The Effects of Consistency among Simultaneous Forecasts on Weather-Related Decisions

CHEN SU,[a] JESSICA N. BURGENO,[a] AND SUSAN JOSLYN[a]

[a] *University of Washington, Seattle, Washington*

(Manuscript received 8 July 2019, in final form 6 August 2020)

ABSTRACT: People access weather forecasts from multiple sources [mobile telephone applications (''apps''), newspapers, and television] that are not always in agreement for a particular weather event. The experiment reported here investigated the effects of inconsistency among forecasts on user trust, weather-related decisions, and confidence in user decisions. In a computerized task, participants made school-closure decisions on the basis of snow forecasts from different sources and answered a series of questions about each forecast. Inconsistency among simultaneous forecasts did not significantly reduce trust, although inaccuracy did. Moreover, inconsistency may convey useful information to decision-makers. Not only do participants appear to incorporate the information provided by all forecasts into their own estimates of the outcome, but our results also suggest that inconsistency gives rise to the impression of greater uncertainty, which leads to more cautious decisions. The implications for decisions in a variety of domains are discussed.

KEYWORDS: Social Science; Forecasting; Decision support

1. Introduction

Information dissemination has changed drastically over the past few decades as a result of many innovative technological developments. The increased accessibility of a wide range of information over the internet and via ''smart'' cellular telephones enables people to access multiple sources of information on any topic, at any moment, changing how they both gather and evaluate decision-relevant information. This is especially true in the domain of weather, for which the information contained in forecasts can influence not only everyday decisions but also critical decisions related to personal safety. However, making weather forecast information more accessible to the public does not necessarily lead to better public understanding. Although mobile-device-based weather applications (''apps'') provide abundant and timely forecasts, few provide guidance on how people should interpret and use the information (Zabini 2016).

In addition, there can be inconsistency in forecasts for the same target event among multiple providers, which may lead to confusion (Weyrich et al. 2019) or reduce trust. Thus, inconsistency could make users hesitant to act upon the information contained in inconsistent forecasts. Indeed, a survey conducted after Hurricane Katrina suggested that inconsistency in evacuation orders contributed to reluctance to evacuate among African Americans (Elder et al. 2007). As a consequence, the maintenance of consistency in forecasts and warning communications is considered important by many (Perry and Green 1982; Quarantelli 1984; Drabek 1999; NOAA 2016) for several reasons, including the maintenance of user trust. Although there are many different kinds of trust (Twyman et al. 2008), here we refer to *calculative* trust, using a construct similar to that defined by Earle (2010), which reflects forecasters' past performance, abilities, or knowledge.[1]

At present, however, the impact of inconsistency on user trust, particularly in the context of weather, is not well understood because there is little experimental evidence addressing this issue. There is some evidence for a reduction in trust when there is inconsistency among sequential weather forecasts from the same source (Losee and Joslyn 2018), but other evidence suggests that it is much less than the reduction of trust due to inaccuracy (Burgeno and Joslyn 2020). It is important to note however, that both of these studies tested the impact of inconsistency in *sequential* forecasts from the *same* source, which may well be different than the impact of inconsistency in *simultaneous* forecasts from *different* sources. At present, there is essentially no existing experimental work of which we are aware on this issue. In other words, no research answers the following question: Must a forecast provider agree with others to be considered trustworthy?

There are a few studies on a related issue examining the integration of information from multiple different sources to make decisions in the financial domain. In one study, confidence in participants' own decisions was reduced when a decision was based on multiple financial experts whose opinions differed (Budescu and Rantilla 2000). In another study, participants were more confident when experts agreed with one another *or* when they were described as more accurate (Budescu et al. 2003). Moreover, participants did not simply rely on a single expert, but instead averaged the predictions from all available sources (Budescu et al. 2003). This research suggests that even though people incorporate all of the available information, they are less confident when there is

[1] Although, unlike Earle (2010), we do not refer to calculative trust as ''confidence,'' which we regard as a separate construct. Here we define confidence with respect to participants' own decisions, based on predictions.

Corresponding author: Jessica N. Burgeno, jburgeno@uw.edu

DOI: 10.1175/WCAS-D-19-0089.1

inconsistency, which may be due to a reduction in trust in the prediction, although trust was not measured directly. In sum, direct evidence for the positive impact of consistent predictions from separate sources (for the same event) on trust is lacking.

In addition, there may be some critical downsides to maintaining forecast consistency. Some providers may have access to better information, for instance more recent model runs, which tend to be more accurate (Lazo et al. 2009; Wilson and Giles 2013). Similarly, some providers may have a greater familiarity with local weather patterns, which can also increase accuracy (Joslyn and Jones 2008). Thus, maintaining consistency with other less well-informed sources might constitute a sacrifice in accuracy under some circumstances. It is clear that inaccuracy decreases trust. An abundance of evidence suggests that inaccuracy in weather forecasts reduces credibility (Ripberger et al. 2015) and has a persisting negative effect on users' trust even after forecast accuracy increases (Joslyn and LeClerc 2012). Similarly, participant investors rated higher competence and trustworthiness in accurate as compared to inaccurate financial analysts and were more likely to purchase future reports from them (Kadous et al. 2009). Mammography patients asked to imagine receiving a false positive breast cancer test result indicated diminished trust and greater likelihood of delaying future mammography relative to patients who imagined receiving accurate test results (Kahn and Luce 2003). In sum, the negative effect of inaccuracy on trust is well established.

In addition, artificially maintaining consistency may deprive users of information that could inform their decisions. For instance, people may interpret inconsistent information as indicating inherent uncertainty in the weather situation (Pappenberger et al. 2011) leading them to be more cautious as it has in other domains (Bloom et al. 2007). Indeed, there is evidence that everyday users have intuitive understanding of numerous principles related to weather forecast uncertainty such as this (Joslyn and Savelli 2010; Morss et al. 2008). Therefore, inconsistency could engender distrust, causing people to rely less on forecast information, or it could be beneficial, alerting people to the inherent uncertainty of some weather situations.

The goal of the research presented here was to determine how people respond, in a simple weather-related decision task, to inconsistency in multiple simultaneous weather forecasts from different sources. We test whether inconsistency reduces trust and if so, how it compares to the reduction in trust due to inaccuracy. To answer these questions, we used a laboratory-based computerized task in which participants were charged with deciding, on the basis of two simultaneously presented forecasts, whether to advise area schools to close because of snowstorms that could lead to dangerous road conditions. We systematically manipulated consistency between the forecasters and the accuracy of the target forecaster, to test the impact of these variables on trust and closure decisions. We also measured confidence in participants' own decisions and uncertainty operationalized as the range of possible outcomes expected. Last, we asked whether participants incorporated information from both forecasts (e.g., averaging them) when making their own estimates.

2. Method

a. Participants

A total of 349 University of Washington psychology students participated in exchange for course credit and the opportunity to earn a cash bonus. The average age of participants was 18.6 years old, and 66% of them were female. In general, the majority of this population has some experience with weather hazards including snow, as well as experience using forecasts to make decisions. They consult weather forecasts every day or nearly every day.[2]

b. Procedure

In this computer-based task, programmed in Excel Visual Basic and administered on standard desktop computers, participants monitored multiple sources of weather forecast information to make snow-based school-closure decisions. The participants' task was a simplified version of that performed by school administrators. In fact, according to prestudy interviews with administrators, several other factors are considered when making school-closure decisions such as road conditions, public transit operations, impacts on student life and school operations, and whether other schools in the district are closing. However, in this experiment, the decision was based on snow accumulation forecasts alone.

Participants first gave informed consent and provided their age and gender. Then they read instructions describing the task and performed four practice trials under the guidance of the experimenter. Participants were told to advise closing school if they expected 6 or more inches of snow (1 in. = 2.54 cm) accumulation. Although this threshold is realistic in general, in fact, the threshold varies depending on the location and other conditions specific to the situation. Thus, for this experiment, participants were given a simplified decision rule with a single threshold. Participants were to provide school-closure decisions over two hypothetical winter seasons, each with 12 weeks. A different school district was affected each week.

For each of the 24 trials, participants based their school-closure decision on two snow accumulation forecasts for the following day (Wednesday) provided by two different forecasters simultaneously. In all, there were eight fictitious forecast providers: TruWeather, Weather Now, Weather Direct, Weather Radar, Sky Watch, WeatherPro, Weather Bug, and AccuCast. Although a pilot study showed no significant difference in trust resulting from provider name alone, they were randomized across blocks and across participants. There were four blocks of six consecutive trials in which the same set of two forecast providers supplied forecasts within a block. Before each new block, participants were notified of the new pair of providers' names. Each forecast provider name was used once per participant.

To motivate participants and standardize decision goals, a point system was used. Participants began the task with a virtual budget of 120 points. Their goal was to retain as many of those points as possible by giving the best advice. A school-closure

[2] This is based on an unpublished survey of undergraduates from the University of Washington conducted in 2019.

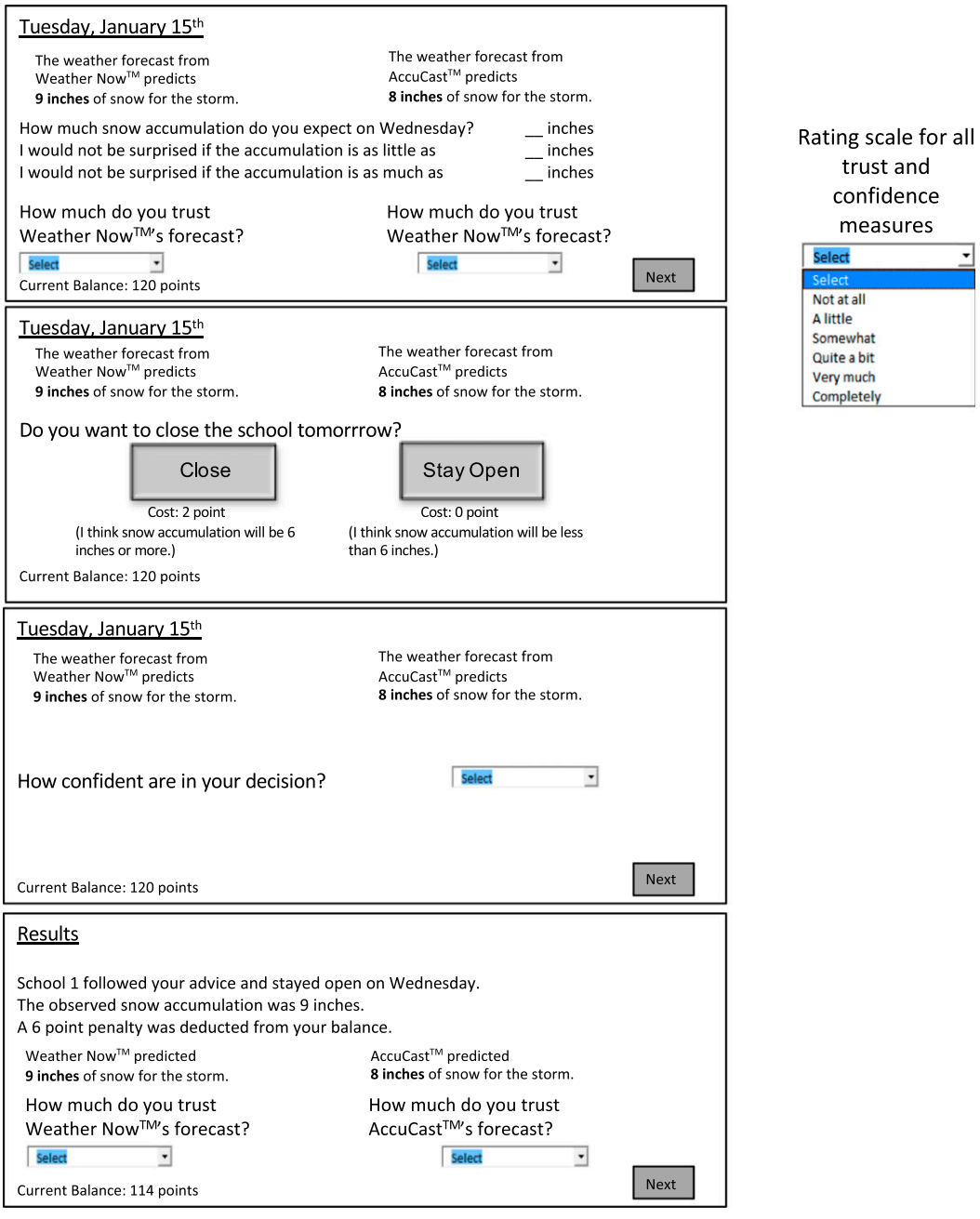

FIG. 1. Screens shown in a single trial in order from top to bottom.

recommendation cost 2 points to reflect the cost of makeup days. There was no cost per se for recommending that a school stay open; however, if 6 or more inches of snow accumulation were observed, a 6-point penalty was deducted from their budget to reflect the risk of damage and injuries. At the end of the experiment, cash was awarded for final budget balances over 72 points at the rate of $1 for every 4 points. This payment threshold was chosen to avoid rewarding the unrealistic strategy of recommending school closure for every trial, which would result in a balance of 72 points.

Each trial began with a screen representing a Tuesday time period that showed two snow-accumulation forecasts for a Wednesday storm (see Fig. 1). One forecast was presented at the top left of the screen, and the other at the top right. On the same screen, participants indicated the number of inches of snow accumulation they expected for Wednesday as well as the least and greatest number of inches that they would not be surprised to observe. These questions were presented in the middle of the screen below the two forecasts. At the bottom of the same screen, participants rated their trust in each

TABLE 1. Forecast and outcome values for all experimental (expt) and filler trials in the consistency blocked condition (the same trials were used in the unblocked condition).

Block	Consistency	Accuracy	Forecaster 1	Forecaster 2	Outcome	Type
A	Consistent	Accurate	*6*	6	6	Expt
	Consistent	Accurate	*5*	5	5	Expt
	Consistent	Accurate	*3*	3	3	Filler
	Consistent	Inaccurate	*9*	9	8	Filler
	Consistent	Inaccurate	*7*	7	5	Expt
	Consistent	Inaccurate	*4*	4	6	Expt
B	Inconsistent	Accurate	6	4	6	Expt
	Inconsistent	Accurate	5	7	5	Expt
	Inconsistent	Accurate	3	2	3	Filler
	Inconsistent	Inaccurate	9	8	8	Filler
	Inconsistent	Inaccurate	7	5	5	Expt
	Inconsistent	Inaccurate	4	6	6	Expt
C	Consistent	Accurate	7	7	7	Expt
	Consistent	Accurate	4	4	4	Expt
	Consistent	Accurate	9	9	9	Filler
	Consistent	Inaccurate	3	3	2	Filler
	Consistent	Inaccurate	6	6	4	Expt
	Consistent	Inaccurate	5	5	7	Expt
D	Inconsistent	Accurate	7	5	7	Expt
	Inconsistent	Accurate	4	6	4	Expt
	Inconsistent	Accurate	9	8	9	Filler
	Inconsistent	Inaccurate	3	2	2	Filler
	Inconsistent	Inaccurate	6	4	4	Expt
	Inconsistent	Inaccurate	5	7	7	Expt

individual forecast on a 6-point drop-down menu from ''not at all'' to ''completely.'' These questions were presented at the bottom left and bottom right of the screen directly beneath the forecast to which they referred. Next, participants saw a screen on which they indicated their school-closure advice by clicking the ''close'' or ''stay open'' button presented at the center of the screen. To ensure that participants remembered the threshold for closure, text below each button stated that close meant, ''I think the snow accumulation will be 6 in. or more'' and that stay open meant, ''I think the snow accumulation will be less than 6 in.'' On the next screen, participants rated their confidence in their decision on a 6-point drop-down menu from ''Not at all'' to ''Completely.'' The two forecasts from the previous screen remained in the same position as a reminder for participants.

After making their school-closure decisions, participants learned that the school followed their advice. On the same screen, they were informed of the observed snow accumulation on Wednesday as well as the subsequent cost or penalty (if any), on the line below. The two forecasts remained in the same position on the screen as a reminder, and participants again rated their trust in each one on the same 6-point drop-down menu. The current point balance was displayed in the lower-left corner of all the screens.

In sum, each trial consisted of four screens: 1) Tuesday forecasts by two forecasters and trust ratings for each, 2) Tuesday-night school-closure decision, 3) Tuesday-night confidence-in-decision rating, and 4) Wednesday outcome and trust ratings for each forecast. Thus, participants reported four trust ratings per trial: two trust ratings (one for each forecast) when they were first given the forecasts and two trust ratings (one for each forecast) when they learned the outcome, for a total of 96 trust ratings. We also collected an additional trust rating for each of the forecasters in the final block, at the end of the experiment where the final point balance was displayed.

c. *Stimuli*

The forecasts and observed snow accumulations were modeled on realistic values for Seattle, Washington (24-h snowfall: $M = 1.38$ in., min = 0.1 in., max = 6.8 in.; NCEI 2019),[3] where the experiment was conducted. Table 1 shows the forecasts and observed snow accumulations (outcomes) for the 24 trials. Because it was possible to control all relevant extraneous variables for only one of the two forecasters, only responses to that forecaster (forecaster 1) were analyzed below, although participants were not aware of this focus. In the experiment, the screen position (left or right) of the target forecaster was counterbalanced to neutralize any potential right- or left-side bias.

Accuracy was defined as an exact match between the forecaster-1 prediction and the observed accumulation (Table 1). By this definition, one-half of trials within each block was accurate and one-half was inaccurate. All inaccurate experimental trials were inaccurate by 2 in. and crossed the 6-in. closure threshold. In one-half of the inaccurate experimental trials (defined in terms of forecast 1), forecaster 2's prediction was also inaccurate by 2 in. and the inaccuracy crossed the 6-in. closure threshold. One-half of inaccurate experimental trials in

[3] These data are based on the last 30 years of snowfall data from NOAA's Seattle Tacoma International Airport station.

each block were misses, in which the forecaster-1 prediction was below the 6-in. decision threshold and the observed accumulation was at or above the threshold. The other half were false alarms (FA) in which the forecaster-1 prediction was above the 6-in. decision threshold and the observed accumulation was below the threshold. Similarly, one-half of accurate experimental trials in each block were correct rejections (CR), in which both the forecaster-1 prediction and the observed accumulation values were below the 6-in. decision threshold. One-half of accurate experimental trials in each block were hits, in which the forecaster-1 prediction and the observed accumulation values were above the threshold.

One-half of experimental trials were consistent, and one-half were inconsistent (Table 1). Consistency was defined as an exact match between the prediction of forecaster 1 and forecaster 2. All inconsistent experimental trials were inconsistent by a 2-in. discrepancy that crossed the 6-in. closure threshold. To determine whether the impact of consistency built up over trials, in one between-group condition (consistency blocked), blocks of six trials included only consistent or only inconsistent forecasts. In the other condition (consistency unblocked) one-half of forecasts in each six-trial block were consistent and one-half were inconsistent. The same trials were used in both blocking conditions. Trial order was randomized within a block.

While these constraints may detract from ecological validity, tight control of all of the relevant variables allows for a direct comparison between accuracy and consistency of the same magnitude. However, the debriefing portion of a pilot study revealed that participants were on the verge of recognizing patterns in the forecasts and observations resulting from manipulating and controlling for primary variables. To distract from these patterns, two filler trials were added to each block. Filler trials were inaccurate by a 1-in. discrepancy between the forecaster-1 prediction and the observed accumulation and did not cross the 6-in. closure threshold. Filler-trial forecast values were either lower (2 or 3 in.) or higher (8 or 9 in.) than values for experimental trials, which hovered around the 6-in. threshold (4–7 in.). Therefore, each block contained six trials: four experimental trials and two filler trials. Filler trials were not analyzed below because they were not subject to the same controls as the experimental trials. Nor was the final trust rating analyzed because the main independent variables were manipulated within groups and affected this trust rating equally. Instead it serves as a reference point (forecaster 1: mean $M = 2.99$, std dev = 0.85, min = 1, and max = 5; forecaster 2: $M = 2.99$, std dev = 0.90, min = 1, and max = 5).

d. Design

The experiment used a 2 (accuracy) × 2 (consistency) × 2 (blocking) mixed-model design. Accuracy and consistency were both within-groups variables with two levels each, accurate and inaccurate, and consistent and inconsistent, respectively. Blocking was a between-groups variable with two levels: blocked consistency and unblocked consistency.

3. Results

Our primary goal was to determine the impact of accuracy and consistency on trust in the forecast. Therefore, in the analyses reported below, we first examined trial-by-trial postoutcome trust ratings. Then we examined confidence in decisions comparing consistent to inconsistent trials to determine whether we replicated the previously observed impact of consistency on confidence. Next, we examined participants' snow accumulation estimates to ascertain the contribution of each forecast to participants' understanding of the situation as well as to their impression of the degree of uncertainty. Last, we examined participants' decisions to determine how these factors impacted cautiousness. All analyses use an alpha level of 0.05. Effect sizes are reported as partial eta squared and Cohen's *d*.

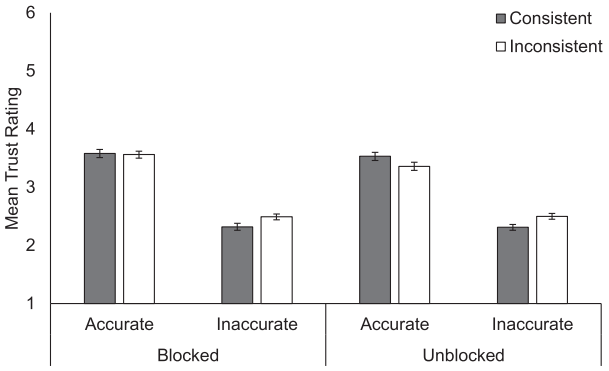

FIG. 2. Mean postoutcome trust rating by consistency between forecasts and accuracy of forecaster 1.

To determine the impact of forecast consistency and accuracy on trust, we first examined trial-by-trial postoutcome trust ratings. A mixed-model analysis of variance (ANOVA) was conducted on mean trust rating with the within-groups independent variables accuracy (accurate/inaccurate) and consistency (consistent/inconsistent) and the between-groups independent variable, blocking (blocked/unblocked). Trust rating (1–6 scale) was significantly higher for accurate forecasts ($M = 3.51$; std dev = 0.84) than for inaccurate forecasts [$M = 2.40$, std dev = 0.64), $F(1, 347) = 677.00$, $p < 0.001$, and $\eta_p^2 = 0.66$—a very large effect size (Cohen 1988)]. Although the main effect of consistency did not reach significance [$F(1, 347) = 3.33$, $p = 0.069$, and $\eta_p^2 = 0.01$], the interaction between accuracy and consistency was significant [$F(1, 347) = 45.23$, $p < 0.001$, and $\eta_p^2 = 0.12$]. When forecaster 1's prediction was inaccurate, surprisingly, postoutcome trust ratings were higher for *in*consistent forecasts ($M = 2.49$; std dev = 0.66) than for consistent forecasts ($M = 2.31$; std dev = 0.73), contrary to our hypothesis. There was also a significant accuracy × consistency × blocking interaction [$F(1, 347) = 4.05$, $p = 0.045$, and $\eta_p^2 = 0.01$], suggesting that only when forecaster 1 was accurate in the unblocked condition was the predicted negative effect of inconsistency observed: trust was lower for inconsistent forecasts ($M = 3.36$; std dev = 0.89) than for consistent forecasts ($M = 3.53$, std dev = 0.96) (see Fig. 2). Thus, forecast accuracy had a greater impact on user trust than did forecast consistency.

However, the effect of consistency might be stronger prior to learning the outcome while participants were making their decisions. To examine this effect, a mixed-model ANOVA was conducted on participants' preoutcome trust with the within-groups

variable consistency (consistent/inconsistent) and the between-groups variable blocking (blocked/unblocked). Because participants were unaware of the outcome when they reported this trust rating, accuracy was not included in this analysis. Confirming our hypothesis, mean trust was significantly higher for consistent forecasts ($M = 3.11$; std dev = 0.75) than for inconsistent forecasts ($M = 2.95$; std dev = 0.65) [$F(1, 347) = 40.50$, $p < 0.001$, and $\eta_p^2 = 0.11$]. However, notice that this effect size was much smaller than that of accuracy in the postoutcome analysis. No other effects reached significance.

The contrast between the pre- and postoutcome trust analyses reported above suggests that trust changes when the outcome is learned. To further examine this effect, the difference between pre- and postoutcome trust was calculated. A positive score indicates an increase in trust, while a negative score indicates a decrease in trust after learning the outcome of the target event. A mixed-model ANOVA was conducted on the mean difference score with the within-groups independent variables accuracy (accurate/inaccurate) and consistency (consistent/inconsistent) and the between-groups independent variable, blocking (blocked/unblocked). There was a significant main effect of accuracy [$F(1, 347) = 853.15$, $p < 0.001$, and $\eta_p^2 = 0.71$] such that trust increased postoutcome for accurate forecasts ($M = 0.62$; std dev = 0.53) and decreased for inaccurate forecasts ($M = -0.77$; std dev = 0.51). There was a significant main effect of consistency [$F(1, 347) = 84.86$, $p < 0.001$, and $\eta_p^2 = 0.20$] such that trust increased postoutcome for inconsistent forecasts ($M = 0.03$; std dev = 0.33) and decreased for consistent forecasts ($M = -0.18$; std dev = 0.35). We return to this issue in the discussion. The interaction between accuracy and consistency was also significant [$F(1, 347) = 89.61$, $p < 0.001$, and $\eta_p^2 = 0.21$], indicating that, after learning the outcome, there was a similar increase in trust for consistent and inconsistent forecasts when forecaster 1 was accurate. However, when forecaster 1 was inaccurate, there was a larger decrease in trust for consistent forecasts than for inconsistent forecasts (see Fig. 3). Again, there was no significant main effect for the between-groups factor blocking [$F(1, 347) = 0.03$, $p = 0.874$, and $\eta_p^2 = 0.00$]. Nor was the interaction of accuracy × consistency × blocking significant [$F(1, 347) = 0.39$, $p = 0.393$, and $\eta_p^2 = 0.001$].

To examine whether consistency had an effect on participants' confidence in their school-closure decisions, we conducted a mixed-model ANOVA on mean confidence ratings with the within-groups independent variable consistency (consistent/inconsistent) and the between-groups variable blocking (blocked/unblocked). Accuracy was not included because the confidence ratings were measured after the decision but before the outcome was presented. There was a significant main effect of consistency [$F(1, 347) = 165.98$, $p < 0.001$, and $\eta_p^2 = 0.32$] such that mean confidence was significantly higher for consistent forecasts ($M = 3.27$; std dev = 0.86) than for inconsistent forecasts ($M = 2.93$; std dev = 0.77)—a large effect size (Cohen 1988). No other effects reached significance.

This set of analyses suggests that, although consistency has little impact after the outcome is known, it does affect both trust in the forecast and confidence in one's decision prior to learning the outcome. Therefore, it might have an impact on participants' understanding of the weather situation as well as the decision they make.

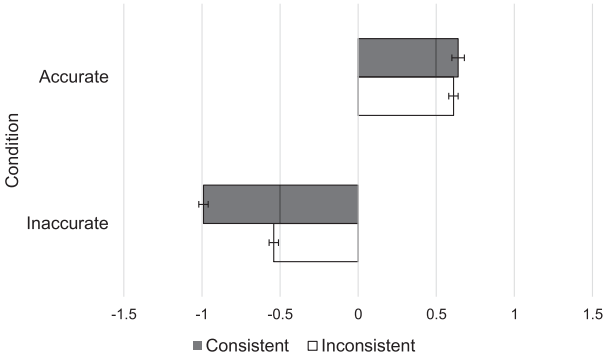

FIG. 3. Mean change in trust rating from pre- to postoutcome by consistency between forecasts and accuracy of forecaster 1.

We next examined whether participants' estimates of snow accumulation were affected by both forecasters. We hypothesized that when the forecasts were inconsistent, participants' estimates would represent an average of the two predictions as had been seen in previous research (Budescu et al. 2003). Therefore, we subtracted participants' snow estimate from the mean of the two forecast values (a single value in the consistent condition) for each trial and calculated an absolute mean difference score for both the consistent and inconsistent conditions. Although the differences from the mean of forecasts were small, less than 0.5 in., they were significantly different from 0 in both conditions [consistent, $t(348) = 10.82$, with $p < 0.001$; inconsistent, $t(348) = 18.39$, with $p < 0.001$]. Moreover, a paired-samples t test revealed that participant estimates differed significantly less from the forecast mean in the consistent condition ($M = 0.15$; std dev = 0.26) than in the inconsistent condition ($M = 0.27$; std dev = 0.28) [$t(348) = 7.93$, with $p < 0.001$; Cohen's $d = 0.43$].

To determine whether there was a bias suggesting that participants either over- or underestimated snow accumulation relative to the forecasts provided, the signed values of the difference between the estimate and the mean value were then calculated. A paired-samples t test revealed that the difference was slightly lower than the mean forecast value in the consistent condition ($M = 0.01$; std dev = 0.18) and slightly higher in the inconsistent condition ($M = -0.04$; std dev = 0.24) [$t(348) = 3.79$, with $p < 0.001$; Cohen's $d = 0.24$]. However, two 1-sample t tests (one for consistent trials and one for inconsistent trials) demonstrate that the mean difference was only significantly different than zero in the inconsistent condition [$t(348) = 3.42$, with $p = 0.001$], suggesting that, rather than averaging the two forecasts in the inconsistent condition, participants were weighting the larger value slightly more.

It is important to note that the forecast values from forecasters 1 and 2 were identical in the consistent and inconsistent conditions (the same forecast values were presented the same number of times by each forecaster within each block; see Table 1). Therefore, we can be confident that the observed effect is due to consistency alone rather than a difference in the forecast values. Thus, participants' own estimates varied to a

greater degree from the forecast mean and were slightly higher when forecasts were inconsistent than when they were consistent.

This might be due in part to a perception of greater uncertainty when forecasts are inconsistent. To determine the amount of uncertainty that participants expected, we subtracted participants' least number from the greatest number of inches of snow accumulation that they would not be surprised to observe for each trial to create a range of values. A paired-sample t test revealed that the range was significantly wider in the inconsistent condition ($M = 3.18$; std dev = 0.07) than in the consistent condition ($M = 2.71$; std dev = 0.07) [$t(348) = 10.51$, with $p < 0.001$; Cohen's $d = 0.37$], suggesting greater perceived uncertainty in inconsistent forecasts. Because the forecast values were identical in the consistent and inconsistent conditions, we can be confident that this effect is due to consistency alone.

A perception of greater uncertainty and a high bias in expected snow accumulation might translate into more cautious decisions. To determine whether this was the case, we conducted an ANOVA on mean proportion of closure decisions with two independent variables, consistency (consistent/inconsistent) and participants' snow accumulation estimates, coded as either above or below the 6-in. closure threshold. Indeed, participants advised closing schools significantly more often when forecasts were inconsistent ($M = 0.63$; std dev = 0.23) than when they were consistent ($M = 0.57$; std dev = 0.15) [$F(1, 340) = 29.11$, $p < 0.001$, and $\eta_p^2 = 0.08$]. In addition, participants closed schools more when their estimates were above the threshold ($M = 0.90$; std dev = 0.30) than when their estimates were below the threshold ($M = 0.28$; std dev = 0.45) [$F(1, 340) = 2172.02$, $p < 0.001$, and $\eta_p^2 = 0.87$]. Thus, although participants were clearly using the 6-in. threshold to make their decision, they were more cautious, closing more often when forecasts were inconsistent.

Participants ended the task with an average final balance of 72.28 points (std dev = 5.68). About 30% of participants earned the cash bonus. The final point balances of participants who experienced blocked trials ($M = 72.48$; std dev = 6.01) and those who experienced unblocked trials ($M = 72.08$; std dev = 5.35) did not differ significantly [$t(347) = 0.66$, with $p = 0.51$].

4. General discussion

The results reported here suggest that although inconsistency among predictions has a negative impact on user trust, it occurs mainly prior to learning the outcome. Moreover, it is smaller than one might expect, and smaller than the reduction due to inaccuracy. Similar results were found in a series of experiments testing consistency among *sequential* forecasts from the *same* source. In that experiment as well, inaccuracy had a far greater negative impact on trust than did inconsistency (Burgeno and Joslyn 2020). Here, with simultaneous forecasts from the different sources, inaccuracy virtually wiped out any effect of inconsistency between them on postoutcome trust. Only when inconsistencies were unblocked and accurate was the expected reduction in postoutcome trust observed. It is difficult to explain why this effect would occur in the unblocked rather than the blocked condition, and we are reluctant to draw any conclusions from the small effect observed here. In sum, the postoutcome trust analysis suggests that consistency among forecasts has only limited effects on trust.

Forecast consistency was more important as participants were making their decisions prior to learning the outcome. In line with previous research in the financial domain (Budescu and Rantilla 2000; Budescu et al. 2003), participants had significantly higher confidence (a large effect size) in their decisions when forecasters agreed with one another. In addition, consistent forecasts were rated significantly more trustworthy than inconsistent forecasts. However, the effect size of consistency on trust preoutcome was intermediate making it smaller than the large effect of accuracy on trust postoutcome.

It is interesting to note that, as soon as participants learned the outcome, the effect of consistency on trust in the forecast dissipated. Analyses of change in trust indicated that trust increased postoutcome when forecasts were accurate and decreased when forecasts were inaccurate. In addition, despite the fact that there were equal numbers of accurate and inaccurate trials in each condition, trust tended to decrease overall postoutcome when the forecasts were consistent and increase when the forecasts were inconsistent. As can be seen in Fig. 2, inaccuracy had a smaller negative impact on trust in the inconsistent than the consistent trials, leveling out any advantage gained by consistency in preoutcome trust. This may have been because participants anticipated greater uncertainty with inconsistent forecasts, for which there was evidence in the wider range of expected outcome values when forecasts were inconsistent. Anticipating greater uncertainty may have served as a protective factor making inaccuracy seem less "wrong" and leading to a smaller loss in trust. This explanation could potentially account for the interaction showing higher trust in inconsistent than consistent forecasts when the forecast was inaccurate.

We were also interested in how participants used the forecasts to inform their own snow estimates and weather-related decisions. Participants' estimates were systematically greater than the mean of forecasts when they were inconsistent, suggesting overweighting of the larger value. The slight positive (high) bias may have been due to the task that participants were assigned, in which the more costly error was not closing schools when 6 or more inches of snow accumulation were observed. For this reason, participants may have decided to err on the side of caution as has been seen in previous research (Weber 1994) and to rely more heavily on the forecast with more accumulation. Indeed, participants were significantly more likely to close schools when the forecasts were inconsistent.

Granted the tight control of extraneous variables and systematic manipulation of experimental variables exercised here resulted in some loss to ecological validity. In addition, this was a vastly simplified decision task compared to its real-world counterpart. However, these techniques allowed us to make a direct comparison between the impact of consistency and accuracy on trust and decision-making when all else is equal. In other words, using this approach allowed us to pinpoint the causes of reduction in trust. Future studies should test these principles in more naturalistic contexts to better understand how they interact with other factors.

The work reported here suggests that the overall impact of inconsistency in simultaneous forecasts may not be as detrimental as was previously thought. The effect on postoutcome trust was minimal, and participants appeared to make good use of the information from both forecasts to gain a better understanding of the weather situations that confronted them. It is important to note, however, that only two sources of information were evaluated here. The picture may change as the number of sources increases, further increasing cognitive load. This is a potentially fruitful line of research for future studies. Nonetheless, the results of the current study suggest that those providing predictions to end users, in the weather domain or more generally, should prioritize accuracy over consistency to preserve users' trust in the long term. Not only is the effect of inconsistency on trust minimal once the outcome is known, inconsistency may also be an important additional source of information to decision-makers.

Acknowledgments. This research was supported by National Science Foundation Grant SES-1559126. Data for this experiment can be found online (https://osf.io/famcg). We thank Chao Qin in particular for assistance with coding of the program for our study and all members of Decision Making with Uncertainty Laboratory at University of Washington who provided insight and expertise that greatly assisted the research.

REFERENCES

Bloom, N., S. Bond, and J. Van Reenen, 2007: Uncertainty and investment dynamics. *Rev. Econ. Stud.*, **74**, 391–415, https://doi.org/10.1111/j.1467-937X.2007.00426.x.

Budescu, D. V., and A. K. Rantilla, 2000: Confidence in aggregation of expert opinions. *Acta Psychol.*, **104**, 371–398, https://doi.org/10.1016/S0001-6918(00)00037-8.

——, ——, H.-T. Yu, and T. M. Karelitz, 2003: The effects of asymmetry among advisors on the aggregation of their opinions. *Organ. Behav. Hum. Decis. Process.*, **90**, 178–194, https://doi.org/10.1016/S0749-5978(02)00516-2.

Burgeno, J. N., and S. L. Joslyn, 2020: The impact of weather forecast inconsistency on user trust. *Wea. Climate Soc.*, **12**, 679–694, https://doi.org/10.1175/WCAS-D-19-0074.1.

Cohen, J., 1988: *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. N. J. Hillsdale, Ed., Lawrence Erlbaum, 567 pp.

Drabek, T. E., 1999: Understanding disaster warning responses. *Soc. Sci. J.*, **36**, 515–523, https://doi.org/10.1016/S0362-3319(99)00021-X.

Earle, T. C., 2010: Trust in risk management: A model-based review of empirical research. *Risk Anal.*, **30**, 541–574, https://doi.org/10.1111/j.1539-6924.2010.01398.x.

Elder, K., S. Xirasagar, N. Miller, S. Bowen, S. Glover, and C. Piper, 2007: African Americans' decisions not to evacuate New Orleans before Hurricane Katrina: A qualitative study. *Amer. J. Public Health*, **97** (Suppl. 1), S124–S129, https://doi.org/10.2105/AJPH.2006.100867.

Joslyn, S., and D. Jones, 2008: Strategies in naturalistic decision-making: A cognitive task analysis of naval weather forecasting. *Naturalistic Decision Making and Macrocognition*, J. M. Schraagen, Ed., Ashgate, 183–201.

——, and S. Savelli, 2010: Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteor. Appl.*, **17**, 180–195, https://doi.org/10.1002/met.190.

——, and J. LeClerc, 2012: Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl.*, **18**, 126–140, https://doi.org/10.1037/a0025185.

Kadous, K., M. Mercer, and J. Thayer, 2009: Is there safety in numbers? The effects of forecast accuracy and forecast boldness on financial analysts' credibility with investors. *Contemp. Account. Res.*, **26**, 933–968, https://doi.org/10.1506/car.26.3.12.

Kahn, B. E., and M. F. Luce, 2003: Understanding high-stakes consumer decisions: Mammography adherence following false-alarm test results. *Mark. Sci.*, **22**, 393–410, https://doi.org/10.1287/mksc.22.3.393.17737.

Lazo, J. K., R. E. Morss, and J. L. Demuth, 2009: 300 billion served: Sources, perceptions, uses, and values of weather forecasts. *Bull. Amer. Meteor. Soc.*, **90**, 785–798, https://doi.org/10.1175/2008BAMS2604.1.

Losee, J. E., and S. Joslyn, 2018: The need to trust: How features of the forecasted weather influence forecast trust. *Int. J. Disaster Risk Reduct.*, **30**, 95–104, https://doi.org/10.1016/j.ijdrr.2018.02.032.

Morss, R. E., J. L. Demuth, and J. K. Lazo, 2008: Communicating uncertainty in weather forecasts: A survey of the US public. *Wea. Forecasting*, **23**, 974–991, https://doi.org/10.1175/2008WAF2007088.1.

NCEI, 2019: NCEI data accessed 12.2.19.csv.NOAA/NCEI, accessed 2 December 2019, https://osf.io/cu6dp/.

NOAA, 2016: Risk communication and behavior: Best practices and research findings. NOAA Social Science Committee Doc., 66 pp., https://www.performance.noaa.gov/wp-content/uploads/Risk-Communication-and-Behavior-Best-Practices-and-Research-Findings-July-2016.pdf.

Pappenberger, F., H. Cloke, A. Persson, and D. Demeritt, 2011: HESS Opinions "On forecast (in)consistency in a hydrometeorological chain: Curse or blessing?" *Hydrol. Earth Syst. Sci.*, **15**, 2391–2400, https://doi.org/10.5194/hess-15-2391-2011.

Perry, R. W., and M. R. Green, 1982: The role of ethnicity in the emergency decision-making process. *Sociol. Inq.*, **52**, 306–334, https://doi.org/10.1111/j.1475-682X.1982.tb01257.x.

Quarantelli, E. L., 1984: Perceptions and reactions to emergency warnings of sudden hazards. *Ekistics*, **51**, 511–515.

Ripberger, J. T., C. L. Silva, H. C. Jenkins-Smith, D. E. Carlson, M. James, and K. G. Herron, 2015: False alarms and missed events: The impact and origins of perceived inaccuracy in tornado warning systems. *Risk Anal.*, **35**, 44–56, https://doi.org/10.1111/risa.12262.

Twyman, M., N. Harvey, and C. Harries, 2008: Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. *Judgm. Decis. Making*, **3**, 111–120.

Weber, E. U., 1994: From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychol. Bull.*, **115**, 228–242, https://doi.org/10.1037/0033-2909.115.2.228.

Weyrich, P., A. Scolobig, and A. Patt, 2019: Dealing with inconsistent weather warnings: Effects on warning quality and intended actions. *Meteor. Appl.*, **26**, 569–583, https://doi.org/10.1002/MET.1785.

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

Zabini, F., 2016: Mobile weather apps or the illusion of certainty. *Meteor. Appl.*, **23**, 663–670, https://doi.org/10.1002/met.1589.