Environment Systems and Decisions (2022) 42:450–461
https://doi.org/10.1007/s10669-022-09863-1

The false alarm/surprise trade-off in weather warnings systems: an expected utility theory perspective

Ramón de Elía[1]

Accepted: 7 May 2022 / Published online: 28 May 2022

Abstract
Early warning systems for weather events are becoming widespread as technological capacities develop. For warnings to be effective, they must allow enough lead time to deploy protective measures yet the earlier a warning is broadcast the greater may be its uncertainty. In dichotomous warning systems (i.e., warning-no warning), a measure of this uncertainty is the number of wrong messages issued in terms of "surprises" (events missed by the warning system) or "false alarms." Given the range of repercussions of errors of either kind, warning system users can be expected to have different reactions to this uncertainty. Some will cope better with false alarms, others with surprises. This will affect preferences with respect to system sensitivity; that is, the threshold of threat evidence required for the realization of the warning, each threshold having a given false alarm/surprise trade-off. This article adopts an expected utility theory perspective to define different false alarm/surprise trade-offs for users of a warning system. An analytical expression for a cost function is proposed, which under certain conditions depends only on one parameter under the control of forecasters (i.e., number of tolerated surprises). We show quantitatively how optimal trade-offs depend on what is at stake for users and their capability to react to warnings, and how users' varying needs represent a dilemma for a weather service regarding false alarm/surprise trade-off settings. In particular, it is shown that unbiased warnings—a condition often rewarded at the verification stage—do not hold any specific virtue for minimizing losses. A general discussion follows regarding the need to better understand and better communicate this dilemma to policy makers, users, and the public.

Keywords Warning system · False alarm · Surprise · Ambiguity · Expected utility theory · Weather events

1 Introduction

Warning systems are a means of communicating an impending threat to specific users or the general public. In technologically developed societies, warning systems are ubiquitous and range from uses such as alerts for home trespassing, detecting illnesses in automated medical diagnoses, signaling insolvency in banks, to announcing the imminence of a tsunami (Choo 2009). These systems have grown in importance in the last decades, developing into what are now called Multi-Hazard Early Warning Systems (MHEWSs), the aim of which is to centralize information about multiple threats within a single system (WMO 2020).

The Sendai Framework for Disaster Risk Reduction 2015–2030 recognizes the significant benefits of MHEWSs by incorporating them into one of its seven global targets (see UN 2015). In the case of weather services, warnings apply to several different phenomena at different time scales, ranging from tornados in the very-short scale to droughts at a seasonal scale (WMO 2020). In general, warning systems have been shown to be beneficial for the community (see for example Rogers and Tsirkunov 2011).

For a warning system to have any success, it must first and foremost be in fact capable of detecting the threat for which it is designed (for a list of main requirements for an effective warning system see Table 2 of Choo 2009). The degree of this ability and its capacity to assist users to make relevant decisions will in the long term define the value of the system. But all systems are imperfect at least to some degree, and hence irrespective of accuracy, developers and users must basically agree on the parameter of system sensitivity. That is, does the user need or prefer a system in which the alarm goes off as

✉ Ramón de Elía
rdelia@smn.gov.ar

[1] Servicio Meteorológico Nacional, Av. Dorrego 4019, C1425 Ciudad de Buenos Aires, Argentina

soon as any suspicion of threat is perceived ("trigger happy"), or one whose operators "hold their nerve," and require a sizable bulk of threat evidence for system activation? [Note: the colloquial terminology used in much of the literature describing these two states seems to this author to be equivocal. We follow that of Young (2017), which seems more intuitive].

Between these extremes lies a spectrum of possible choices for calibration with a middle position which has the virtue of having no "frequency bias" (ratio of number of predicted events over number of observed events, see Wilks 2006), that is, there are as many warnings triggered as events occurring in reality (although, alas, not all warnings result in events nor are all events forewarned). In the common lexicon of warning systems, this is usually described as the trade-off between false alarms and surprises or misses (see Table 1; see also for example Sättele et al. 2016). Verification studies show that different applications tend to have different trade-off settings, presumably because users are more or less resistant to false alarms and surprises for different kinds of threats (Swets et al. 2000a). This naturally puts the onus at least partially on the organization responsible for issuing the warning, which creates for the issuer the challenge to determine a reasonable and justifiable trade-off setting that will ensure the accomplishment of its mission.

During the development phase, it is possible to hedge a warning system toward either of the two extremes mentioned above, yet a truly functional understanding of its behavior can only be gained after a period of use and a thorough verification process. Before a verification stage is reached, other influences may also intervene—for example, through peer/public/user pressure—to shift parameters toward a trade-off different from the initial settings. As a result, warning system calibration is often a result of science-based policy mixed with a public relations component (see Choo 2009). In the case of weather events, it can be seen that many warning systems tend toward the "trigger happy" end of the spectrum favoring false alarms over surprises, presumably because much of the general public is badly affected by unannounced serious weather events (see Brooks and Correia 2018).

For example, after a serious missed event that had far-reaching consequences, Météo-France was encouraged to establish a "doctrine" defining how certain events were to be framed in terms of warnings. Among the targets set, it was established that for all cases assessed within a 24-h forecast, surprises could not be higher than 2% and false alarms could not be greater than 16% (Gillet-Chaulet 2020). Similarly, Brooks and Correia (2018) observed that in 2012, with respect to tornado warnings in the US, an apparent emphasis on reducing FAR led to a change in the threshold for issuing warnings.

A descriptive summary of how weather warnings are calibrated in different systems around the world (i.e., their different false alarm/surprise trade-offs) can help to guide institutions approaching this issue for the first time. But adopting a system calibration that has been strongly affected by a single event, or a small number of highly publicized events, has the weakness of potentially costly and stressful false alarms, and is therefore limited as an argument for new stakeholders. It is in this situation that *prescriptive*—instead of *descriptive*—approaches are needed to find the most appropriate rationale for a weather office to solve this dilemma.

In order to accomplish this, Sect. 2 introduces a conceptual model based on the minimization of a cost function, and different approximations are proposed to simulate cases relevant to weather information users. Section 3 presents results for three cases using a set of different parameter values to establish the robustness of results. In Sect. 4, conclusions and suggestions for future work are discussed.

2 A simple conceptual model

We analyze this dilemma through a prescriptive approach, using the conceptual model discussed by Sättele et al. (2016), and by Didier et al. (2017). A good primer on related theories of risk and expected utility as well as a discussion on their implications can be found in Baron (2008, Chap. 10). There are, however, some differences between the models addressed there and our adaptation for the issue under discussion. As we shall see, the problem is reduced here to its bare bones, based on the minimization of a cost function or risk function, considering only a minimal number of parameters of cost and loss in a given situation and whether a warning was issued or not. It does not consider any other parameter used in standardized warning systems, such as timeliness (see OASIS 2010).

The chosen cost function structure assumes that actions triggered by a warning are not correlated in time or space, and that actions are equally effective every time they are taken.

Table 1 Contingency table for the verification of a simple warning system (see, for example, Wilks 2006)

	Event	No event	
Warning	Hit (correct prediction of event)	False alarm	Total number of warnings
No warning	Surprise or Miss	Correct negative	Total number without warning
	Total number of events	Total number of no events	Total number of cases

2.1 The cost/loss function

Table 2 shows the same as Table 1 but in this time is expressed in terms of probabilities. Using these expressions, we can write a cost function for a warning system as

$$R_W = P(W,E)L_H + P(\neg W,E)L_S + P(W,\neg E)L_{FA} + P(\neg W,\neg E)L_{CN}, \tag{1}$$

where R_W represents the expected cost (or risk) of an event E with the existence of an operational warning system that generates individual warnings W, and whose performance we know (i.e., the probability P of each term is known). Each term refers to the four sectors in the contingency table (Table 2) associated with the subscripts, H for Hit, S for Surprise, FA for False alarm, and CN for Correct negative. The four L parameters refer to the losses or costs associated with each of the four cases. The first term represents, then, a successful warning (an event forewarned, a Hit) and its associated loss L_H. This loss can be thought of as $L_H = L_c + L_p$, where L_c represents the cost of the protective measures taken, and L_p the loss suffered despite those measures. The second term represents the cases when the event occurs without forewarning (surprise), where the associated loss is L_u, the result of absence of protective measures. The third term (false alarm) describes the activation of protective actions with their associated costs, in the absence of the event, and this value is identical to the cost L_c, introduced above. The fourth term represents the system when no event is registered and no warning has been activated—which is most of the time—and we can take advantage of this term to represent basic and ongoing costs of system development and maintenance as in Didier et al. (2017). This definition of losses and costs is somewhat simpler than what is presented by Lopez et al. (2020), who considered more specific situations such as that "the additional cost of transporting back to headquarters non-perishable food that had been prepositioned, would be an additional cost of acting in vain that would not have been incurred if the extreme event had occurred." This simplification, we believe, is an advantage to better understand the issue.

With these definitions, expression (1) can be rewritten as

$$R_W = P(W,E)\left(L_c + L_p\right) + P(\neg W,E)L_u + P(W,\neg E)L_c + L_{\text{sys}}, \tag{2}$$

where L_{sys} corresponds to the cost of the development and maintenance of the system during the period of exploitation, independent of event activity. Applying properties of the conditional probabilities we can rewrite (2) as

$$\begin{aligned} R_W = {} & P(W|E)P(E)\left(L_c + L_p\right) + P(\neg W|E)P(E)L_u \\ & + P(\neg E|W)P(W)L_c + L_{\text{sys}}. \end{aligned} \tag{3}$$

Table 2 Contingency table for the verification of a simple warning system written in terms of joint and marginal probabilities, where the symbol "¬" stands for "no". For example, the joint probability P(W,¬E) is to be interpreted as the joint probability of warnings (W) which coincided with no events (¬E)

	Event	No event	Marginal probability
Warning	$P(W, E)$	$P(W,\neg E)$	$P(W)$
No warning	$P(\neg W, E)$	$P(\neg W, \neg E)$	$P(\neg W)$
Marginal probability	$P(E)$	$P(\neg E) = 1 - P(E)$	1

The conditional probability in the first term can be taken from verification exercises as being equal to the probability of detection (POD), which is defined as hits/(hits + surprises) in Table 1 (see Wilks 2006). The POD—also called Hit rate—is directly associated with the rate of surprises (simply 1-POD), and at the same time a very popular categorical statistic. The probability factors in the second term can be rewritten as $(1 - P(W|E))P(E) = (1 - POD)P(E)$. The conditional probability in the third term can be associated with the False Alarm Ratio (FAR which is defined as false alarms/(hits + false alarms), not to be confused with False Alarm rate, used in the ROC diagram –see Wilks 2006). In addition, given that

$$\frac{P(W)}{P(E)} = \frac{\text{POD}}{1 - \text{FAR}}, \tag{4}$$

(which can be obtained from Table 2 and the definitions of POD and FAR), we finally get

$$R_W = P(E)\left[\text{POD}\left(L_c + L_p\right) + (1 - \text{POD})L_u + \frac{\text{FAR POD}}{1 - \text{FAR}}L_c\right] + L_{\text{sys}}. \tag{5}$$

It is interesting to note that this expression can also be derived with identical results without the need for "correct negatives" of the contingency table (Table 1). This is important because many warning verification studies ignore the "correct negatives," since they are often ill-defined (see, for example, discussion in Stephenson et al. 2010).

2.2 Preliminary analysis of the cost/loss function

Expression (5) is the main tool for what follows, particularly the expression between square brackets. The first thing to notice is that among the parameters in the expression, neither $P(E)$ nor L_{sys} play a part in the minimization of the cost/loss function once the event is defined and the warning system available. This should not be interpreted as saying that the term between square brackets is independent of the event since losses are usually correlated with the rarity of an

event. The term L_{sys} is, for reasons of its independence from the management of the system sensitivity, also not affected by the minimization of the cost function.

Given a warning system that triggers an intervention for each warning emitted, each with its estimated losses and costs, we can see that the minimization of (5) is, as expected, a function only of POD and the false alarm ratio FAR.

Let us suppose that some years into operation, the verification process produces a given pair of (POD, FAR) values. This pair of values tells us that, overall, the system displays a certain trade-off, perhaps one that is desired or one that is simply accidental. The question that may arise is whether this trade-off between POD and FAR is in fact optimal for the threat in question with its associated losses and capabilities of response. To answer this question, we need to know what path in a given warning system the pair (POD, FAR) follows when we vary the evidence threshold to trigger a warning. For example, the more demanding we become in terms of evidence prior to warning emission, the more we reduce incidence of false alarms (FAR), but also the hits (POD). But at what rate each?

In practice, what we want is for FAR to be written in expression (5) as a function of POD and a parameter of quality that describes changes in FAR given a change in POD for a specific warning system. There is no obvious theoretical expression that relates POD and FAR, hence addressing this question requires that we make some assumptions.

Once we define this relation, we will have a function depending on five parameters: the three factors defining the losses, the skill parameter of the warning system (see next subsections), and finally POD, the only parameter in which the forecasters can have a clear influence by simply changing the threshold—through measures that may be subjective or objective—used to trigger warnings.

2.3 The POD–FAR relation

Here, we will introduce two different versions of this relation, each one with a specific rationale, the aim being to try to sweep a number of families large enough to make our results more general. A number of other relations were also tested but these did not add new insights.

2.3.1 The CSI isoline

In the verification of warning systems, the Critical Success Index [also known as Threat Score and defined as CSI = hits/(hits + false alarms + misses)] is widely used, mostly for being independent of the category of "correct negatives" but also because of its presence in the "Performance diagram" (Roebber 2009), a much used visualization tool. However, the CSI is not always the best choice for verification purposes (see for example Schaefer 1990) which has led some users to adopt other scores. For the case under study, we will see that it has some interesting properties. Following Roebber (2009), we can see that for a given value of CSI we get

$$\mathrm{FAR} = 1 - \frac{1}{\left(\frac{1}{\mathrm{CSI}}\right) - \left(\frac{1}{\mathrm{POD}}\right) + 1},$$

where CSI lies in the interval [0,1], with the upper limit defining a perfect warning system. We can then rewrite expression (5) as

$$R_W = P(E)\left[\mathrm{POD}\left(L_c + L_p\right) + (1 - \mathrm{POD})L_u + \left(\frac{\mathrm{POD}}{\mathrm{CSI}} - 1\right)L_c\right] + L_{\mathrm{sys}}. \tag{6}$$

Figure 1 illustrates different curves of CSI in a Performance diagram, defining possible paths of the POD–FAR relation. A notable characteristic of this parametrization is that expression (6) is now linear in POD.

2.3.2 The power law

Despite the advantages of obtaining a linear relation with the previous parameterization illustrated in Fig. 1, many real-life examples show that, to the contrary, the relation between POD and FAR does not seem to follow isolines of CSI. Using ensemble forecasting, which produces probabilities for any desired forecast parameter, several authors have varied the warning trigger threshold to estimate the POD-FAR relation (see, for example, Adams-Selin et al. 2019; Flora et al. 2019; Gagne et al. 2017, 2019).

One distinctive issue lacking in the previous parametrization is that—irrespective of the warning system quality—when POD becomes very close to 1, FAR should continue

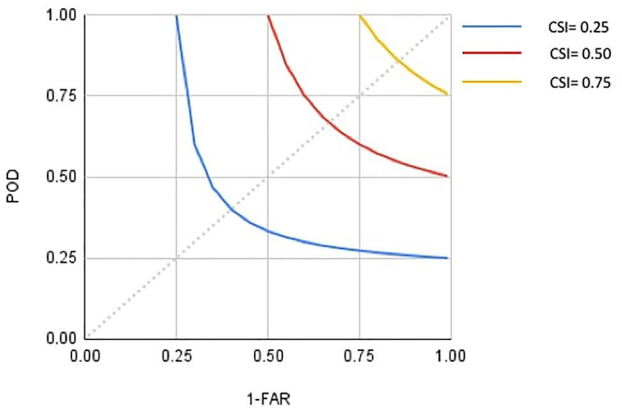

Fig. 1 CSI-based POD–FAR relation represented in a Performance Diagram (POD versus 1-FAR). Colored lines (CSI isolines) are paths for the relation considering that increasing or decreasing POD does not impact the CSI. The highest skill is reached in the upper right corner of the performance diagram

increasing. That is, a desired property of the curve should be that when we demand a very small level of missed events, false alarms have to be high too. But to what extent? This depends on the quality of the warning system, but we want that our parameterization represents that FAR → 1 when POD → 1, each approaching at a different speed depending on the skill of the system. This is a very important issue, because the third term in expression (5) diverges as FAR → 1.

There are in fact many functions that exhibit this characteristic and the authors have tested a large variety of such functions. We have chosen here to illustrate the one with the simplest form that allows us to control the speed of approximation of FAR to 1, when POD reaches the vicinity of 1. A simple expression is $FAR = POD^r$, where r is a positive number. Figure 2 shows its aspect in a performance diagram.

For $r = 1$, we can see that both approach at the same speed. When $r > 1$, we obtain a POD greater than FAR. The opposite is the case when $r < 1$. A larger r represents systems with more skill.

2.4 Losses and costs

As we have said, our focus here is on loss minimization. And as also mentioned earlier, the losses present in this function have their clearest interpretation when measured in terms of monetary units, or even in any other quantitative unit that allows for algebraic operations. Beyond this simple interpretation, if used for considerations of losses of lives or other non-quantifiable impacts, the cost function becomes on the one hand more useful, but on the other much less precise (for a pledge to use cost functions as a tool for a thought process, see Sunstein 2000). In what follows we will pursue a predominantly quantitative approach but we will also consider some broader interpretations.

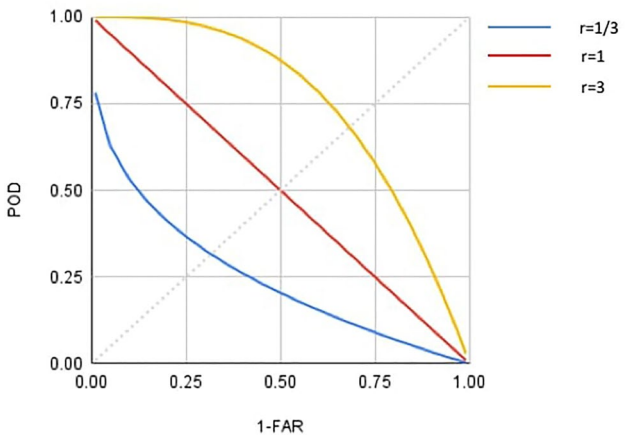

Fig. 2 Power-law based POD–FAR relation represented in a Performance Diagram (POD versus 1-FAR). Colored lines are paths for the relation considering that increasing or decreasing POD maintains r constant. The highest skill is reached in the upper right corner of the performance diagram

As we saw above, the standing cost of the system [last term in expression (5)] does not play a role in loss minimization of the kind attempted here, so although it is at the core of many decisions regarding development of warning systems, it will not be further discussed here. The other three losses are relevant to the minimization.

We will not dwell on the details nor on the role of these losses in the cost function, but it is important to keep in mind that they are interrelated. For example:

1. It is reasonable to have a costly warning system only for grave losses L_u that can be diminished by a timely warning.
2. The loss of a forewarned event L_p should be smaller than that provoked by a surprise event L_u.
3. The cost of a protective intervention L_c should be lower than the loss associated with a surprise event L_u.

The cases discussed in the results section take these relations into account.

It is important to mention that the cost of an intervention with protective measures may have a subjective component. For example, on learning of a hailstorm warning, some members of the public may react by covering their cars with some material already at hand. This is probably a near cost-free action. But, if the warning arrives in the middle of the night, already under rain, some owners may wonder whether the action is worth the effort. An economist might suggest trying to estimate the cost of such an action by simply imagining how much money the person would have paid in order not to perform the work themselves. This is usually referred to as "willingness to pay" (see Baron 2008). This interpretation is useful to understand that some apparently cost-free interventions can eventually be represented in this formulation by specific non-zero values. And that this action carried out by a large population may be finally represented by a large overall cost.

2.5 The impact of repeated false alarms

The discussion in the previous sections assumes, as a prescriptive approach, that users follow warnings in an identical way whether the warning system is very accurate or if it generates a large number of false alarms. But, as Breznitz (1984) puts it, ''Each false alarm reduces the credibility of a warning system. The credibility loss following a false alarm episode has serious ramifications to behavior in a variety of response channels.'' This seems so obvious that we generally assume as a fact that the public will tend to react like villagers in the fable *The boy who cried Wolf* from Aesop,

who, tired by the boy's false cries, fatally decided not to protect the sheep.

Simmons and Sutter (2009) found that for tornado warnings, there is a strong relation between FAR and fatalities as well as injuries. Regarding hurricanes, Hallegatte (2012) states that the evacuation prior to Katrina in 2005 was hampered by unnecessary ones associated with hurricanes George [September 1998] and Ivan [September 2004]. Citing several sources, Sättele et al. (2016) argue that, furthermore, frequent false alarms can lead to excessive intervention costs as well as reduce compliance with future warnings. LeClerc and Joslyn (2015) presented a thorough review of the topic and carried out an experiment that showed that very high false alarm rates led to inferior decision making.

However, other studies, in different contexts, show less public sensitivity to FAR. For example, Dow and Cutter's (1998) study of hurricane evacuation found no evidence of a lower evacuation rate for hurricane Fran in 1996, which occurred just weeks after a false alarm evacuation for hurricane Bertha. Later, Barnes et al. (2007) commented that ''Evidence for the cry-wolf effect in natural hazards research, however, has not been forthcoming," and suggest that the public is by and large not dissuaded from action by known false alarms. Lately, Lim et al. (2019) after a study of tornado warnings in the southeastern United States suggest that concerns about high false alarm ratios generating a complacent public may be overblown. Trainor et al. (2015) go further and exhibit the varied definitions of "false alarm" among the public, arguing and showing evidence of the difference between the perceived false alarms and actual false alarms. This difference is neither negligible nor only of academic interest, since the cry wolf effect is dependent on the former.

As we can see, the existence, impact, and measurement of a "cry-wolf effect" are still an open debate. Nonetheless, in what follows we will assume its existence and proportionality with actual false alarms to study its probable impact.

In order to include the "cry-wolf effect" in the expression of risk R_W presented in (5), we assume that when false alarms abound, protective actions are less thorough (both less expensive and less effective).

This effect can be represented by replacing the constants L_c and L_p in (5) by the expressions of the form $L_c^* = (1-\mathrm{FAR})^k L_c$ and $L_p^* = L_p + \mathrm{FAR}^k(L_u - L_p)$. The exponent k only controls the public´s tolerance to false alarms, with larger values of k representing the more intolerant ones. The cost L_c^* now diminishes for increasing FAR, while the protective loss L_p^* nears the value of the unprotected loss L_u as FAR approaches 1. For the sake of simplicity, we will take $k=1$, which implies a rather high level of mistrust toward the warning system. This choice will help us appreciate clearly its impact in the user's decision making. With this we can rewrite (5) as

$$R_W = P(E)\big[\mathrm{POD}\big((1-\mathrm{FAR})L_c + L_p + \mathrm{FAR}\big(L_u - L_p\big)\big) + (1-\mathrm{POD})L_u + \mathrm{FAR}\,\mathrm{POD}L_c\big] + L_{\mathrm{sys}}. \tag{7}$$

As in expression (5), we also need to write FAR as a function of POD.

3 Results

In what follows we present three distinct cases where the cost function is defined by three different sets of costs and losses. At the same time, each case will be treated with the two POD–FAR relations presented in Sects. 2.3.1 and 2.3.2. Since each of these functions is associated with a parameter, three different values are presented for each parameter in order to span the more important characteristic of its behavior.

The three cases present different combinations of losses that can be associated with typical real-life events. In order to simplify our treatment, we have assumed a fixed value to unprotected losses L_u, which is relatively independent from the warning system management, while L_c and L_p will be modified. Terms such as "expensive" or "ineffective" refer, then, to the relative cost of a measure and its relative capacity to reduce losses as compared to L_u.

The three cases for different sets of cost/loss are:

1. Inexpensive but effective response measures: An example of such a case would be a warning system against hailstorms, where users put their cars under cover. For this case, we have chosen the triad$L_p = 10, L_u = 100, L_c = 5$.
2. Expensive, effective response measures: An example of such a case would be a warning system against frost in sensitive crops, where the users turn on heaters as protective measures (see Snyder and Melo-Abreu 2005). For this case, we have chosen the triad $L_p = 10, L_u = 100, L_c = 50$.
3. Inexpensive but barely effective response measures: An example of such a case could be a warning system against strong winds, calling for the public to take minor mitigating steps to cover their windows. For this case, we have chosen the triad $L_p = 50, L_u = 100, L_c = 5$.

Table 3 lists the parameters used in the mentioned experiments.

Table 3 List of parameters used for solving the cost function defined in (5) for each of the cases analyzed. The three columns on the left define the losses and costs, while the two on the right display the different values of the parameters defining the POD–FAR relation

	L_p	L_u	L_c	Parameter representing skill	
				CSI (as defined in 2.3.1)	r (as defined in 2.3.2)
Case 1	10	100	5	0.25; 0.50; 0.75	1/3; 1; 3
Case 2	10	100	50	0.25; 0.50; 0.75	1/3; 1; 3
Case 3	50	100	5	0.25; 0.50; 0.75	1/3; 1; 3

3.1 Case 1: Inexpensive but effective response measures

The upper panels of Fig. 3 show the core of the cost function (factor between brackets) presented in expression (5) for Case 1 as defined in Table 3. The left panel depicts the linear relation through the use of the CSI-based POD–FAR relation defined in expression (6), while the right panel illustrates the cost function under the power law discussed in Sect. 2.3.2.

The linear relation (upper left panel of Fig. 3) gives an idea of the general behavior of the curve. For the different levels of skill plotted—different values of CSI—it can be seen that maximizing the POD minimizes losses, while minimizing FAR (POD near minimum) seems to be the worst strategy. This simply tells us that for the case of inexpensive but effective response measures, the best strategy is to *over-forecast as much as you can* and cope with false alarms. This is quite an unnatural result due to the linearity of the expression.

The figure also tells us that, as expected, the more accurate the warning system is—where CSI is closer to 1—the less will be the overall cost for any particular POD chosen. It is interesting to see, though, that according to this figure, increases in the warning system skill in fact bring very little gain (shown by the proximity of the yellow and red lines). This would suggest that there may be a limit beyond which further improving the warning system does not bring practical gains.

The upper right panel presents the cost function for the power-law POD–FAR relation and here it can be seen that, as in the linear relation, the better option is still to refrain from minimizing FAR. Unlike what was seen in the left-hand panel, however, maximizing POD now seems to be a very bad strategy. The vertical colored lines show the value of POD that a system should have if no frequency bias is acceptable (if the average number of events must coincide with the average number of predicted events). As we can see for the three different values of warning system skill in this figure, the minimum of the cost function is clearly still to be found on the *over-forecasting* side, that is, with a positive frequency bias ("trigger happy").

Compared to the linear solution, the power law POD-FAR suggests not only a strategy change of POD-FAR trade-off for different skill levels (minima are obtained at different levels of POD), but also that the minimum of the cost function in fact reacts strongly to improvements in warning system skill.

3.2 Case 2: Expensive and effective response measures

The middle panels of Fig. 3 show the core of the cost function presented in expression (5) for Case 2 as defined in Table 3. The left panel depicts the linear relation through the use of the CSI-based POD–FAR relation defined in expression (6), while the right panel uses the power law discussed in Sect. 2.3.2.

Contrary to what was observed in Case 1, and for all three different levels of skill plotted, the middle left panel shows that the lowest cost for the operations is associated with a minimized POD, whereas maximizing FAR (POD near maximum) seems to be the worst strategy. This is simply telling us that for the case of expensive and effective reactive measures, the best strategy is to avoid false alarms. It is interesting to note that, according to this formulation, the better the warning system (higher CSI), the lower its sensitivity to the trade-off between POD and FAR. Other experiments show that this is a consequence of the set of losses chosen, and we will develop this further in the next section.

The middle right panel presents the power law POD–FAR relation and it can be seen that, as in Case 1, the impact of the diverging term for POD approaching 1 is large. Still, minima location is dependent on the quality of the warning system (value of r), moving toward higher levels of POD for better systems.

The vertical colored lines again show the value of POD for a warning system without frequency bias. As we can see, the three variations displayed in this figure indicate that the correct strategy for loss reduction is to *under-forecast*, that is, a negative frequency bias (you should "hold your nerve").

It is also worth noticing the large sensitivity to the quality of the warnings. As we can see, the term associated with the response measures in expression (5) is the only dependent on forecast quality (forecast quality being the link between FAR and POD), and in this case, L_c takes the largest value of all cases studied.

3.3 Case 3: Inexpensive but barely effective response measures

The lower panels of Fig. 3 show that the behavior here is very similar to that of Case 1, both for the CSI-based

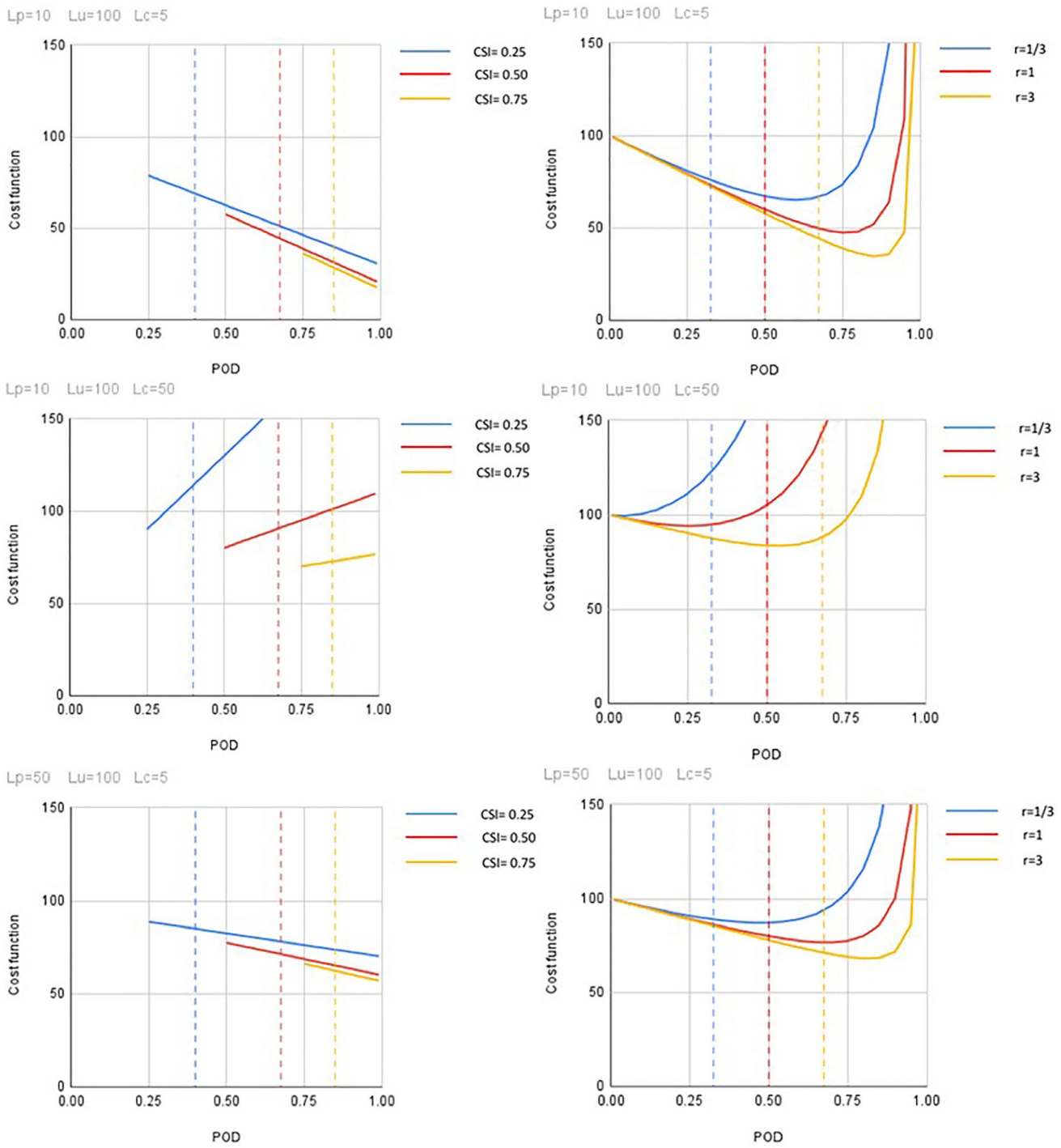

Fig. 3 Cost function as defined in (5) for the three cases under study (rows, from top to bottom). Left side shows results for the use of the CSI-based POD–FAR relation, while the right side shows results for the power-law based POD–FAR relation. Each panel depicts three curves in different colors obtained with different parameters of the POD–FAR relation used. The vertical lines of the same color define the POD values corresponding to an unbiased warning system. To the left of these lines, the system under-forecasts ("holds its nerve"), and to the right over-forecasts ("trigger happy"). Note that only the factor between brackets of the cost function defined in (5) is presented

POD–FAR and for the power-law POD–FAR relation, and for that reason will not be discussed in detail. One visible difference is that before POD nears 1, the cost functions are less dependent on the trade-off between POD and FAR.

From expression (6), we can see that there exists a combination of cost and losses and CSI for which the cost function is fully independent of the POD-FAR trade-off. This can be expressed as

$$L_c\left(1+\frac{1}{\mathrm{CSI}}\right)+L_p=L_u. \tag{8}$$

This expression allows for a large combination of cases, but clearly eliminates response measures that are both inexpensive (relatively low L_c), and effective (relatively low L_p) as an option (our Case 1).

3.4 Impact of repeated false alarms

As we can see on the right panels of Fig. 3, it is highly suboptimal under most conditions for a warning system to produce a large FAR. In practice, as discussed in Sect. 2.5, if overwhelmed by false alarms, it can be expected that users may modify their reactions to the warning system. Consciously or unconsciously, they may behave like the villagers in the fable and become less thorough in the application of protective measures.

Figure 4 depicts the same curves as seen in the solid lines of the top right panel of Fig. 3, but now also plotted is a representation of a strongly distrustful public (dotted lines). As we can see, the desired effect of reducing the undesired costs of a large number of false alarms by not paying attention to them is successful. Still, the impact on the cost function is quite large and the net effect is to drastically reduce

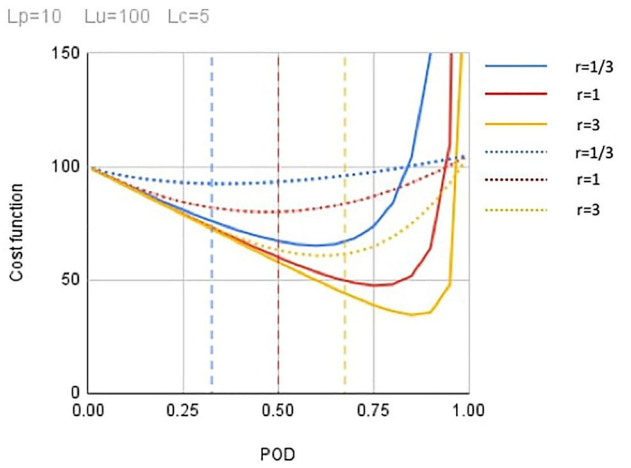

Fig. 4 Cost function as defined in (5) for two different cases. Solid lines depict the same curves as seen in the top right panel of Fig. 3. Dotted lines show the impact on these of a user whose protective actions are dampened due to abundance of false alarms

the benefit of the warning system—shown by the higher cost-function values overall—and to move the optimal POD toward lower values. It is interesting to see, furthermore, that for the lowest skill (r = 1/3), the cost function has a near-flat shape, indicating that the warning system has lost both its sensitivity to POD and its practical value.

As was discussed in Sect. 2.5, we have set the "cry-wolf effect" to quite a high level to explore its effects, and hence, a more realistic case would fall somewhere between the curves with and without the effect.

4 Conclusions and future work

We presented here a mathematical expression that is a function of costs, losses, and POD and FAR scores, with the aim of analyzing the false alarm/surprise trade-off for an early warning system. In order to write the cost function with a single forecaster-controlled parameter, we have chosen different mathematical relations between POD and FAR.

Of the two POD–FAR relations presented here, the first is based on the CSI score, and generates a cost function linear in POD. The second involves a power-law relation that takes a more realistic approach and shows a very significant impact in the general shape of the cost function.

The linear representation shows that minima in the cost function are reached either by issuing "trigger happy" warnings or by the opposite approach, with forecasters "holding their nerve," depending on the set of costs and losses chosen. The more realistic power-law relation between POD and FAR modifies the linear results, making them less prone to extremes while maintaining their general slant.

Where the relation between protective costs and expected losses is such that a response measure can be described as "inexpensive but effective," over-forecasting seems an optimal approach. The same is seen to be the case even for "inexpensive but barely effective" response measures. Where the relation between protective costs and losses can be described as "expensive and effective," however, under-forecasting would appear to be the optimal approach.

A tendency to set system thresholds toward one or the other extreme has already been noticed. As Swets et al. (2000b) put it, "… a high prevalence of a problem in a population or a large benefit associated with finding true cases generally argues for a lenient threshold; conversely, a low prevalence or a high cost for false alarms generally calls for a strict threshold."

As other researchers in this field have found, we find that *unbiased* warnings systems—neither under nor over forecasting—do not have any general "virtue" for minimizing losses. That is, an equality between the number of real events and the number of warnings does not necessarily represent a convenient feature for a warning system. For those interested

and trained in numerical forecasting, where modeled phenomena (e.g., extremes) should have the same recurrence rate as those observed in reality (see for example Guan and Zhu 2017), this may be somewhat surprising.

In this simple analysis, the optimal trade-off between POD and FAR was shown to be case dependent—some users need "trigger-happy" warnings, others that forecasters "hold their nerve"—which suggests that weather services do not have an easy choice. They have at least three different options: they can satisfy some users and frustrate others (see for example Samenow 2013), with the former not even knowing that they are receiving preferential treatment; they can opt for an intermediate position, perhaps choosing to have no frequency bias; or they can try to build one general cost function that maximizes the overall benefit to the community (although this is probably a daunting and futile task). An example of the first case, in which a warning system setting focuses on a specific user in northwestern Peru can be found in Lopez et al. (2020).

Our study of the impact of large FAR values, in the context of a distrusting public, suggests that if a warning system could guarantee an optimal POD-FAR trade-off (that is, the minimum in the cost function), strict user compliance would be the rational action. When this is not the case, it may in fact be rational for individual users to mistrust warnings and not to comply with recommended measures. It is important to recall that we have assumed here a strong relation between perceived and actual false alarms; when this is not the case, as, for example, in the cases described by Trainor et al. (2015), our results may not apply. Further, these authors also caution that "even simple concepts like false alarm are significantly more complex than they appear, and good policy needs extensive, detailed analysis to understand these phenomenon and in turn their implications." Future studies should take into consideration how this could affect simple models like the one presented here.

Given that a warning system operator cannot be fully cognizant of the stakes of any particular user—i.e., their personal cost function—he/she cannot act fully in the user's best interests. It is up to the user to "debias" the warnings—and this is what distrustful users are trying to do—in order to make generic warnings more useful for their individual needs. The responsibility of producing an optimal warning for each user cannot, hence, be transferred solely to the institution in charge of issuing the warning. This is unfortunate, and is not a very satisfying conclusion for the weather office or for the public.

There are, however, ways in which the repercussions and perceptions of a warning system within a given community could be better understood by the institutions issuing those warnings. Through social networks, perhaps a proxy of the cost function of an entire community can be found by "measuring" user complaints and modifying warning system settings in response (as if paraphrasing Baron's (2008) "our choices reveal our utilities", with "our complaints reveal our utilities"). Consciously or not, this is in fact what many weather services already do in an effort to avoid negative publicity. And perhaps, this attitude is not as unsophisticated as it may seem at first sight; it is clear, for example, that although the losses in expression (5) can be thought of in terms of monetary units, this does not preclude other interpretations such as a general malaise, for which an amount of complaining "tweets" can become a proxy (see, for example, social media data analytics carried out by Lee and Kim 2020).

The experiments presented here and the rationale behind them seem to share some objectives with that of "Probabilistic forecast games." A number of these games are being developed at the Technology Collaboration Program in the wind power division of the International Energy Agency (IEA Wind TCP). Under Task 36, Work Package 3, devoted to the optimal use of forecasting, they exploit the potential of games for experimentation in decision-making (see Giebel et al. 2021). The analytical model presented here could be used in a game format to further the needs of weather services to communicate the forecaster dilemma.

It is important to recall that the dilemma discussed here is solely the product of a dichotomous form of broadcasting warnings (warning, no warning). This issue is one among the many reasons why some would rather shift to probabilistic warnings, leaving the final decision—to act or not to act depending on a probability threshold—to the individual receiving the message (see for example; Roulston and Smith 2004; Joslyn and LeClerc 2012, Fundel et al. 2019). A number of studies suggest that users with a business background can take advantage of probability forecasting (see, for example, Howard et al. 2021 and Howard et al. 2022). Even for cases with imminent threats to public safety—where at first sight a probability format may seem to promote hesitation—studies have shown promising results (see for example Miran et al. 2018).

However, it should be kept in mind that probability warnings open the door to the difficulty of understanding the slippery concept of probability (see de Elía and Laprise 2005), as well as to the psychological and cognitive biases not fully controlled by those in charge of issuing the warning (see, for example, Chater et al. 2020).

These obstacles may hinder probability forecasting from becoming a universal solution, and the POD-FAR dilemma seems to be unavoidable for some cases. For the time being, then, weather services must do their best to arrive at good system settings given what they know of their user communities, and make all possible efforts to explain to policy makers, specialized users, and the general public the imperfect trade-off between surprises and false alarms.

Acknowledgements The author would like to thank his colleagues at the Servicio Meteorólogico Nacional in charge of producing weather warnings for stimulating discussions. I would like to thank Dr. Jillian Tomm for her contribution to make this text more readable. Comments by three anonymous reviewers have helped to improve and clarify the manuscript.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Adams-Selin RD, Clark AJ, Melick CJ, Dembek SR, Jirak IL, Ziegler CL (2019) Evolution of WRF-HAILCAST during the 2014–16 NOAA/hazardous weather testbed spring forecasting experiments. Weather Forecast 34(1):61–79

Barnes LR, Gruntfest EC, Hayden MH, Schultz DM, Benight C (2007) False alarms and close calls: a conceptual model of warning accuracy. Weather Forecast 22(5):1140–1147

Baron J (2008) Thinking and deciding, 4th edn. Cambridge University Press, Cambridge, p 570

Breznitz S (1984) Cry wolf: the psychology of false alarms. Lawrence Earlbaum Associates, Hillsdale, p 265

Brooks HE, Correia J Jr (2018) Long-term performance metrics for national weather service tornado warnings. Weather Forecast 33(6):1501–1511

Chater N, Zhu J-Q, Spicer J, Sundh J, León-Villagrá P, Sanborn A (2020) Probabilistic biases meet the Bayesian brain. Curr Dir Psychol Sci 29(5):506–512. https://doi.org/10.1177/0963721420954801

Choo CW (2009) Information use and early warning effectiveness: perspectives and prospects. J Am Soc Inf Sci Technol 60(5):1071–1082

Dow K, Cutter SL (1998) Crying wolf: Repeat responses to hurricane evacuation orders. Coast Manage 26:237–252

de Elía R, Laprise R (2005) Diversity in interpretations of probability: implications for weather forecasting. Mon Weather Rev 133(5):1129–1143

Didier D, Bernatchez P, Dumont D (2017) Systèmes d'alerte précoce pour les aléas naturels et environnementaux : virage ou mirage technologique ? Revue Des Sciences De L'eau/journal of Water Science. https://doi.org/10.7202/1042922ar

Flora ML, Skinner PS, Potvin CK, Reinhart AE, Jones TA, Yussouf N, Knopfmeier KH (2019) Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental warn-on-forecast system. Weather Forecast 34(6):1721–1739

Fundel VJ, Fleischhut N, Herzog SM, Göber M, Hagedorn R (2019) Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users. Q J R Meteorol Soc 145(S1):210–231. https://doi.org/10.1002/qj.3482

Gagne DJ, McGovern A, Haupt SE, Sobash RA, Williams JK, Xue M (2017) Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. Weather Forecast 32(5):1819–1840

Gagne DJ, Haupt SE, Nychka DW, Thompson G (2019) Interpretable deep learning for spatial analysis of severe hailstorms. Mon Weather Rev 147(8):2827–2845

Giebel G, Shaw W, Frank H, Draxl C, Zack J, Pinson P, Möhrlen C, Kariniotakis G, Bessa R (2021) IEA wind task 36—international collaboration on forecast improvements, EGU general assembly 2021, online, 19–30 Apr 2021, EGU21-13417. https://doi.org/10.5194/egusphere-egu21-13417

Gillet-Chaulet B (2020) Expected utility, a benefit for the forecaster. In: The European Forecaster. Newsletter of theWGCEF N° 25, pp 39–41. Accessed10 Nov 2021

Guan H, Zhu Y (2017) Development of verification methodology for extreme weather forecasts. Weather Forecast 32(2):479–491

Hallegatte S (2012) A cost effective solution to reduce disaster losses in developing countries: hydro-meteorological services, early warning and evacuation. World Bank, Washington, DC

Howard SP, Klockow-McClain KE, Boehmer AP, Simmons KM (2021) Firm behavior in the face of severe weather: economic analysis between probabilistic and deterministic warnings. Weather Forecast 36(3):757–767

Howard SP, Boehmer AP, Simmons KM, Klockow-McClain KE (2022) Business behavior in the face of severe weather: studying the effects of deterministic and probabilistic warning systems. Weather Clim Soc 14(1):39–50

Joslyn SL, LeClerc JE (2012) Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. J Exp Psychol Appl 18(1):126–140

LeClerc J, Joslyn S (2015) The cry wolf effect and weather-related decision making. Risk Anal 35(3):385–395

Lee KK, Kim IG (2020) Social media data analytics to enhance sustainable communications between public users and providers in the weather forecast service industry. Sustainability 12(20):8528

Lim JR, Liu BF, Egnoto M (2019) Cry wolf effect? Evaluating the impact of false alarms on public responses to tornado alerts in the southeastern United States. Weather Clim Soc 11(3):549–563

Lopez A, de Perez EC, Bazo J, Suarez P, van den Hurk B, van Aalst M (2020) Bridging forecast verification and humanitarian decisions: a valuation approach for setting up action-oriented early warnings. Weather Clim Extrem 27(100167):2020

Miran SM, Ling C, Gerard A, Rothfusz L (2018) The effect of providing probabilistic information about a tornado threat on people's protective actions. Nat Hazards 94:743–758

OASIS (2010) Common alerting protocol version 1.2, OASIS standard. http://docs.oasis-open.org/emergency/cap/v1.2/CAP-v1.2-os.pdf. Accessed 7 Nov 2021

Roebber PJ (2009) Visualizing multiple measures of forecast quality. Wea Forecast 24:601–608. https://doi.org/10.1175/2008WAF2222159.1

Rogers D, Tsirkunov V (2011) Costs and benefits of early warning systems, Glob. Assess. Rep. United Nations International Strategy for Disaster Reduction, Washington DC. www.preventionweb.net/english/hyogo/gar/2011/en/bgdocs/Rogers_&_Tsirkunov_2011.pdf

Roulston MS, Smith LA (2004) The boy who cried wolf revisited: the impact of false alarm intolerance on cost-loss scenarios. Weather Forecast 19(2):391–397

Samenow J (2013) Is the National Weather Service issuing too many thunderstorm warnings?. In: The Washington Post. https://www.washingtonpost.com/news/capitalweather-gang/wp/2013/08/14/is-the-weather-service-issuing-too-many-thunderstorm-warnings/. Accessed 25 Nov 2021

Sättele M, Bründl M, Straub D (2016) Quantifying the effectiveness of early warning systems for natural hazards. Nat Hazards Earth Syst Sci 16:149–166. https://doi.org/10.5194/nhess-16-149-2016

Schaefer JT (1990) The critical success index as an indicator of forecasting skill. Weather Forecast 5:570–575

Simmons KM, Sutter D (2009) False alarms, tornado warnings, and tornado casualties. Weather Clim Soc 1(1):38–53

Snyder RL, Melo-Abreu JP (2005) Frost protection: fundamentals, practice and economics, FAO Environment and Natural Resources Service Series, No. 10, vol. 1. FAO, Rome, p 240. https://www.fao.org/3/y7223e/y7223e.pdf. Accessed 10 Sept 2021

Stephenson DB, Jolliffe IT, Ferro CAT, Wilson CA, Sharpe M, Mittermaier M, Hewson TD (2010). White paper review on the verification of warnings, met office forecasting research technical report no.546, p 22. https://library.metoffice.gov.uk/Portal/DownloadImageFile.ashx?objectId=465

Sunstein CR (2000) Cognition and cost-benefit analysis. J Legal Stud 29(2):1059–1103

Swets JA, Dawes RM, Monahan J (2000a) Psychological science can improve diagnostic decisions. Psychol Sci Public Interest 1(1):1–26

Swets JA, Dawes RM, Monahan J (2000b) Better decisions through science. Sci Am 2000(283):82–87

Trainor JE, Nagele D, Phillips B, Scott B (2015) Tornadoes, social science, and the false alarm effect. Weather Clim Soc 7:333–352

UN (2015) Sendai framework for disaster risk reduction 2015–2030. In: Adopted at the third UN world conference on disaster risk reduction in Sendai, Japan, on March 18, 2015.

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Academic Press, London

WMO (2020) 2020 state of climate services: risk information and early warning systems. WMO-no. 1252

Young MV (2017) The human element in forecasting—a personal viewpoint. In: Newsletter of the WGCEF N° 22. https://asr.copernicus.org/articles/17/29/2020/. Accessed 10 Nov 2021