

Jarell, J. D., B. M. Mayfield, E. N. Rappaport, and C. W. Landsea. 2021. *Hurricane FAQ*. AOML. https://www.aoml.noaa.gov/hrd-faq/.

Joslyn, S. L., and J. E. LeClerc. 2012. "Uncertainty Forecasts Improve Weather-Related Decisions and Attenuate the Effects of Forecast Error." *Journal of Experimental Psychology: Applied* 18: 126–140.

Kahan, D. M., E. Peters, M. Wittlin, et al. 2012. "The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks." *Nature Climate Change* 2: 732–735.

Knaflic, C. N. 2015. *Storytelling With Data: A Data Visualization Guide for Business Professionals*. 1st ed. Wiley.

Lin, S., J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. 2013. "Selecting Semantically-Resonant Colors for Data Visualization." *Computer Graphics Forum* 32: 401–410.

Lindner, B. L., F. Alsheimer, and J. Johnson. 2019. "Assessing Improvement in the Public's Understanding of Hurricane Storm Tides Through Interactive Visualization Models." *Journal of Coastal Research* 35: 130–142.

Liu, L., L. Padilla, S. H. Creem-Regehr, and D. H. House. 2019. "Visualizing Uncertain Tropical Cyclone Predictions Using Representative Samples From Ensembles of Forecast Tracks." *IEEE Transactions on Visualization and Computer Graphics* 25: 882–891.

Millet, B., A. P. Carter, K. Broad, A. Cairo, S. D. Evans, and S. J. Majumdar. 2020. "Hurricane Risk Communication: Visualization and Behavioral Science Concepts." *Weather, Climate, and Society* 12: 193–211.

Millet, B., S. J. Majumdar, A. Cairo, B. D. McNoldy, S. D. Evans, and K. Broad. 2024. "Exploring the Impact of Visualization Design on Non-Expert Interpretation of Hurricane Forecast Path." *International Journal of Human-Computer Interaction* 40, no. 2: 425–440.

Morrow, B. H., J. K. Lazo, J. Rhome, and J. Feyen. 2015. "Improving Storm Surge Risk Communication: Stakeholder Perspectives." *Bulletin of the American Meteorological Society* 96: 35–48.

National Hurricane Center. 2024. *National Hurricane Center Product Description Document: A User's Guide to Hurricane Products*. NHC. https://www.nhc.noaa.gov/pdf/NHC_Product_Description.pdf.

Noar, S. M., P. Palmgreen, R. S. Zimmerman, M. L. A. Lustria, and H. Lu. 2010. "Assessing the Relationship Between Perceived Message Sensation Value and Perceived Message Effectiveness: Analysis of PSAs From an Effective Campaign." *Communication Studies* 61: 21–45.

Peters, E., D. Västfjäll, P. Slovic, C. K. Mertz, K. Mazzocco, and S. Dickert. 2006. "Numeracy and Decision Making." *Psychological Science* 17: 407–413.

Ripberger, J., A. Bell, A. Fox, et al. 2022. "Communicating Probability Information in Weather Forecasts: Findings and Recommendations From a Living Systematic Review of the Research Literature." *Weather, Climate, and Society* 14: 481–498.

Rosen, Z., M. J. Krocak, J. T. Ripberger, et al. 2021. "Communicating Probability Information in Hurricane Forecasts: Assessing Statements That Forecasters Use on Social Media and Implications for Public Assessments of Reliability." *Journal of Operational Meteorology* 9, no. 7: 89–101.

Ruginski, I. T., A. P. Boone, L. M. Padilla, et al. 2016. "Non-Expert Interpretations of Hurricane Forecast Uncertainty Visualizations." *Spatial Cognition & Computation* 16, no. 2: 154–172.

Ruin, I., C. League, M. Hayden, B. Goldsmith, and J. Estupiñán. 2008. *Differential Social Vulnerability and Response to Hurricane Dolly Across the US-Mexico Border*. Weather. https://www.weather.gov/media/bro/research/pdf/Hurricane_Dolly_AMSannual2009_Final.pdf.

Schott, T. 2012. *The Saffir–Simpson Hurricane Wind Scale. NOAA/NHC Technical Report*. NOAA. https://www.nhc.noaa.gov/pdf/sshws.pdf.

Sellnow, D. D., J. Iverson, and T. L. Sellnow. 2017. "The Evolution of the Operational Earthquake Forecasting Community of Practice: The L'Aquila Communication Crisis as a Triggering Event for Organizational Renewal." *Journal of Applied Communication Research* 45: 121–139.

Sellnow, D. D., L. M. Jones, T. L. Sellnow, P. Spence, D. R. Lane, and N. Haarstad. 2019. "The IDEA Model as a Conceptual Framework for Designing Earthquake Early Warning (EEW) Messages Distributed via Mobile Phone Apps." In *Earthquakes-Impact, Community Vulnerability and Resilience*, edited by J. Santos-Reyes. IntechOpen.

Sellnow, D. D., D. Lane, R. S. Littlefield, et al. 2015. "A Receiver-Based Approach to Effective Instructional Crisis Communication." *Journal of Contingencies & Crisis Management* 23: 149–158.

Sellnow, D. D., and T. L. Sellnow. 2019. "The IDEA Model for Effective Instructional Risk and Crisis Communication by Emergency Managers and Other Key Spokespersons." *Journal of Emergency Management* 17: 67–78.

Sellnow, T. L., M. W. Seeger, and R. R. Ulmer. 2002. "Chaos Theory, Informational Needs, and Natural Disasters." *Journal of Applied Communication Research* 30: 269–292.

Shulman, H. C., G. N. Dixon, O. M. Bullock, and D. Colón Amill. 2020. "The Effects of Jargon on Processing Fluency, Self-Perceptions, and Scientific Engagement." *Journal of Language and Social Psychology* 39: 579–597.

Strack, F., and T. Mussweiler. 1997. "Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility." *Journal of Personality and Social Psychology* 73: 437–446.

Tak, S., A. Toet, and v. E. Jan. 2015. "Public Understanding of Visual Representations of Uncertainty in Temperature Forecasts." *Journal of Cognitive Engineering and Decision Making* 9: 241–262.

Toet, A., J. B. v. Erp, M. B. Erik, and v. B. Stef. 2019. "Graphical Uncertainty Representations for Ensemble Predictions." *Information Visualization* 18: 373–383.

Trabing, B. C., and M. M. Bell. 2020. "Understanding Error Distributions of Hurricane Intensity Forecasts During Rapid Intensity Changes." *Weather and Forecasting* 35: 2219–2234. https://doi.org/10.1175/WAF-D-19-0253.1.

Trabing, B. C., K. D. Musgrave, M. DeMaria, B. C. Zachry, M. J. Brennan, and E. N. Rappaport. 2023. "The Development and Evaluation of a Tropical Cyclone Probabilistic Landfall Forecast Product." *Weather and Forecasting* 38: 1363–1374. https://doi.org/10.1175/WAF-D-22-0199.1.

Trypke, M., F. Stebner, and J. Wirth. 2023. "Two Types of Redundancy in Multimedia Learning: A Literature Review." *Frontiers in Psychology* 14: 1148035. https://doi.org/10.3389/fpsyg.2023.1148035.

Tversky, A., and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science American Association for the Advancement of Science* 185: 1124–1131.

Wallsten, T. S., D. V. Budescu, A. Rapoport, R. Zwick, and B. Forsyth. 1986. "Measuring the Vague Meanings of Probability Terms." *Journal of Experimental Psychology. General* 115: 348–365.

Wenger, E. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

Whitmer, D. E., and V. K. Sims. 2021. "An Implicit—Not Explicit—Understanding of Hurricane Storms." *Weather, Climate, and Society* 13: 1043–1053.

Windschitl, P. D., and E. U. Weber. 1999. "The Interpretation of 'Likely' Depends on the Context, but '70%' is 70%-Right? The Influence of Associative Processes on Perceived Certainty." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 25: 1514–1533.

Zikmund-Fisher, B. J., D. M. Smith, P. A. Ubel, and A. Fagerlin. 2007. "Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations." *Medical Decision Making* 27: 663–671.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Study 1 survey. **Data S2:** Study 2 survey.