Q Open, 2024, **4**, 1–22
https://doi.org/10.1093/qopen/qoad031
Advance access publication date: 5 December 2023
Article

OXFORD

False and missed alarms in seasonal forecasts affect individual adaptation choices

Katharina Hembach-Stunden[1], Tobias Vorlaufer[2,*] and Stefanie Engel[1]

[1]School of Business Administration and Economics and Institute for Environmental Systems Research (IUSF), Osnabrueck University, Osnabrück, Germany
[2]Leibniz Centre for Agricultural Landscape Research (ZALF), Working Group Governance of Ecosystem Services, Müncheberg, Germany

*Corresponding author: Leibniz Centre for Agricultural Landscape Research (ZALF), Working Group Governance of Ecosystem Services, Eberswalder Str. 84, 15374, Müncheberg, Germany. E-mail: tobias.vorlaufer@zalf.de

Received: May 8, 2023. **Accepted:** November 28, 2023

Abstract

Facing climate change, seasonal forecasts, and weather warnings are increasingly important to warn the public of the risk of extreme climate conditions. However, being confronted with inaccurate forecast systems may undermine individuals' responsiveness in the long run. Using an online experiment, we assess how false alarm and missed alarm-prone forecast systems influence individuals' adaptation behaviour. We show that exposure to false alarm-prone forecasts decreases investments if a warning is issued (the 'cry-wolf effect'). Exposure to missed alarm-prone forecasts increases adaptation investments if no warning, but also if a warning has been issued. Yet, individuals exposed to both false and missed alarm-prone forecasts still adjust their adaptation investments depending on the forecasted probability of extreme climate conditions. Individuals with missed alarm-prone forecasts are, however, less sensitive to the forecasted probability if a warning has been issued. In case of low probability warnings, overshooting investments in adaptation hence becomes more likely.

Keywords: Climate information, Decisions under uncertainty, Economic online experiment, Forecast design, Early warning signals

JEL codes: Q54, D81, D83, C91

1. Introduction

Climate change globally increases the frequency and intensity of weather and climate extremes such as heatwaves, droughts, and heavy precipitation (IPCC 2022). Individual adaptation decisions that are solely guided by experience runs the risk of severely underestimating the need to adapt to these new conditions. In this context, seasonal forecasts and (regional) climate information are increasingly recognised as important sources of information and guiding tools for governments', private sectors', and households' adaptation actions (Stainforth et al. 2007; Bruno Soares et al. 2018; Knudson and Guido 2019; Webber 2019; Pacchetti et al. 2021). While climate information models, which predict long term changes in

© The Author(s) 2023. Published by Oxford University in association with European Agricultural and Applied Economics Publications Foundation. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

(local) climates, are potentially useful in guiding long-term private and public investment in climate adaptation (Stainforth et al. 2007; Pacchetti et al. 2021), seasonal forecasts, and warnings of extreme weather events can be especially useful to inform individual adaptation choices in the context of recurring decisions (e.g., cropping choices in agriculture) or temporary behavioural responses (e.g., in response to hurricane or flood warnings).

However, seasonal forecasts, as well as extreme event warnings, are often highly uncertain and inaccurate, which poses a challenge for both the communication of their predictions and their use (Zommers 2012; Taylor et al. 2015, 2018; National Institute of Water and Atmospheric Research (NIWA) 2016; Katzav et al. 2021). Besides adverse direct effects of following inaccurate forecasts in the short run, for example, experiencing a loss due to the failure to take preparatory measures based on a forecast, the repeated exposure to inaccurate forecasts may also have longer-term effects. Specifically, inaccurate forecasts may erode decision-makers' trust in the forecast system and lead them to ignore future forecasts or warnings (e.g., LeClerc and Joslyn 2015; Ripberger et al. 2015; Burgeno and Joslyn 2020). Thus, policymakers and agencies that are responsible for forecast and warning system design should take the consequences of inaccuracies into careful consideration. If not, they could risk limiting the overall contribution of forecast systems to climate change adaptation.

This paper uses an online experiment to address the question to what extent exposure to inaccurate forecast systems—that result in the repeated experience of false or missed alarms—affects individuals' adaptation decisions. Prior literature has identified two types of potential errors that individuals—who base their adaptation decisions on possibly inaccurate warnings—face (Losee and Joslyn 2018). On the one hand, they may experience a *false alarm*, where a warning is issued but an extreme event does not occur. In this case, decision-makers may comply with the warning and invest in costly adaptation, which later turns out to have been unnecessary. The effect that experiencing false alarms more frequently in the past decreases individuals' responsiveness to warnings is also known as the *cry-wolf effect* (e.g., LeClerc and Joslyn 2015; Trainor et al. 2015). Some studies find that high rates of prior false alarms reduced individuals' compliance with extreme weather warnings, leading individuals, for example, to be less likely to seek shelter in response to tornado warnings (Donner et al. 2012; Jauernic and Van Den Broeke 2017; Ripberger et al. 2015; Simmons and Sutter 2009) or react to extreme weather warnings (LeClerc and Joslyn 2015). However, other studies of behavioural responses to hurricanes (Dow and Cutter 1998) and tornado warnings (Schultz et al. 2010) do not find clear evidence of such a *cry-wolf effect*, or yield mixed results (Lim et al. 2019; Trainor et al. 2015).

On the other hand, individuals may experience a *missed alarm*, i.e., a situation where no warning is issued but extreme weather conditions strike. In this case, decision-makers might rely on the forecast and decide against adaptation, but then experience losses from extremes they were unprepared for. There are only a few empirical studies that focus on the impact of missed alarms on behaviour. Experimental findings suggest that larger forecast inaccuracies (increasing both the false and missed alarm rate at the same time) lead to lower trust in forecasts and compliance (Burgeno and Joslyn 2020; Joslyn and LeClerc 2012). Experimental studies that focus on fast responses to automatic machine alerts found that experiencing missed alarms more frequently affects behaviour. In these experiments, participants can acquire more information after either receiving an alarm or not. Experiencing missed alarms more frequently increases an individual's propensity to acquire more information in case no alarm has been issued (Chancey et al. 2015; Wiczorek and Meyer 2016).[1]

According to frequency-based probability learning theory (Estes 1976), individuals are able to approximate risks over time through repeated experiences. In the case of weather forecasts, one would expect individuals to learn about the false and missed alarms frequency over time and adapt their decisions accordingly. Forecast users would become less reactive to warnings and no-warnings, if they were exposed to a false- and missed-alarm prone system, respectively. But recent research has also shown that inaccuracies can cause a general decrease of trust in the forecast system (Joslyn and LeClerc 2012; LeClerc and Joslyn 2015; Burgeno and Joslyn 2020; Ripberger et al. 2015) and thereby may also affect how individuals react to the opposite signal. A history of false alarms may affect how individuals respond if no warning is issued. And vice versa, missed alarm experience may affect individuals' response if a warning is issued. We refer to these two effects as the *cross-effects* of inaccurate forecast systems. Empirical evidence concerning these *cross-effects* is scarce. LeClerc and Joselyn (2015) find that a decrease of false alarms, while simultaneously increasing missed alarms, decreases compliance suggesting the presence of a missed alarm cross-effect. Using survey data on hypothetical behaviour in case of a tornado warning, Ripberger et al. (2015) find that an increase in the perceived missed alarms rate decreases trust in warnings and subsequently lowers compliance. Experimental studies on automated machine alerts yield mixed results, finding evidence of negative cross-effects of false alarms (Wiczorek and Meyer 2016) or no evidence for any cross-effects (Manzey et al. 2014).

With a few exceptions (Burgeno and Joslyn 2020; Joslyn and LeClerc 2012; LeClerc and Joslyn 2015; Losee and Joslyn 2018), studies regarding forecasts or extreme event warnings are observational and mostly rely on self-reported data (Dow and Cutter 1998; Simmons and Sutter 2009; Schultz et al. 2010; Donner et al. 2012; Ripberger et al. 2015; Taylor et al. 2015; Trainor et al. 2015; Lindell et al. 2016; Jauernic and Van Den Broeke 2017; Lim et al. 2019). Controlling confounding factors is thus challenging and poses a challenge for identifying causal effects. For example, the frequency of accurate and inaccurate forecasts likely correlates with the location of residence, which in turn likely correlates with many other confounding factors (such as socio-economic characteristics, background risk, risk preferences, etc.) that also influence individual adaptation decisions. It is consequently challenging to identify the causal effect of repeated false and missed alarms with observational data. We circumvent this problem, by reporting the results of an incentivised online experiment. In the experiment, we randomly assign respondents to accurate, false, or missed alarm-prone forecast systems. In contrast to observational studies, our experiment allows to identify causal effects with high internal validity.

Our contribution to the existing literature is fourfold: First, we conceptually differentiate between false and missed alarm-prone forecast systems, a trade-off that decision-makers face when designing warning systems. For example, in the context of tornado warnings, governmental agencies responsible to publish warnings need to decide on a threshold when to issue an alarm. Under the same forecasting system, decreasing the missed alarm rates will ultimately result in more false alarms as more low-probability warnings are issued (Brooks and Correia 2018). Our experimental design allows us to identify the main and cross-effects of missed and false alarm-prone forecast systems, which to our knowledge has not been done in the context of weather forecasts and warnings. We evaluate separately whether exposure to increased missed alarm and false alarm rates does affect investments in protective adaptation measures when a warning as well as no warning is issued. While the existing experimental literature predominantly focused on the reaction to warnings, we extend the analysis to behaviour when no warnings are issued.

Second, in contrast to existing experimental studies (Burgeno and Joslyn 2020; Joslyn and LeClerc 2012; LeClerc and Joslyn 2015; Losee and Joslyn 2018), we elicit behaviour on a continuous scale instead of focusing on binary decisions, and are thus able to capture more nuanced differences in behaviour. The cry-wolf effect, for example, may reduce the willingness to react to a warning, but possibly not to an extent of complete inaction. Such nuances would be potentially lost with a binary decision, but are relevant in the context of decisions that allow for different degrees of protection, such as actions to reduce potential damages from hurricanes or floods, or cropping decisions of farmers at the beginning of a growing season.

Third, our experimental design allows disentangling the short- and long-term effects of inaccurate forecasts on adaptation behaviour. Prior to assessing the impact of error-prone

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

forecast systems, participants are, in our experiment, exposed to nine seasons. This allows us to assess whether the impact of false and missed alarms in the early seasons fades over time. Few researchers have explored the temporal dimension of false and missed alarms. One exception is Joslyn and LeClerc (2012), who conclude that trust is lost in the long run when exposed to a large-error forecast system early on even if forecast accuracy improves over time. Their design, however, does not differentiate between false and missed alarms, but instead focuses on forecasts errors in general.

Fourth, we assess whether exposure to a false and missed-alarm prone forecast systems also affects the sensitivity to the forecasted probabilities. Prior research indicates that false alarm rates, and error rates in general, affect trust and compliance less when forecasts are probabilistic instead of deterministic (Joslyn and LeClerc 2012; LeClerc and Joslyn 2015). To our knowledge, no study has evaluated to what extent forecast inaccuracies affect the actual sensitivity of adaptation behaviour to forecast probabilities. If inaccurate forecast systems undermine general trust in forecasts, individuals may not only be less responsive to warnings but also to the communicated probabilities.

Our results show that exposure to a false alarm-prone forecast system decreases individuals' willingness to invest in adaptation if they receive a forecast warning of extreme climate conditions in the future (the 'cry-wolf effect'), but does not influence future adaptation investments in the absence of a warning. Thus, we find no evidence for a false alarm cross-effect. Exposure to a missed alarm-prone forecast system, as expected, increases individuals' adaptation investments if no warning is issued. Surprisingly, missed alarm-prone systems also increase adaptation in the case when a warning is issued. Overall, we find that the main and cross-effects of false and missed alarm-prone forecast systems on adaptation investments are relatively small compared to the effect of the forecasted probabilities themselves. Individuals still react with an increase in their adaptation investment to increasing forecasted probabilities of upcoming climate extremes. Ultimately, our experimental results suggest that integrating probabilities into the forecast design is potentially more relevant than the long-term rate of false and missed alarms.

In the following section, we present the experimental design, our pre-registered hypotheses, and the analytic approach. Section 3 presents our results, and we conclude with a discussion of our main findings in Section 4.

2. Material and methods

This paper is based on an online experiment with multiple decision rounds, and was conducted with a sample from the general population in the UK. In the experiment, individuals decide over ten rounds—representing ten seasons—whether to invest in protection from extreme climate-related losses after receiving a probabilistic forecast that the upcoming season is of an extreme or normal climate. While protection minimises losses to zero if an extreme season occurs, it also decreases the final payout in a normal season. We systematically manipulate the accuracy of the forecast system to be either false or missed alarm-prone, which results in more false or missed alarm experiences. In the following, we describe the experiment along the three stages that participants face in each of the ten seasons.

2.1 Experimental design

At the beginning of each season, individuals have an endowment of 500 points and receive a probabilistic forecast that the upcoming season is of an extreme or normal climate, without knowing the true underlying risk of facing extreme climate conditions. The risk for the season to be extreme is randomly drawn from an underlying probability distribution shown in Table 1, Column 1 (CTRL—accurate). Participants know that the outcomes are independent across seasons. In each season, the computer draws one of twelve possible risk options

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Table 1. Forecast probability design and outcome in the three treatments for Round 1 to 9.

	(1) CTRL—accurate (= true underlying risk) forecast probability			(2) FA—false alarm-prone forecast probability			(3) MA—missed alarm-prone forecast probability		
Option	extreme	normal	warning	extreme	normal	warning	extreme	normal	warning
1	0.15	0.85	no	0.6	0.4	yes	0.6	0.4	yes
2	0.2	0.8	no	0.65	0.35	yes	0.65	0.35	yes
3	0.25	0.75	no	0.7	0.3	yes	0.25	0.75	no
4	0.3	0.7	no	0.75	0.25	yes	0.3	0.7	no
5	0.35	0.65	no	0.8	0.2	yes	0.35	0.65	no
6	0.4	0.6	no	0.85	0.15	yes	0.4	0.6	no
7	0.6	0.4	yes	0.6	0.4	yes	0.15	0.85	no
8	0.65	0.35	yes	0.65	0.35	yes	0.2	0.8	no
9	0.7	0.3	yes	0.7	0.3	yes	0.25	0.75	no
10	0.75	0.25	yes	0.75	0.25	yes	0.3	0.7	no
11	0.8	0.2	yes	0.35	0.65	no	0.35	0.65	no
12	0.85	0.15	yes	0.4	0.6	no	0.4	0.6	no

	A priori probabilities[a]	*A posteriori* probabilities[b]	*A priori* probabilities[a]	*A posteriori* probabilities[b]	*A priori* probabilities[a]	*A posteriori* probabilities[b]
accurate warning	0.36	0.38	0.36	0.35	0.03	0.03
false alarm	0.14	0.13	0.47	0.48	0.14	0.13
accurate no-warning	0.36	0.35	0.03	0.03	0.36	0.38
missed alarm	0.14	0.13	0.14	0.14	0.47	0.47

Note: Each season the computer randomly chooses one of the 12 options, which determines the communicated forecast probabilities and whether or not individuals receive a warning. In CTRL, the values from the twelve options representing the true underlying risks are shown as forecasts. Participants in treatments FA and MA are shown the matching false or missed alarm-prone probabilities instead of the true underlying risks.

[a] ***A priori* probabilities** refer to the *a priori* probabilities to receive an accurate warning, an accurate no-warning, a false alarm or a missed alarm in seasons 1 to 9. A priori probabilities for season 10 are the same for all treatments and follow the shown *a priori* probabilities in the CTRL treatment.

[b] ***A posteriori* probabilities** are calculated based on seasons one to nine, excluding season 10.

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

with a 1/12 probability. The risk options each have a probability between 15 and 85 per cent that an extreme season occurs, with an expected probability of 50 per cent.

To simplify the understanding of our experiment, we introduce two forecast categories, normal and extreme seasons. Each forecast shows the probabilities of both extreme and normal conditions for that upcoming season (Fig. 1A). If the forecasted probability for an extreme season is 60 per cent or higher, participants receive an alarm in form of a *warning forecast* (Fig. 1A.1). In contrast, if the forecasted probability for an extreme season is lower than 60 per cent, participants receive a standard forecast message (i.e., *no-warning forecast*, Fig. 1A.2). This design of our experimental forecast is based on common seasonal forecasts of precipitation or temperature. Such seasonal forecasts present the probability of whether the upcoming season is likely to be normal, below normal, or above normal (see, e.g., the seasonal forecasts provided by the International Research Institute for Climate and Society at Columbia University, U.S.A., https://iri.columbia.edu/, accessed 17 July 2023). Seasonal forecasts are inevitably probabilistic due to the uncertainties in climate models and imperfect knowledge of the atmosphere and climate system (Smith et al. 2019). Previous research also indicates that communicating the probabilistic nature of forecasts (versus stating forecasts as deterministic) reduces the perceived inconsistency of warnings and have a positive effect on trust in the forecast system in general (Fundel et al. 2019; LeClerc and Joslyn 2015; Losee and Joslyn 2018).

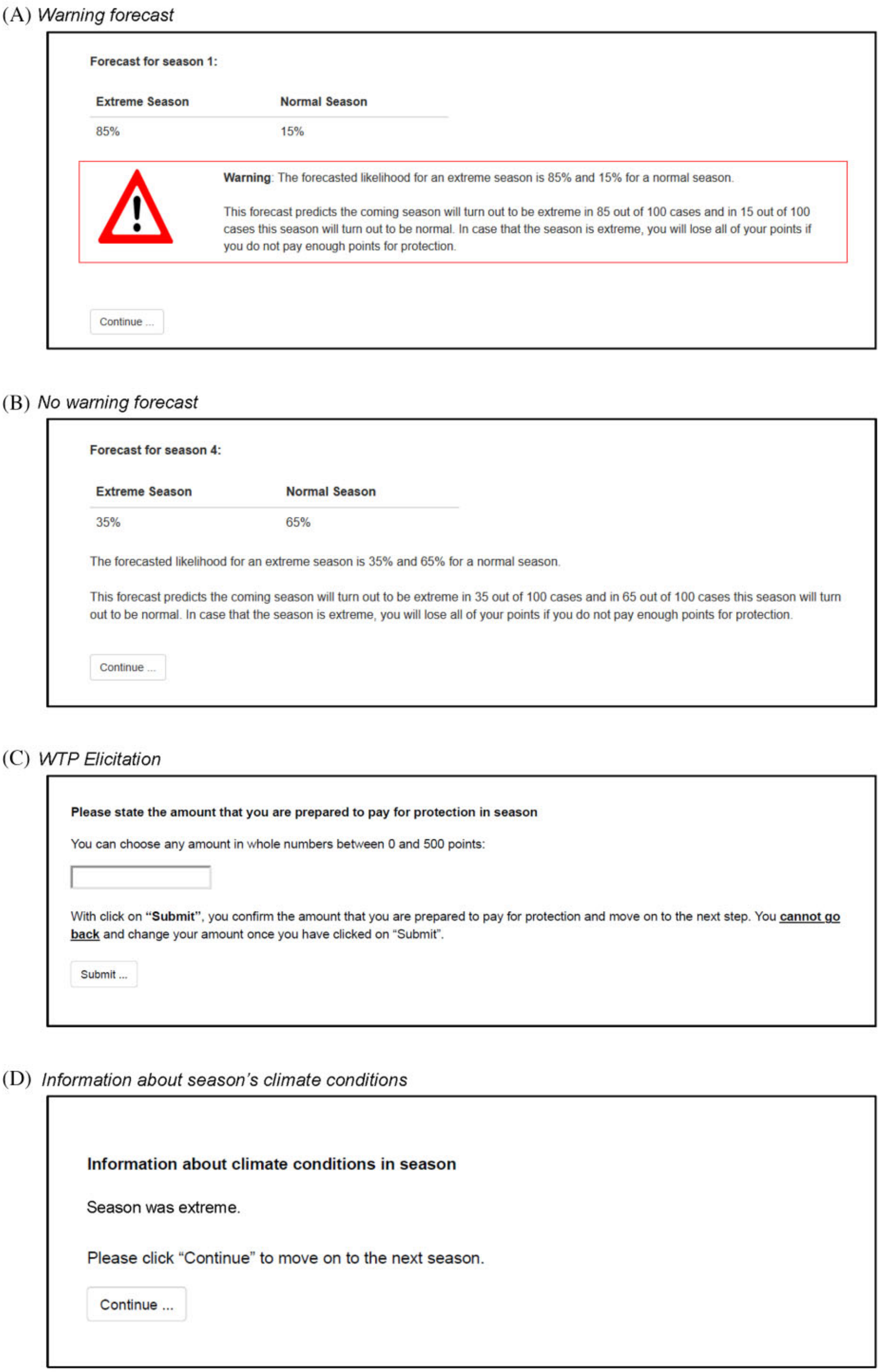
Secondly, after participants receive the forecast, we elicit their adaptation behaviour in form of their willingness to pay (WTP) for protection from extreme climate conditions, with a minimum of 0 and a maximum of 500 points (see Fig. 1B). Experiencing extreme climate conditions without any protection would result in losing all of their 500-point endowment. If participants invest in adaptation and experienced a normal season, they paid for unneeded protection. We opted for eliciting individuals' WTP based on the Becker–DeGroot–Marschak (BDM) method (Becker et al. 1964) to yield a continuous measure for adaptation behaviour instead of simply eliciting a binary decision of whether to adapt (i.e., follow the warning) or not. With this, we can measure fine, more nuanced treatment effects.[2] The BDM method has the advantage that it is incentivised and motivates individuals to state their true WTP because their stated values do not influence the final price (Schmidt and Bijmolt 2019).

Thirdly, after participants have stated their WTP, the computer randomly determines the climate conditions of the season based on the underlying risk option and the climate outcome is revealed to participants (see Fig. 1C). Thereby, participants receive indirect feedback on the accuracy of their forecast system each season and the underlying risk of an extreme season. They can update their beliefs about the accuracy of the forecasts over time and adjust their stated WTP in response to warning and no-warning forecasts accordingly. Respondents did not receive information on their earnings after each season. This would require to determine a randomly selected price for protection (in accordance with the BDM method) and could have induced anchoring effects for the WTP in the upcoming seasons.

After the tenth and final season, the computer randomly chooses the payout-relevant season. Following the BDM method, it then randomly determines the price for protection in that season between 0 and 500 points. As participants know from the beginning that only one season is randomly selected as payout-relevant at the end, it is not possible to hedge risks over the ten seasons. Participants also know that all prices for protection, between 0 and 500 points, were equally likely and that none of the events during the ten seasons would influence the randomly determined price set by the computer in the end.

The final payout in points is calculated as follows and converted to Great British Pounds (£1.00 = 100 points). If participants had indicated a WTP equal to or higher than the price, they purchase the protection for the determined price, irrespective of the climate conditions of the season. Consequentially, they receive the rest of their endowment as payout, irrespective of an extreme or normal season. However, if participants had indicated a WTP lower

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024



(A) *Warning forecast*

Forecast for season 1:

Extreme Season	Normal Season
85%	15%

Warning: The forecasted likelihood for an extreme season is 85% and 15% for a normal season.

This forecast predicts the coming season will turn out to be extreme in 85 out of 100 cases and in 15 out of 100 cases this season will turn out to be normal. In case that the season is extreme, you will lose all of your points if you do not pay enough points for protection.

(B) *No warning forecast*

Forecast for season 4:

Extreme Season	Normal Season
35%	65%

The forecasted likelihood for an extreme season is 35% and 65% for a normal season.

This forecast predicts the coming season will turn out to be extreme in 35 out of 100 cases and in 65 out of 100 cases this season will turn out to be normal. In case that the season is extreme, you will lose all of your points if you do not pay enough points for protection.

(C) *WTP Elicitation*

Please state the amount that you are prepared to pay for protection in season

You can choose any amount in whole numbers between 0 and 500 points:

With click on **"Submit"**, you confirm the amount that you are prepared to pay for protection and move on to the next step. You **cannot go back** and change your amount once you have clicked on "Submit".

(D) *Information about season's climate conditions*

Information about climate conditions in season

Season was extreme.

Please click "Continue" to move on to the next season.

Figure 1. Screenshots of exemplary forecasts (A and B), the WTP elicitation (C) and the seasonal outcome information screen (D) in the experiment. Panel A presents the forecast design of the experiment with a

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Figure 1. warning forecast, and Panel B presents a no-warning forecast. The forecast probabilities for an extreme and normal season are presented in both cases. Panel C presents the screen where participants are asked to state their willingness to pay for protection (WTP; between 0 and 500 points). Panel D depicts the screen at the end of each season, informing participants about the outcome of the season. The example here is an example where the season turned out to be extreme and thus states 'Season was extreme'. In case of a normal season, it states 'Season was normal'.

than the randomly determined price, they do not buy protection and are not protected from extreme climate-related losses. Thus, their payout depends on the climate conditions of the payout-relevant season. They lose their full endowment if the season is extreme, but keep their full endowment if the season is normal.

We decided to partially frame the experiment in the context of seasonal climate conditions and forecasts to increase familiarity with the basic experimental decision environment: *'In every season, you are at risk of losing all of your 500 points through extreme climate conditions. Extreme climate conditions could come in the form of heatwaves, droughts, heavy storms and flash flooding events. However, you can protect yourself from extreme climate-related losses by paying with your points for protection from extreme climate conditions.'* We opted, however, to not frame the specific adaptation behaviour at hand, since the availability of specific adaptation behaviours largely differs between individuals (e.g., between house owners and tenants). This partial framing allows us to associate the overall design with climate change adaptation and warnings of extreme weather events. At the same time, our design allows learning about general behavioural patterns by not being too specific with respect to adaptation behaviours.

2.2 Treatment manipulations

We systematically vary the accuracy of the forecast systems in three treatment conditions (represented in the three columns of Table 1). The underlying probabilities of an extreme season and, thus, the risk options are identical across treatments. However, the three treatments differ in the forecasted probabilities, and consequently the number of warning and no-warning forecasts that participants receive. In the *control treatment (CTRL)*, the forecast system is accurate, and the forecasted probabilities show the true underlying risk of an extreme season based on the drawn risk option for the specific season. Thus, on average, half of the forecasts issue a warning of an extreme season (i.e., probability of an extreme season being 60 per cent or higher) and the other half do not (i.e., probability of an extreme season being lower than 60 per cent).

In addition, we include two treatments with inaccurate forecast systems that either issue too many or too few warnings and systematically over- or underrate forecast probability. In the *false alarm treatment (FA)*, options 1 to 6 of the underlying risk options are matched with an overrating forecast probability and consequently, with too many warnings being issued (see Table 1, Column 2 [FA—false alarm-prone]). In the *missed alarm treatment (MA)*, the forecasted probabilities matched with the risk options 7 to 12 understate the risk of an extreme season, with too few warnings being issued (see Table 1, Column 3 [MA—missed alarm-prone]).[3] As a result, both the MA and FA forecast systems are expected to generate missed and false alarms in 47 per cent of the seasons, respectively.[4]

For the analysis of treatment differences, we solely focus on behaviour in season 10. To balance the number of warning and no-warning forecasts between treatments in season 10 and to assure that the communicated probabilities of extreme seasons are the same across treatments, the forecasts for season 10 show the true underlying risk of an extreme season in all forecast systems. We can verify our treatment designs regarding the forecast inaccuracy by analysing the frequency at which accurate forecasts, false and missed alarms occurred in the first nine seasons (lower section of Table 1). In CTRL ($N = 667$), 38 per cent of the warning forecasts and 35 per cent of the no-warning forecasts were accurate, 13 per cent

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

were false alarms, and 13 per cent were missed alarms. In FA ($N = 667$), 35 per cent were accurate warnings, 3 per cent were accurate 'no-warnings', 48 per cent were false alarms, and 14 per cent were missed alarms. Whereas in MA ($N = 666$), 3 per cent were accurate warnings, 38 per cent were accurate 'no-warnings', 13 per cent were false alarms, and 47 per cent were missed alarms. Thus, the *a posteriori* probabilities for the four different forecast cases match the *a priori* probabilities that we aimed for, and both the false and missed alarm-prone forecast systems led to the desired rate of inaccuracy.

At the beginning of the experiment, participants only know that they are randomly assigned to a forecast model (treatment) that generates all forecasts that they receive during the experiment. Participants do not know up front how (in-)accurate the forecasts generated by their assigned model are and whether forecast models differed in the degree of (in-)accuracy, but just receive the following information: '*Before the first season, you will be randomly assigned a forecast model, which will generate all forecasts you receive for all 10 seasons. You will not be informed of how accurate (or inaccurate) the forecasts generated by your assigned model are.*' Over the course of the first nine seasons, participants are able to get a sense of the accuracy of their assigned forecast model, in terms of false and missed alarm frequencies. We do not explicitly explain what we mean by accuracy to reduce the cognitive load for subjects at this stage. We consider the first nine seasons sufficient to get an understanding of what accuracy in this experimental context implies.

2.3 Treatment assignment

The computer assigned participants to one of the three treatments based on the order in which participants finished the instructions. The first participant to finish was assigned to the control treatment CTRL, the second participant to the FA treatment, the third participant to the MA treatment, the fourth to CTRL again, and so on. This was done because entirely random treatment assignment may have resulted by chance in treatment imbalances within a given time period (e.g., the first hour of the data collection or at a specific time of the day). Particularly, the time of participation could be correlated with behaviour, and thus could induce a confounder for the identification of treatment effects. For example, participants who are more eager to earn money by participating in experiments, are potentially more experienced on Prolific (the online recruitment platform used for the data collection), monitor available studies more frequently, and are more likely to participate early on. If the random treatment assignment would not be balanced in the beginning, any treatment differences may be an artefact of these imbalances. Since participants were not aware of the treatment assignment mechanism and could not control the timing of finishing the instructions to an extent that would allow them to influence their treatment assignment, we consider our procedure quasi-random. We do not find statistically significant differences at the 5 per cent level in the socio-economic characteristics between treatments (Table S1), indicating that our assignment procedure resulted in comparable treatment groups.

2.4 Hypotheses

As with real weather forecasts, participants face two different uncertainties in our experiment.[5] Firstly, they do not know the true risk of an extreme season, nor the outcome of the next season. According to frequency-based probability learning (Estes 1976), individuals are expected to approximate their risk of being exposed to an extreme season over the course of the ten experimental seasons. Secondly, participants are uncertain of the accuracy of the forecasts themselves. Again, based on frequency-based probability learning (Estes 1976) subjects likely learn over time how accurate their forecasts were, based on the experienced frequency of false and/or missed alarms. As outlined above, learning over time regarding forecast inaccuracy may result in either lower trust in the forecasts in general

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

or lower trust in specific forecast signals (warnings and no-warnings).[6] In both cases, this would lead to the following hypothesised behaviour:

Hypothesis 1: Being exposed to a false alarm-prone as opposed to an accurate forecast system decreases adaptation investments in response to a warning forecast ('cry-wolf effect').

Hypothesis 2: Being exposed to a missed alarm-prone as opposed to an accurate forecast system increases adaptation investments in response to a no-warning forecast.

Prior research has indicated that forecast inaccuracies decrease trust in the general forecast system. Based on survey data, Ripberger et al. (2015) found that perceived false and missed alarm ratios are negatively correlated with stated trust in the US National Weather Service. Burgeno and Joslyn (2020) report that lower forecast accuracy (including both false and missed alarms) results in lower post-outcome trust in an experimental setting. Based on these studies, we would expect false and missed alarm-prone forecast systems to undermine overall trust in forecasts and lead to following two cross-effects:

Hypothesis 3: Being exposed to a false alarm-prone as opposed to an accurate forecast system increases adaptation investments in response to a no-warning forecast.

Hypothesis 4: Being exposed to a missed alarm-prone as opposed to an accurate forecast system decreases adaptation investments in response to a warning forecast.

2.5 Data collection

For our experiment, we recruited 2,000 residents of the United Kingdom (UK) via the online crowdsourcing platform Prolific (Prolific 2021) in July 2020. Participants received a fixed payment of £2 and a variable payment between £0 and £5 (0 and 500 points) depending on the experimental outcomes. The average payout, including the variable payment was £4.95 (SD = 1.99, $N = 2{,}000$). Our study was programmed using the 'Software Platform for Human Interaction Experiments' (SoPHIE) (Hendriks 2012). The study consisted of two parts, an economic experiment followed by a post-experimental questionnaire. The average completion time was 14 minutes (SD = 7, $N = 1{,}996$) of which 10 minutes (SD = 5, $N = 2{,}000$) were spent on the experiment itself. Participation was voluntary and we followed the common ethical standards of data confidentiality and anonymity. All participants gave their consent at the beginning of the study. We preregistered our study and hypotheses at 'AsPredicted.org' (https://aspredicted.org/ay5zm.pdf) (Wharton Credibility Lab 2017). The experimental material, the pre-analysis plan, and the replication data and analysis scripts are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/TMESK).

We successfully assured gender balance (50 per cent female participants). The average age of participants was 34.5 years (SD = 12.7), with the majority living in urban areas with 10,001 to 100,000 inhabitants (49 per cent). The average household size was three individuals (SD = 1.4), and 76 per cent of all participants had a disposable monthly household income below £4,500, and 41 per cent were the owners of the house they live in. The majority of participants had either a college degree (27 per cent) or a bachelor's degree (38 per cent). Please see Table S1 for further details of participants' socio-economic characteristics, Section 2 of the Supplementary Information (SI) for a discussion on the correlation between

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

socio-economic characteristics, respondents' attitudes, and behaviour in the experiment, and SI, Section 3 for the experimental instructions, including the post-experimental questionnaire.

We implemented control questions as part of the instructions and added a trap question to assure participants' understanding and attentiveness during the experiment (Berinsky et al. 2014; Malone and Lusk 2018). Our trap question was a colour screener similar to the one presented in Berinsky et al. (2014), where respondents are asked for their favourite colour with an explanatory note to not answer the question and instead enter 'none'. Overall, the results of our control and attention questions point out that the majority of participants read the instructions carefully and paid attention to the experiment (see Table S1).

2.6 Identification strategy

Our outcome variable of interest for the data analysis is individuals' WTP for protection in the final season (season 10). In the analysis, we test for treatment effects in the two sub-samples who received and did not receive a warning in this season 10, as we specified separate hypotheses for these two sub-samples (see also the pre-registration).

In our analysis, we first focus on a treatment-level comparison between FA and MA relative to CRTL. As indicated in Table 1, both treatments systematically vary the accuracy of the forecasts by increasing either the FA or MA rate. Please note, however, that participants in all three treatments experienced FA, MA, and accurate forecasts. As such, our main analysis—in accordance with our hypotheses—estimates the aggregated impact of being exposed to a FA and MA prone-forecast system, rather than the impact of actual FA and MA rates. We estimated Tobit models to account for the censoring of the outcome between 0 and 500 points:

$$WTP|\,(no)\,warning = \alpha + \beta_1 * MA + \beta_2 * FA + \beta_3 * Prob + \epsilon, \tag{1}$$

where *MA* and *FA* are dummy variables, taking on 1 if the participant was assigned to the MA and FA treatment, respectively, and otherwise 0, *Prob* specifies the probability of the forecast in season 10. While the forecasted probability of an extreme season is by design not independent of the warning itself, we can control for the forecasted probabilities, conditional on having received or not received a warning. We estimate four separate Tobit regression models: We used the observations of CTRL repeatedly for the comparisons to FA and MA and have two sub-samples per treatment depending on whether or not a warning was issued in season 10. The estimates for the FA main and cross effect are represented by β_1 for the sub-samples who received and did not receive a warning, respectively. The estimates for the MA main and cross effect are represented by β_2 for the sub-samples who did not receive and received a warning, respectively.

In a second step, we focus—instead of comparing outcomes between treatment groups—on the history of experienced FA and MA. Please note that this analysis is exploratory and was not part of the pre-registration. We aggregate the frequency of false and missed alarms in seasons 1–3, 4–6, and 7–9. While an analysis strategy that takes into account the more nuanced differences in alarm histories would be ideal, we aggregated the outcomes in order to maintain sufficient statistical power. We can, therefore, identify whether the effects of FA and MA early on in the experiment persist and affect behaviour until the end. Again, we focus here on WTP in season 10 as outcome variable. The corresponding Tobit model was specified as follows:

$$\begin{aligned} WTP|\,(no)\,warning = \alpha + \beta_1 * MA_{1-3} + \beta_2 * MA_{4-6} + \beta_3 * MA_{7-9} + \beta_4 * FA_{1-3} \\ + \beta_5 * FA_{4-6} + \beta_6 * FA_{7-9} + \beta_3 * Prob + \epsilon, \end{aligned} \tag{2}$$

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Table 2. Summary statistics of the main outcome, WTP in season 10.

	Warning		No Warning	
Treatment	Mean WTP (SD)	*N*	Mean WTP (SD)	*N*
CTRL	330 (120)	342	170 (130)	325
FA	305 (131)	316	170 (137)	351
MA	349 (129)	353	202 (147)	313

Note: N denotes the number of observations. Mean WTP is the mean WTP in season 10. Given the bonus of 500 points, the minimum possible WTP was 0 and the maximum was 500 in all treatments. Standard deviations (SD) are presented in parentheses. 'Warning' includes only the observations where individuals received a warning in season 10, while 'no warning' includes only the observations of individuals receiving no warning in season 10.

where the six variables *MA* and *FA* indicate the frequency of false and missed alarms in seasons 1–3, 4–6, and 7–9, respectively (ranging between 0 and 3), and *Prob* specifies the forecast probability that was provided in season 10. We estimate two separate Tobit models depending on whether a warning or no warning was issued in season 10.

3. Results and discussion

In season 10, the average WTP for protection across treatments was 329 points if a warning (SD = 128, $N = 1{,}011$) and 180 points if no-warning was issued (SD = 139, $N = 989$). The average WTP per treatment is shown in Table 2. The development of average WTP from seasons 1 to 9 is provided in Figure S1. We observe that WTP remains relatively constant in CTRL, decreases over time in FA and increases over time in MA, indicating that participants infer the accuracy of the forecast system based on their experiences over time and adapt their WTP accordingly.

We will first focus on testing the four hypotheses outlined in Section 2. Figure 2 presents the treatment coefficients of the four Tobit models (see Eq. 1) (see Table S2 for the full regression results), one for each hypothesis.

Exposure to a false alarm-prone forecast system decreases average adaptation investments in response to a warning. Individuals' WTP if they received a warning forecast is on average significantly lower in FA than in CTRL (Fig. 2B), confirming Hypothesis 1. However, calculating Cohen's *d* as a measure of the effect size shows that the effect is only 0.2 standard deviations and thus, small (Cohen's $d = -0.20$).[7] Our finding of a *cry-wolf effect* is in line with previous experimental studies that focus on binary decisions whether to follow a warning (LeClerc and Joslyn 2015) and observational studies focusing on tornado warnings in the US (Donner et al. 2012; Jauernic and Van Den Broeke 2017, Ripberger et al. 2015; Simmons and Sutter 2009).

Exposure to a missed alarm-prone forecast system increases adaptation investments in the absence of a warning. We find a significant increase in the WTP among individuals who experience no warning in MA compared to CTRL (Fig. 2A). We therefore also confirm Hypothesis 2, and again, the treatment effect is small (Cohen's d = 0.231). This finding is in line with previous studies that focus on deterministic machine warning systems (Chancey et al. 2015; Wiczorek and Meyer 2016). We find that also with a probabilistic forecast, individuals rely less on no-warning forecasts if they are exposed to a missed alarm-prone forecast system. These results confirm the presence of both false and missed alarms' main effects and are in line with frequency-based probability learning (Estes 1976). Participants are

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

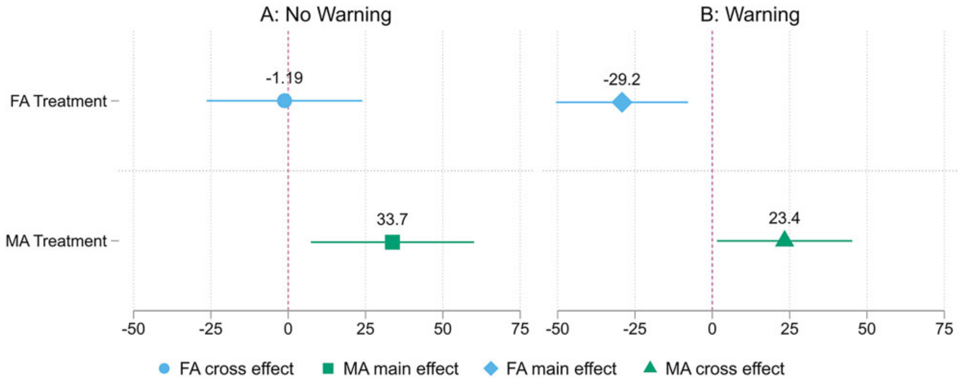


Figure 2. Coefficient plots based on the four separate Tobit regression models with the dependent variable 'WTP in season 10'. The coefficient plots display the point estimates for the coefficients 'False alarm treatment' (FA) and 'Missed alarm treatment' (MA) with their 95 per cent-confidence intervals along the *x*-axis. These coefficients represent the treatment effect on WTP and are estimated relative to the control treatment CTRL. Figure 2A shows the coefficient plots for if no warning was issued (Hypotheses 2 and 3), also defined as the MA main and FA cross effect. Figure 2B shows the coefficient plots if a warning was issued (Hypotheses 1 and 4), also known as FA main and MA cross-effect. The dotted, vertical line at zero is a reference line to visualise which coefficients are significantly different from zero at the 0.05 level. See Table S2 for the corresponding Tobit regression models.

less likely to follow the forecast because they have updated their beliefs about the forecast accuracies throughout the experiment.

There is no evidence of a negative cross-effect on adaptation investments in the absence of a warning from being exposed to a false alarm-prone forecast system. We do not observe a significant difference in individuals' WTP when comparing FA to CTRL if they do not experience a warning forecast (Fig. 2A). Our data thus does not support Hypothesis 3, namely that exposure to a false alarm-prone system increases adaptation investments if no warning is issued. This result agrees with Manzey et al. (2014), who also did not find evidence for a FA cross-effect, but is in contrast with Wiczorek and Meyer (2016), who observe a negative cross-effect. Both experimental studies focus on machine warning systems.

Contrary to Hypothesis 4, we find evidence of a positive cross-effect on adaptation investments from exposure to a missed alarm-prone forecast system if a warning is issued. Average WTP after receiving a warning forecast is, surprisingly, significantly higher among individuals in MA compared to CTRL (Fig. 2B). Thus, our result does not concur with previous studies on deterministic warning systems that find evidence for a negative cross-effect (LeClerc and Joslyn 2015; Ripberger et al. 2015) or evidence for no effect (Manzey et al. 2014; Wiczorek and Meyer 2016). Nonetheless, the MA cross-effect in our study is smaller than the two main effects (Cohen's $d = 0.15$). Consequently, it is potentially negligible, even though it is statistically significant.

Our cross-effect results are partially in line with the frequency-based probability learning theory. Accordingly, individuals in the MA treatment have experienced that the forecasts systematically underestimate the risks of an extreme season and thus assume that this is also the case if a warning is issued. Surprisingly, this is, however, not the case in the FA treatment. One possible explanation would be that individuals exposed to false alarm-prone systems only update their beliefs regarding the accuracy of the warning; hence, merely the trust in the warning signal is undermined. In contrast, individuals exposed to missed alarm-prone systems update their beliefs about the accuracy of the warning and the forecast probabilities as well, leading to lower trust in the overall forecast system. Both mechanisms could be—in principle—compatible with the frequency-based probability learning theory (Estes 1976).

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

However, the different inaccurate signals would lead to the updating of different probabilities: the probability of a correct warning in the case of false alarms and the probability of a correct probabilistic forecast in the case of missed alarms. We provide additional analysis of trust in the forecast that was elicited after the last season in Section 3 of the SI. While we find that the FA and MA treatments lead to lower trust compared to CTRL, trust in the forecast only affects decisions if a warning is issued. These findings echo Chancey et al. (2015), who analysed deterministic a warning system and found that trust does not mediate the relationship between missed alarms and behaviour in the absence of a warning.

Our results are robust to specifications, including a broad set of controls (sociodemographics, prior experiences of climate extremes, risk preferences, attention during and understanding of the experiment) (Table S3), a restricted sub-sample analysis based on our comprehension questions (Table S4), and pooling observations of respondents who received and did not receive a warning in season 10 (Table S5).

In addition to the treatment-level comparison, we also estimated the effect of the relative false and missed alarms frequency in seasons 1–9 on behaviour in season 10 (Table S6). These findings corroborate the treatment level analysis. Being exposed to more false alarms decreases adaptation investments if a warning is issued. Being exposed to more missed alarms increases adaptation investments, both if a warning is issued and not issued. Including both false and missed alarms frequencies and treatment indicators in regression models, renders the treatment coefficients insignificant, which indicates that the treatment effects are entirely mediated through the frequency of inaccurate forecasts (Table S7).

3.1 Persistence of false and missed alarm effects

In the next step, we again depart from the treatment level comparison. Instead, we focus on the frequency of false and missed alarms at different stages of the experiment (see Eq. 2) to test for the persistence of false and missed alarm effects over time. Figure 3 illustrates the results of the corresponding regression analysis that focuses again on adaptation decisions in season 10.

Experiencing more frequent false alarms has a persistent negative impact on adaptation decisions, while the impact of missed alarms on adaptation decisions is larger for recent than earlier seasons. Similar to the results reported above, we find that the main effects of both false and missed alarms lead to less and more adaptation if a warning and no warning is issued, respectively. Interestingly, the impact of false alarms early on in the experiment (seasons 1–3 and 4–6) have a similar impact on decisions as false alarms in later seasons (seasons 7–9), suggesting that the cry-wolf effect is relatively persistent and does not fade over time. Once trust in the warning signal is lost, it is not regained. We do not observe this dynamic for missed alarms in case that no warning was received. Here, we find that missed alarms in the three most recent seasons (7–9) increase adaptation investments to a larger extent than missed alarms in earlier seasons (1–3 and 4–6). Joslyn and LeClerc (2012) found that trust—lost due to inaccurate forecasts early in their experiment—is difficult to regain with more accurate forecasts later on. Our findings complement theirs and suggest that missed and false alarms may be perceived differently. False alarms have more persistent effects on behaviour, while more recent missed alarms weigh heavier than earlier ones.

3.2 Treatment effects relative to forecasted probabilities

Even though we observe statistically significant treatment effects for three of our four hypotheses, one key question is how strong these effects are relative to other determinants of adaptation. In this section, we, therefore, compare the treatment effects to the effects of the forecasted probabilities in season 10.[8] In addition, we also assessed if exposure to a false or missed alarm-prone forecast system affects the sensitivity to the forecasted probabilities. To do this, we estimate separate Tobit models for the sub-samples who received and did

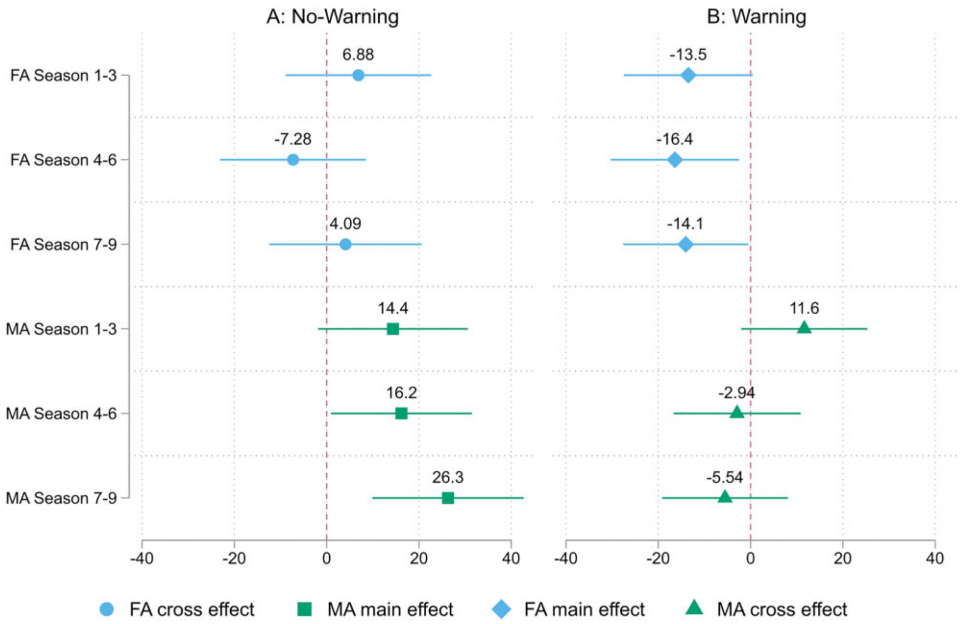


Figure 3. Coefficient plots based on the two separate Tobit regression models with the dependent variable 'WTP in season 10' and the frequency of MA and FA as explanatory variables. The coefficient plots display the point estimates for the coefficients 'False alarm' (FA) and 'Missed alarm' (MA) frequency in Seasons 1–3, 4–6, and 7–9 with their 95 per cent-confidence intervals along the *x*-axis. Figure 3A shows the coefficient plots if no warning was received. Figure 3B shows the coefficient plots if a warning was received. The dotted, vertical line at zero is a reference line to visualise which coefficients are significantly different from zero at the 0.05 level. See Table S8 for the corresponding Tobit regression models.

not receive a warning in season 10 and include the interaction of the treatment dummy and forecasted probability (see Table S9). Based on these regression models, Fig. 4 illustrates the predicted adaptation investments by treatment in relation to the different forecast probabilities in season 10. The figure also illustrates the optimal WTP of a risk-neutral individual, if the forecasted probabilities are expected to reflect the underlying probability of an extreme season (red line). We use this as a benchmark to which we compare the predicted WTP of our estimated models. If no warning is issued (with a forecast probability < 0.6), predicted WTPs are mostly above the benchmark, which is expected as most individuals are risk averse (Dohmen et al. 2011). However, if a warning is issued (with forecast probabilities ≥ 0.6), predicted WTPs are mostly below the benchmark, suggesting that subjects systematically anticipate lower probabilities than communicated in the forecasts (under the assumption that individuals are on average risk averse).

The treatment effects are relatively small compared to the effects of the forecasted probabilities. We observe that overall, the higher the forecasted probability, the higher the predicted WTP (Fig. 4). These relative effects of the forecasted probabilities are overall stronger than the treatment effects measured as Cohen's d ($0.2 < d > 1.09$, see Table S10; treatment effects range between $-0.15 < d > 0.23$). The behavioural responses to the forecasted probabilities are thus stronger than the impact of false or missed alarm-prone forecast systems on individual behaviour.

More frequent false alarms do not affect the sensitivity to forecasted probabilities, while more frequent missed alarms to some extent do. In case that no warning is issued, the near-identical slopes of the treatments indicate that the sensitivity to the forecast probabilities is very similar across treatments (see Fig. 4, left section). Similarly, if a warning is issued,

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

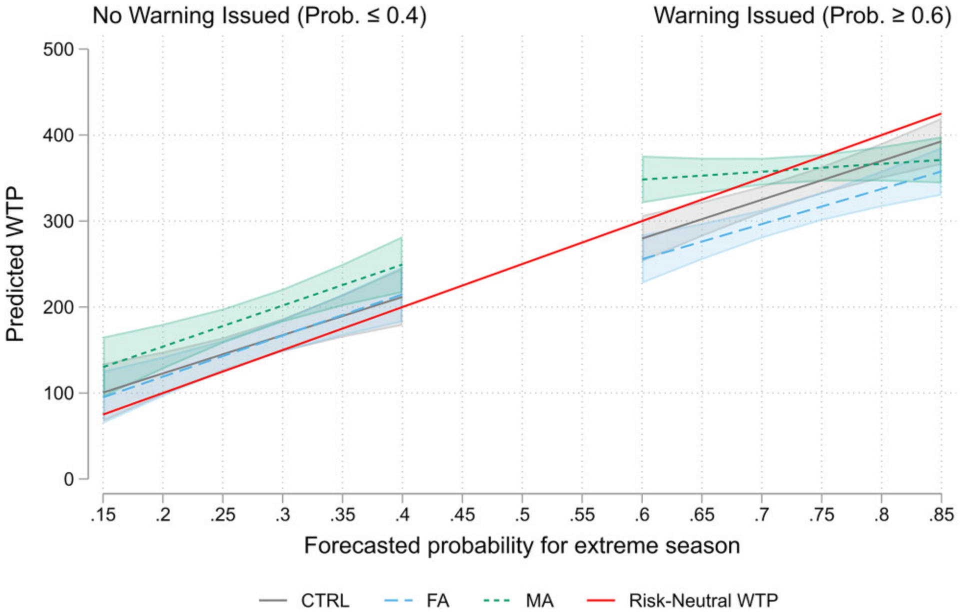


Figure 4. Predicted mean WTP in season 10 per treatment with 95 per cent confidence interval bands. CTRL treatment is the control treatment with accurate forecasts, FA the false alarm-prone and MA the missed alarm-prone treatment. The red line shows the theoretically predicted WTP for a risk-neutral, rational individual for reference. The left section refers to the cases if no warning was issued with a forecasted probability for an extreme season below 0.4. The right section refers to the cases if a warning was issued with forecasted probability for an extreme season above 0.6. See Table S9 for the corresponding Tobit regression models.

the predicted WTP in FA and CTRL is nearly parallel indicating that the sensitivity to the probability in both treatments is similar (Fig. 4, right section). In contrast, the slope of MA compared to both CTRL and FA is flatter, and thus WTP is, on average, less sensitive to the forecasted probability if individuals receive a warning in season 10 in MA. The corresponding interaction term between the MA treatment and the forecast probability in Table S9, Models 3 and 4, is negative and statistically significant as opposed to the remaining treatment-probability interactions. In the MA treatment, WTP is generally higher independent of the forecast probabilities (the MA cross-effect found above).

Our results on the sensitivity to forecast probabilities emphasise the importance of communicating probabilities as part of forecasts (LeClerc and Joslyn 2015; Taylor et al. 2015). Even if repeated false alarms lead to lower investment in adaptation if a warning is issued, individuals still increase their investment in adaptation with higher forecasted probabilities of extreme climate conditions (Fig. 4, right section). This remaining sensitivity to the forecast probabilities potentially limits their economic losses as high probability forecasts still motivate individuals to invest more in adaptation, even if they are exposed to a false alarm-prone system. However, individuals with a history of missed alarms are less likely to adjust their adaptation to the forecasted probabilities if they receive a warning, which makes them prone to overshoot their investments and potentially leads to welfare loss in the long run. The relative advantage of probabilistic forecasts as discussed in the literature (Fundel et al. 2019; LeClerc and Joslyn 2015; Losee and Joslyn 2018) may therefore not materialise if forecast systems have provided frequent missed alarms in the past.

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

4. Conclusions

This paper uses an online experiment to explore the extent to which exposure to inaccurate forecast systems affects individual adaptation decisions in response to forecasts. To this end, we systematically manipulated the accuracy of the forecast systems that result in the repeated experience of either false or missed alarms and explore: (1) the existence and magnitude of both main effects and cross effects of these inaccuracies, (2) the persistence of these effects, and (3) the magnitude of these effects relative to the responsiveness to forecasted probabilities.

Overall, we find evidence that exposure to inaccurate forecast systems affects individuals' responsiveness to forecasts. The first main finding of this study is related to the systematic analysis of false and missed alarm-prone forecasts on behaviour when a warning and no warning is issued (relative to a more accurate forecast system). False alarm-prone forecast systems decrease individual adaptation investment if a warning is issued (the *cry-wolf effect*). In contrast, missed alarm-prone forecast systems increase individual investments irrespective of whether or not a warning is issued. Disentangling the effect of false and missed alarm histories, we find that false alarms have a relatively persistent negative impact on adaptation behaviour, while more frequent missed alarms affect behaviour relatively stronger than earlier ones.

Taken together, these results suggest that there are systematic differences in the mechanisms how false and missed alarms affect behaviour. One compatible explanation would be that false alarms lead individuals to update their beliefs about the accuracy of the warning. As such, only trust in the warning itself is affected, but not trust in the forecasted probability. As a result, individuals are less likely to comply with an issued warning but still rely on the forecast in the absence of a warning. Individuals exposed to missed alarms, on the other hand, potentially update their beliefs about the forecasted probabilities and assume that the forecasts generally underestimate the risk of an extreme season. Further research is warranted to systematically test this hypothesis. Ideally, such research would elicit the beliefs about the accuracy of the warnings and forecast probabilities separately, to assess if false and missed alarm-prone forecast systems have different impacts on the two types of beliefs. This would also allow to explore how beliefs ultimately affect adaptation behaviour. In the broader literature, it has also been debated to what extent individuals attach more weight to positive than negative signals in belief updating (e.g., Coutts 2019; Sharot et al. 2012), which could provide an interesting avenue for future studies.

Our second main finding relates to the importance of forecast system inaccuracies relative to the impact of the communicated forecast probabilities. We find that the observed treatment effects are relatively small in relation to the effects of the forecasted probabilities. Even if the forecast system is prone to false alarms, individuals respond to an increase in the forecasted probabilities with larger adaptation investments. However, if individuals experience frequent missed alarms, their sensitivity to the forecasted probabilities is affected. Once a warning has been issued, they exhibit relatively high adaptation investments irrespective of the forecast probability. Overshooting adaptation investments is, hence, becoming more likely in cases of low probability warnings.

Our findings provide insights for the design of forecast and warning systems, for example, of extreme weather events such as heavy rain and hurricane warnings, or for seasonal weather forecasts (e.g., for farmers). Firstly, practitioners who are designing forecast systems commonly face the decision whether to decrease the likelihood of missed alarms at the costs of increasing the frequency of false alarms (Brooks and Correia 2018). Survey evidence from potential users of seasonal weather forecasts in Europe indicates that there is a substantial willingness to accept high false alarms rates, in order to be warned of extreme conditions (Taylor et al. 2015). Our research highlights that both types of inaccuracies affect behaviour in the long run, indicating that no panacea exists when designing forecast systems. Whether

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

more frequent false or missed alarms cause more harm in the long run inevitably depends on the case-specific stakes at risk and the level of adaptation costs. Hence, our experiment can provide only limited insights. Nonetheless, it highlights that decision-makers, such as forecast designers should carefully assess the users and their risk profiles, as well as potential losses and adaptation costs before deciding on the communication and design of forecasts and warnings. To this end, also more user-specific research is needed, for example, with farmers as potential users of seasonal weather forecasts.

Secondly, our findings highlight the benefits of using probabilistic forecasts that communicate the probabilistic nature of forecasts. This recommendation is in line with an emerging literature on the advantages of probabilistic instead of deterministic forecasts (Fundel et al. 2019; LeClerc and Joslyn 2015; Losee and Joslyn 2018). Despite exposure to false and missed alarm-prone forecasts, individuals remained overall receptive towards the forecast probability. The higher the likelihood of an extreme season, the higher the adaptation investment. This effect is relatively strong compared to the long-term impact of missed and false alarms.

Lastly, we would like to discuss the external validity of our study. We decided to use an online experiment instead of a lab experiment because the former are known to be more cost-efficient, allow for much larger samples with higher power, and participants are more representative of the general population than students (Palan and Schitter 2018; Peer et al. 2017). Nonetheless, it should be acknowledged that the composition of the participants' sample and the constructed experimental environment limit the external validity of the outcomes of this study (Al-Ubaydli and List 2015). Yet, we see our experiment as a valuable first step to the understanding of general behavioural patterns that are important to understand the implications of forecast inaccuracies on climate change adaptation behaviour. A better understanding of these general behavioural patterns can improve forecast and warning systems' design. In the future, additional (more strongly framed) experiments with users of domain-specific seasonal forecasts or early warning signals (e.g., with farmers) will allow us to expand our knowledge of the implications of forecast inaccuracy on adaptation behaviour.

Acknowledgements

This paper greatly benefited from helpful comments of two anonymous reviewers, Maximilian Burger, Ann-Kathrin Kössler, Matthias Mayer, Marco Nilgen, Juan-Felipe Ortiz-Riomalo, Achim Schlüter, and Ivo Steimanis. We also received helpful comments to the experimental design from our colleagues of the LaER group at Osnabrueck University. Special thanks go to all participants who took the time to take part in our experiment and María Angélica López Ardila for her research assistance. All remaining errors in the manuscript are our own.

Supplementary material

Supplementary data are available at *Q Open* online.

Author contributions

Katharina Hembach-Stunden, Tobias Vorlaufer, and Stefanie Engel devised and designed the project and experiment. Katharina Hembach-Stunden carried out the experiment. Katharina Hembach-Stunden and Tobias Vorlaufer conducted the statistical tests. Katharina Hembach-Stunden, Tobias Vorlaufer, and Stefanie Engel wrote the manuscript.

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Funding

This work was funded by the Ministry of Science and Culture of Lower Saxony (Germany), and by the Alexander von Humboldt-Foundation in the framework of the Alexander von Humboldt-Professorship endowed by the German Federal Ministry of Education and Research. The publication of this article was funded by the Open Access Fund of the Leibniz Association, Alexander von Humboldt Stiftung, Germany, Leibniz-Gemeinschaft, Germany and Niedersächsisches Ministerium für Wissenschaft und Kultur, Germany.

Conflict of interest

The authors declare no competing interests.

Data Availability

The experimental material, replication data and analysis scripts are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/TMESK).

End Notes

[1] Overall, we believe that caution is warranted when generalising findings from automated machine alerts to another behavioural domain. Machine alerts require immediate attention, so it is an intuitive decision made within seconds and often not based on conscious deliberation (e.g., alarms of life support machines in hospitals). In contrast, responses to seasonal forecasts or extreme weather warnings are typically slower and more deliberate (cf Kahneman 2011).

[2] Alternatively, we could have allowed for continuous adaptation decisions that would either reduce the risk of being affected by an extreme season or the damage caused by an extreme season. However, this would require explaining the function relating investment amounts to risk/damage reduction and would have increased the complexity of the experiment. We therefore opted for the BDM method, which allows to elicit a continuous measure with a relatively simple mechanism.

[3] To balance the *a priori* probabilities such that participants in all three treatments have a chance to experience all four forecast cases (accurate warning, false alarm, accurate no-warning and missed alarm) and to have a level of forecast accuracy similar to real-world forecast systems which is around 33 per cent (National Institute of Water and Atmospheric Research (NIWA) 2016), we match option 11 and 12 in the FA treatment with underrating forecast probabilities and option 1 and 2 in the MA treatment with overrating forecast probabilities.

[4] The false alarm prone system is more sensitive (likelihood to issue a warning prior to an extreme season) with 72 per cent than specific (likelihood to issue no warning prior to a normal season) with 6 per cent. The numbers for the missed alarm-prone system are the opposite (sensitivity 6 per cent, specificity 72 per cent). Both forecast systems have the same accuracy with 39 per cent (likelihood to issue correct warnings and no-warnings) which is lower than in CTRL (72 per cent).

[5] Individuals also face a third uncertainty, namely which season is payout-relevant. This is determined by a lottery at the end of the experiment. Since the probabilities are evenly distributed and this is known by participants, we assume this uncertainty to not affect individual decision-making apart from incentivising the decision in each season. The actual loss of income and wealth in the real world that is hence not present in our experiment may influence future decisions regarding individuals' responses to forecast systems as well. However, this aspect is not the focus of this study and the design of our experiment excludes income and wealth effects as potential drivers.

[6] Our experimental design is related to two streams of experimental research on decision-making under uncertainty in psychology and behavioural economics, but features key differences that may limit the transferability of previous findings. Firstly, a vast number of experiments have focused on the role of biases and heuristics in the context of Bayesian belief updating. These experiments are different to our experiment insofar as our participants do not receive information about possible 'states of the world', the frequency of particular signals in each state, or prior probabilities which state is true (see Benjamin 2019). Furthermore, these experiments typically elicit individuals' *a priori* and *a posteriori* beliefs, whereas we focus on individuals' behavioural response in terms of adaptation behaviour, which

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

is in our view the more relevant outcome when it comes to seasonal forecasts and extreme weather warnings. Secondly, some individuals may erroneously believe that independent events are instead conditional on each other, known as the gambler's fallacy (Rabin and Vayanos 2010). In our experiment, both seasonal outcomes and forecasts are independent from the outcomes and forecasts of the prior seasons. However, in accordance with the gambler's fallacy, individuals could erroneously believe that an inaccurate forecast (or extreme season) is less likely in the future if they have experienced inaccurate forecasts (or extreme seasons) before. This bias has been predominantly observed in individuals facing decisions with risk, where they know the underlying probabilities of events. However, individuals in our experiment do not know the underlying probabilities and instead face several layers of uncertainty, which leads us to assume that the gambler's fallacy is likely not applicable.

[7] We calculated Cohen's d by using the Stata command 'esize' for two independent samples using groups from the Stata 15 software (StataCorp 2017). The subtraction of the means of WTP in season 10 is divided by the pooled standard deviation of the relevant two sub-samples.

[8] Please note that we do not compare the effect of receiving a warning to the effect of an increase in the communicated forecasts. Since forecast probabilities of $\geq$ 60 per cent are always associated with a warning (and vice versa), we cannot disentangle these two effects. We therefore estimate separate models conditional on receiving and not receiving a warning. We can therefore estimate the impacts of an increase in the forecast probability for the two ranges: <60 per cent and $\geq$ 60 per cent.

References

Al-Ubaydli O. and List J. A. (2015) 'On the generalizability of experimental results in economics'. In:Fréchette G. R. and Schotter A. (eds.) *Handbook of Experimental Economic Methodology*. Oxford: OUP.

Becker G. M., DeGroot M. H. and Marschak J. (1964) 'Measuring utility by a single-response sequential method.', *Behavioral Science*, 9: 226–32.

Benjamin D. J. (2019) 'Chapter 2–errors in probabilistic reasoning and judgment biases'. In: Bernheim, B. D., DellaVigna, S., and Laibson, D. (eds.) *Handbook of Behavioral Economics: Applications and Foundations 1, Handbook of Behavioral Economics—Foundations and Applications* 2, Vol. 2, pp. 69–186. Amsterdam, the Netherlands: Elsevier.

Berinsky A.J., Margolis M.F. and Sances M.W. (2014) 'Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys', *American Journal of Political Science*, 58:739–53.

Brooks H. E. and Correia J. (2018) 'Long-term performance metrics for national weather service tornado warnings', *Weather and Forecasting*, 33: 1501–11.

Bruno Soares M., Alexander M. and Dessai S. (2018) Sectoral use of climate information in Europe: a synoptic overview', *Climate Services*, 9: 5–20.

Burgeno J. N. and Joslyn S. L. (2020) 'The impact of weather forecast inconsistency on user trust', *Weather, Climate, and Society*, 12: 679–94.

Chancey E. T. et al. (2015) 'False alarms vs. misses: subjective trust as a mediator between reliability and alarm reaction measures'. *Proceedings of the Human Factors and Ergonomics Society*, 59:647–51. DOI: 10.1177/1541931215591141

Coutts A. (2019) 'Good news and bad news are still news: experimental evidence on belief updating', *Experimental Economics*, 22: 369–95.

Dohmen T. et al. (2011) 'Individual risk attitudes: measurement, determinants, and behavioral consequences', *Journal of the European Economic Association*, 9: 522–50.

Donner W. R., Rodriguez H. and Diaz W. (2012) 'Tornado warnings in three southern states: a qualitative analysis of public response patterns', *Journal of Homeland Security and Emergency Management*, 9: 1547–7355. 1957.

Dow K. and Cutter S. L. (1998) 'Crying wolf: Repeat responses to hurricane evacuation orders', *Coastal Management*, 26: 237–52.

Estes W. K. (1976) 'The cognitive side of probability learning.', *Psychological Review*, 83: 37–64.

Fundel V. J. et al. (2019) 'Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users', *Quarterly Journal of the Royal Meteorological Society*, 145: 210–31. DOI: 10.1002/qj.3482

Hendriks A. (2012) *SoPHIE-Software Platform for Human Interaction Experiments*. Osnabrück, Germany: Working Paper, Osnabrueck University.

IPCC. (2022) 'Summary for policymakers'. In: Pörtner, H.-O., Roberts, D. C., Tignor, M. M. B., Poloczanska, E. S., Mintenbeck, K., Alegría, A., and Craig, M. et al. (eds.) *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 3–33. UK and New York, NY, USA.

Jauernic S. T. and Van Den Broeke M. S. (2017) 'Tornado warning response and perceptions among undergraduates in Nebraska', *Weather, Climate, and Society*, 9: 125–39.

Joslyn S. L. and LeClerc J. E. (2012) 'Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error.', *Journal of Experimental Psychology: Applied*, 18: 126–40.

Katzav J. et al. (2021) 'On the appropriate and inappropriate uses of probability distributions in climate projections and some alternatives', *Climatic Change*, 169: 15.

Kahneman D. (2011). *Thinking, Fast and Slow*. New York: Farrar Straus & Giroux.

Knudson C. and Guido Z. (2019) 'The missing middle of climate services: layering multiway, two-way, and one-way modes of communicating seasonal climate forecasts', *Climatic Change*, 157: 171–87.

LeClerc J. E. and Joslyn S. (2015) 'The cry wolf effect and weather-related decision making', *Risk Analysis*, 35: 385–95.

Lim J. R., Fisher Liu B. and Egnoto M. (2019) 'Cry wolf effect? evaluating the impact of false alarms on public responses to tornado alerts in the Southeastern United States', *Weather, Climate, and Society*, 11: 549–63.

Lindell M. K. et al. (2016) 'Perceptions and expected immediate reactions to tornado warning polygons', *Natural Hazards*, 80: 683–707.

Losee J. E. and Joslyn S. (2018) 'The need to trust: How features of the forecasted weather influence forecast trust', *International Journal of Disaster Risk Reduction*, 30: 95–104.

Malone T. and Lusk J.L. (2018) 'Consequences of participant inattention with an application to carbon taxes for meat products', *Ecological Economics*, 145:218–30. DOI: 10.1016/j.ecolecon.2017.09.010

Manzey D., Gérard N. and Wiczorek R. (2014) 'Decision-making and response strategies in interaction with alarms : the impact of alarm reliability, availability of alarm validity information and workload', *Ergonomics*, 57: 1833–55.

National Institute of Water and Atmospheric Research (NIWA). (2016) 'Improving seasonal climate forecasts'. *2016-02-08*. Retrieved March 11, 2020, from <https://niwa.co.nz/news/improving-seasonal-climate-forecasts>.

Pacchetti M. B. et al. (2021) 'Assessing the quality of regional climate information', *Bulletin of the American Meteorological Society*, 102: E476–91.

Palan S. and Schitter C. (2018) 'Prolific.ac—a subject pool for online experiments', *Journal of Behavioral and Experimental Finance*, 17: 22–7.

Peer E. et al. (2017) 'Beyond the Turk: alternative platforms for crowdsourcing behavioral research', *Journal of Experimental Social Psychology*, 70: 153–63.

Prolific. (2021) 'Prolific'. (Oxford, UK) https://www.prolific.co/ accessed 14 May 2021.

Rabin M. and Vayanos D. (2010) 'The gambler's and hot-hand fallacies: theory and applications', *Review of Economic Studies*, 77: 730–78.

Ripberger J. T. et al. (2015) 'False alarms and missed events: the impact and origins of perceived inaccuracy in tornado warning systems', *Risk Analysis*, 35: 44–56.

Schmidt J. and Bijmolt T. H. A. (2019) 'Accurately measuring willingness to pay for consumer goods: a meta-analysis of the hypothetical bias', *Journal of the Academy of Marketing Science*, 48:499–518.

Schultz D. M. et al. (2010) 'Decision making by Austin, Texas, residents in hypothetical tornado scenarios', *Weather, Climate, and Society*, 2: 249–54.

Sharot T. et al. (2012) 'Selectively altering belief formation in the human brain', *Proceedings of the National Academy of Sciences*, 109: 17058–62.

Simmons K. M. and Sutter D. (2009) 'False alarms, tornado warnings, and tornado casualties', *Weather, Climate, and Society*, 1: 38–53. *American Meteorological Society*.

Smith D. M. et al. (2019) 'Robust skill of decadal climate predictions', *npj Climate and Atmospheric Science*, 2:1–10.

Stainforth D. a. et al. (2007) 'Confidence, uncertainty and decision-support relevance in climate predictions', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365: 2145–61.

StataCorp. (2017) '*Stata Statistical Software: Release 15*'. College Station, TX: StataCorp LLC.

Taylor A. L., Dessai S. and Bruine De Bruin W. (2015) 'Communicating uncertainty in seasonal and interannual climate forecasts in Europe', *Philosophical Transactions of the Royal Society A*, 373: 140454.

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024

Taylor A. L., Kox T. and Johnston D. (2018) ‘Communicating high impact weather: improving warnings and decision making processes’, *International Journal of Disaster Risk Reduction*, 30: 1–4.

Trainor J. E. et al. (2015) ‘Tornadoes, social science, and the false alarm effect’, *Weather, Climate, and Society*, 7: 333–52.

Webber S. (2019) ‘Putting climate services in contexts: advancing multi-disciplinary understandings: introduction to the special issue’, *Climatic Change*, 157: 1–8.

Wharton Credibility Lab. (2017) ‘AsPredicted’. <https://aspredicted.org> accessed 20 October 2021

Wiczorek R. and Meyer J. (2016) ‘Asymmetric effects of false positive and false negative indications on the verification of alerts in different risk conditions’, *Proceedings of the Human Factors and Ergonomics Society*, 60: 289–92.

Zommers Z. (2012) *Climate Early Warning System Feasibility Report: Early Warning Systems and Hazard Prediction*. Nairobi, Kenya: United Nations Environment Programme (UNEP).

© The Author(s) 2023. Published by Oxford University in association with European Agricultural and Applied Economics Publications Foundation. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/qopen/article/4/1/qoad031/7459342 by Leibniz-Zentrum fuer Agrarlandschaftsforschung (ZALF) e.V. - user on 26 March 2024