Methodological and conceptual challenges in rare and severe event forecast-verification

Philip A. Ebert[1,*] and Peter Milne[1,*]

[1]Division of Law and Philosophy, University of Stirling, UK
[*]These authors contributed equally to this work.

Correspondence: Philip A. Ebert (p.a.ebert@stir.ac.uk); Peter Milne (peter.milne@stir.ac.uk).

Abstract. There are distinctive methodological and conceptual challenges in rare and severe event (RSE) forecast-verification, that is, in the assessment of the *quality* of forecasts of rare but severe natural hazards such as avalanches or tornadoes. While some of these challenges have been discussed since the inception of the discipline in the 1880s, there is no consensus about how to assess RSE forecasts. This article offers a comprehensive and critical overview of the many different measures used to capture the quality of categorical, binary RSE forecasts—forecasts of occurrence and non-occurrence—and argues that of skill scores in the literature there is only one adequate to RSE forecasting. We do so by first focusing on the relationship between accuracy and skill and showing why skill is more important than accuracy in the case of RSE forecast-verification. We then motivate three adequacy constraints for a measure of skill in RSE forecasting. We show that of skill scores in the literature *only* the Peirce Skill Score meets all three constraints. We then ~~show~~ how our theoretical investigation has important practical implications for avalanche forecasting, basing our discussion on a study in avalanche forecast-verification using the nearest neighbour method (Heierli et al., 2004). Lastly, we raise what we call the "scope challenge"; this affects all forms of RSE forecasting and highlights how and why working with the right measure of skill is important not only for local binary RSE forecasts but also for the assessment of different diagnostic tests widely used in avalanche risk management and related operations~~. Finally, our discussion is also of relevance to the thriving research project of designing~~ methods to assess the quality of regional multi-categorical avalanche forecasts.

1 Introduction

In this paper, we draw on insights from the rich history of tornado forecast-verification to locate important theoretical debates that arise within the context of binary rare and severe event (RSE) forecast-verification. Since the inception of this discipline many different measures have been used to assess the quality of an RSE forecast. These measures disagree in their respective evaluations of a given sequence of forecasts; moreover, there is no consensus about which one is the best or the most relevant measure for RSE forecast-verification. The diversity of existing measures not only creates uncertainty when performing RSE forecast-verification but, worse, can lead to the adoption of qualitatively inferior forecasts with major practical consequences.

This article offers a comprehensive and critical overview of the different measures used to assess the quality of an RSE forecast and argues that there really is only one skill score adequate to binary RSE forecast-verification. Using these insights, we then show how our theoretical investigation has important consequences for practice, such as in the case of nearest neighbour avalanche forecasting, in the assessment of more localised slope stability tests, and other forms of avalanche management.

We proceed as follows: first, we show that RSE forecasting faces the so-called *accuracy paradox* (which, although only recently so-named, was pointed out at least as far back as 1884!). In the next section, we present this "paradox", explain why it is specific to RSE forecasting, and argue that its basic lesson—to clearly separate merely successful forecasts from genuinely skillful forecasts—raises the challenge of identifying adequacy constraints on a measure of the skill displayed in forecasting.

In the third section, we identify three adequacy constraints to be met by a measure of *skill* in RSE forecasting and assess a variety of widely used skill measures in forecast-verification in terms of these constraints. Ultimately, we argue that the Peirce Skill Score is the *only* score in the literature that meets all three constraints and should thus be considered *the* skill measure for RSE forecasting (with a proviso to be noted).

To highlight the practical implications of our theoretical investigation, in the fourth section we build on a recent study of nearest neighbour avalanche forecast-verification and explain how our discussion has important practical consequences in choosing the best avalanche forecast model.

In the final section, we highlight a wider conceptual challenge for verification of binary RSE forecasts by considering what we call the "scope-problem". We examine this problem in the context of avalanche forecasting and conclude by highlighting how our results are relevant to different aspects of avalanche operations and management.

2 Accuracy Paradox: setting the stage

2.1 Sgt. Finley's tornado predictions

The discipline of *forecast-verification* sprang into existence in 1884. In July of that year, Sergeant John Park Finley of the U.S. Army Signal Corps published the article 'Tornado Predictions' in the recently founded *American Meteorological Journal* (Finley, 1884). Finley reported remarkable success in his predictions of the occurrence and non-occurrence of tornadoes in the contiguous United States east of the ~~Rockies~~ during the three-month period from March to May of 1884. Consolidating his monthly figures, we summarise Finley's successes in Table1.

		Observed		
		Tornado	**No tornado**	*totals*
Predicted	**Tornado**	28	72	100
	No tornado	23	2,680	2,703
	totals	51	2,752	2,803

Table 1. Finley's consolidated tornado predictions March–May 1884, after (Gilbert, 1884).

From the totals in the bottom row, we find that the base rate—the climatological probability, as it is sometimes called—of tornado occurrence is a little under 2%, i.e. 51 observations of tornadoes as against 2752 observations of non-tornadoes, well below the 5% base rate used to classify *rare and severe* event forecasting (Murphy, 1991). Further, combining the figures of the top left and bottom right entries we obtain the total number of verified predictions (of both occurrence and of non-occurrence) in the three-month period. Out of a total of 2,803 predictions, 2,708 were correct, which is an impressive success rate of over 96%. This figure goes by many names: among the more common, it is known as the *percentage-correct* (when multiplied by 100), the *proportion-correct*, the *hit rate*, or simply *accuracy*. In a table laid out as in Table 2—commonly referred to as a *2x2 contingency table* or *confusion matrix*—it is the proportion $\frac{a+d}{a+b+c+d}$, which gives us the proportion of predictions that are successful or "verified". The questions we focus on are (1) what does this type of accuracy tell us about forecast performance or forecasting skill in the case of rare event forecasting and (2) how best to measure and compare different RSE event forecast performances.

		Observed		
		+	$-$	*totals*
Predicted	+	a	b	$a+b$
	$-$	c	d	$c+d$
	totals	$a+c$	$b+d$	$a+b+c+d$

Table 2. Standard characterisation of a 2x2 contingency table.

2.2 The accuracy paradox: accuracy vs skill

A feature of tornadoes and also of snow avalanches, is that they are rare events, that is $a+c \ll b+d$. As a result, a forecaster will exhibit high accuracy or attain a high proportion correct simply by predicting 'No tornado' or 'No avalanche' all the time. This trivial (but often overlooked) observation is nowadays blessed with the name *the accuracy paradox* (e.g., Bruckhaus, 2007; Thomas and Balakrishnan, 2008; Fernandes, 2010; Brownlee, 2014; Valverde, 2014; Akosa, 2017; Uddin, 2019; Brownlee, 2020; Davis and Maiden, 2021). The point at issue was made robustly in a letter to the editor in the 15^{th} August, 1884 issue of *Science*, a correspondent named only as 'G.' writing, "An ignoramus in tornado studies can predict no tornadoes for a whole season, and obtain an average of fully ninety-five per cent" (G, 1884).

Indeed, as Finley makes more incorrect predictions of tornadoes (72) than correct ones (28) ($b > a$ in Table 2), it was quickly pointed out that he would have done better *by his own lights* if he had uniformly predicted 'No tornado' (Gilbert, 1884)—he would then have caught up with the skill-less ignoramus whose accuracy, all else equal, would have been an even more impressive 98.2% ($\frac{b+d}{a+b+c+d}$ in Table 2). Where the prediction of *rare* events is concerned, this strongly suggests that accuracy, the proportion correct, is *not* an appropriate measure of the *skill* involved. While we argue for this in more detail in the next section, two concerns counting against accuracy can be noted immediately.

First, focusing on accuracy in rare event forecasting often rewards skill-less performances and incentivizes “no-occurrence” predictions. Second, where the prediction of *severe* events is concerned such an incentive is hugely troubling as a failure to predict occurrence is usually far more serious than an unfulfilled prediction of occurrence. As Allan Murphy observes,

> Since it is widely perceived that type 2 errors [failures to predict occurrences, c in Table 2] are more serious than type 1 errors [unfulfilled predictions of occurrence, b in Table 2], forecasts of RSEs generally are characterised by overforecasting. That is, over a set of forecasting occasions, more RSEs are usually forecast to occur than are subsequently observed to occur [i.e, in terms of Table 2, $a+b > a+c$]. (Murphy, 1991)

The accuracy measure is, then, *doubly unsuitable* when it comes to assessing the skill involved in RSE forecasting.

But if not by accuracy, how should we assess the quality of a set of RSE forecasts? Immediately after the publication of Finley’s article, a number of U.S. government employees rose to the challenge. In the next section, we outline the skill measures they introduced, and in doing so motivate three adequacy constraints that skill measures in RSE forecasting ought to meet.

3 What is skill? Three adequacy constraints on skill measures for RSE forecasting

3.1 First adequacy constraint: Better than chance

Gilbert (1884) responded immediately to Finley’s article and in doing so made two lasting contributions to forecast-verification. His thought was straightforward. Anybody making a sequence of forecasts, whether skilled or unskilled, is likely to get some right by chance. How many? In Table 2, there are $a+c$ occurrences of tornadoes in the sequence of $a+b+c+d$ forecasting occasions. The forecaster makes $a+b$ forecasts of occurrence. If these $a+b$ forecasts were made “randomly”—by “random prognostication” as Gilbert called it—we should expect a fraction $\dfrac{a+c}{a+b+c+d}$ of them to be correct. So the “number”, a_r, of predictions of occurrence that we might expect the skill-less forecaster to get right by luck or chance is $\dfrac{a+c}{a+b+c+d} \times (a+b)$, i.e., the number in proportion to the base rate. Likewise, in parallel fashion, we work out the “number”, d_r, of predictions of non-occurrence we might expect the skill-less forecaster to get right by chance, the “number”, b_r, of predictions of occurrence we might expect the skill-less forecaster to get wrong by chance, and the “number”, c_r, of predictions of non-occurrence we might expect the skill-less forecaster to get wrong by chance *keeping fixed the marginal totals* $a+b$*,* $c+d$*,* $a+c$ *and* $b+d$. We find:

$$a_r = \frac{(a+b)(a+c)}{a+b+c+d}, \quad b_r = \frac{(a+b)(b+d)}{a+b+c+d}, \quad c_r = \frac{(a+c)(c+d)}{a+b+c+d}, \quad d_r = \frac{(b+d)(c+d)}{a+b+c+d}.$$

(We put ‘number’ in scare quotes because $\dfrac{(a+c)(a+b)}{a+b+c+d}$ need not take a whole number value. ~~It’s~~ a familiar fact that an expected value need not be a realisable value—no undamaged face on a fair die has three and a half spots on it.)

$a - a_r$ is then the number of successful predictions of occurrence that we credit to the forecaster's skill, $d - d_r$ the number of successful predictions of non-occurrence. As Gilbert noted,

$$a - a_r = d - d_r = \frac{ad - bc}{a + b + c + d}.$$

The forecaster does better than chance if $a > a_r$, equivalently, if $d > d_r$, i.e., if $ad > bc$. (For a rigorous derivation of a_r, see Appendix A.)

It is also the case that

$$b_r - b = c_r - c = \frac{ad - bc}{a + b + c + d}.$$

What do $b_r - b$ and $c_r - c$ represent? When the forecaster does better than chance, they are the improvements over chance, thus *decreases*, in, respectively, the making of Type I and the making of Type II errors.

Given these considerations, we can now substantiate our earlier claim that Finley exhibited genuine skill, in contrast to the ignoramus, in issuing his predictions. While Finley's 28 correct out of 100 predictions of occurrence made may not seem impressive, his score is a fraction over *fifteen* times more than he could have expected to get right by chance, given Table 1's numbers—Finley's performance was *skillful*.

That was Gilbert's first contribution. Although the next step we take is not exactly Gilbert's, the idea behind it is his second lasting contribution. Our forecaster makes $a + b + c + d$ predictions. How many do we credit to ~~her~~ skill? Gilbert's suggestion is $(a - a_r) + b + c + (d - d_r)$, a suggestion in effect taken up by Glenn Brier and R. A. Allen (1951) when they give this general form for a skill score:

$$\frac{actual\ score - score\ attainable\ by\ chance}{total\ number\ of\ forecasts\ - score\ attainable\ by\ chance}.$$

Here, in both numerator and denominator, the score attainable by chance is $a_r + d_r$. So, instead of accuracy's $\frac{\text{successes}}{\text{predictions}}$ as a measure that doesn't take into account skill, we instead take

$$\frac{\text{successes owed to skill}}{\text{predictions credited to skill}}, \quad i.e., \quad \frac{(a - a_r) + (d - d_r)}{(a - a_r) + b + c + (d - d_r)}.$$

This is

$$\frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}.$$

This is a skill score that, in contrast to accuracy, meets our first adequacy constraint as it controls for chance and aims at genuinely skillful predictions. In the forecasting literature, this measure is known as the special case of the *Heidke Skill Score* (Heidke, 1926) applicable to binary categorical forecasting. It was first mentioned by M. H. Doolittle (1888) but he dismissed it as not having any scientific value. (As we explain below, we have some sympathy with Doolitte's judgement.)

We can rewrite the Heidke Skill Score as:

$$\frac{(b_r - b) + (c_r - c)}{b_r + c_r}.$$

The score is then revealed to be the proportional improvement (decrease) over chance in the making of errors of both Type I and Type II (and if one says only this about it, we have no complaint to make).

Now, if we take it that the best a forecaster can do is have all ~~her~~ predictions, both of occurrence and non-occurrence, fulfilled then, following Woodcock (1976), we can present the Heidke Skill Score in an interestingly different way:

$$\frac{actual\ score - score\ attainable\ by\ chance}{best\ possible\ score\ - score\ attainable\ by\ chance}.$$

Here, as said, we equate the "best possible score" with correctly predicting all occurrences and non-occurrences—$a+c$ correct predictions of occurrence, $b+d$ correct predictions of non-occurrence (Table 3). Substituting into the above equation, we obtain

		Observed		
		+	−	*totals*
Predicted	+	$a+c$	0	$a+c$
	−	0	$b+d$	$b+d$
	totals	$a+c$	$b+d$	$a+b+c+d$

Table 3. Best possible score relative to Table 2's data on occurrence.

the Heidke Skill Score, for the actual score, the actual number of successful predictions, is $a+d$, the number attributable to chance is a_r+d_r, and the best possible score is $a+b+c+d$.

Focusing on the notion of a *best possible score*, though, gives us a different way to think about skill. On the model of what we did above, someone randomly making $a+c$ predictions of occurrence could expect to get $\frac{(a+c)(a+c)}{a+b+c+d}$ of them right by chance and, likewise, someone randomly making $b+d$ forecasts of non-occurrence could expect to get $\frac{(b+d)(b+d)}{a+b+c+d}$ of them right by chance. So in the case of perfect prediction, in which there are no Type I or Type II errors, the number of successes we credit to the forecaster's skill is

$$(a+b+c+d) - \frac{(a+c)^2+(b+d)^2}{a+b+c+d} = \frac{2(a+c)(b+d)}{a+b+c+d}.$$

Putting a different reading on our rewriting of Brier and Allen's conception of a skill score, namely,

$$\frac{actual\ score - score\ attainable\ by\ chance\ (relative\ to\ actual\ performance)}{best\ possible\ score - score\ attainable\ by\ chance\ (relative\ to\ best\ possible\ performance)},$$

we get

$$\frac{2(ad-bc)}{2(a+c)(b+d)} = \frac{a}{a+c} - \frac{b}{b+d},$$

which is known as the *Peirce Skill Score* (Peirce, 1884), the *Kuipers Skill Score* (KSS, (Hanssen and Kuipers, 1965)), and the *True Skill Statistic* (TSS, (Flueck, 1987)). (Peirce's own way of arriving at the Peirce Skill Score is quite different~~. It~~ is examined in detail in Milne (submitted).)

We can think of the Peirce Skill Score as being this ratio:

$$\frac{successes\ due\ to\ skill\ in\ actual\ performance}{successes\ due\ to\ skill\ in\ perfect\ performance},$$

and we will discuss this way of thinking of the Peirce Skill Score further in what follows.

To summarise, one of the earliest responses to the challenge to identify the skill involved in RSE forecasting was to highlight the need to take into account, in some way, the possibility of getting predictions right "by chance" and thus present the skill exhibited in a sequence of forecasts as relativized to what a "random prognosticator" could get right by chance. As we have just seen, this can be done in different ways which motivate different measures of skill. At this point, we don't have much to say on whether Gilbert's and Brier and Allen's reading or our rewrite is preferable, i.e. whether the Heidke or Peirce Skill Score is preferable. However, we can note that this first adequacy constraint rules out simple scores such as accuracy (proportion correct) as capturing anything worth calling *skill* in forecasting.

We should note that some measures in the literature, in particular those that are functions of a, b and c only, such as Gilbert (1884)'s $\frac{a}{a+b+c}$, the Dice coefficient $\frac{2a}{2a+b+c}$, also called the F-score or F_1, and its generalisation, the adjusted F-measure, $F_\beta = \frac{(1+\beta^2)a}{(1+\beta^2)a+b+\beta^2 c}$, and $\frac{a}{\sqrt{(a+b)(a+c)}}$, sometimes called the Fowlkes–Mallows index or the cosine similarity, do not build in a correction for chance successes. The measures just mentioned are expressible as functions of $\frac{a}{a+b}$ (the frequency of hits, FOH) and $\frac{a}{a+c}$ (the probability of detection, POD, or hit rate), as Gilbert (1884) initially proposed any measure of forecasting quality should be; he tacitly gave up on that commitment when he hit upon the idea of discounting the number of successful predictions attainable by chance. Now, we can evaluate these measures setting a, b and c to the values they take for "random prognostication", i.e., a_r, b_r and c_r (which are all functions of a, b, c *and* d). When we do this we find that each of the measures returns a value greater than it yields for random prognostication when, and only when, $ad > bc$. The argument is straightforward: except at the extremes, i.e., except when $a > 0$ and $b = c = 0$, when each of these measures takes the value 1, or when $a = 0$ and $b + c > 0$ (or $b \times c > 0$ in the case of the Fowlkes–Mallow coefficient), when each of the measures takes the value 0, each of the measures is strictly increasing in a and strictly decreasing in b and in c; as we saw above, $a > a_r$ if, and only if, $b_r > b$ if, and only if, $c_r > c$ if, and only, if $ad > bc$, hence each counts a performance as better than chance ("random prognostication") if, and only if, $ad > bc$. All well and good but a significant weakness of these measures looms exactly here, namely, that what value each assigns random prognostication is highly context dependent—dependent on the values of a, b, c *and* d so that the same value of the measure may on one occasion indicate a better than chance performance, on another a worse than chance performance. With these measures, then, whether a performance is better than chance and to what extent has to be worked out on a case-by-case basis, a significant demerit.

Satisfaction, or not, of the criterion that $ad > bc$, depends, of course, on there being a value for d. Jolliffe (2016) picks up on this when, while acknowledging that the Dice and Gilbert measures just considered 'have undesirable properties', he says that 'the Dice coefficient is of use when the number of "correct rejections", d, is unknown, is difficult to define or is so large that it dominates the calculation of most measures.' The first two render assessment of whether a performance is better than chance at best doubtful, at worst impossible; the third renders it trivial (assuming $a \neq 0$). We take the implication of Jolliffe's remarks to be that measures such as Gilbert's original—the Jaccard coefficient as it is often called—, Dice's and the cosine similarity are *only* of use when, for whatever reason, the number of successful predictions of non-occurrence is either not well defined or not known which however is usually not the case for RSE forecasting.

3.2 Second adequacy constraint: Direction of fit

~~M. H.~~ Doolittle (1885a, b) introduced a measure of "that part of the success in prediction which is due to skill and not to chance" that is the product of two measures now each better known in the forecasting literature than Doolittle's own, the Peirce Skill Score, which we have just introduced, expressed in terms of Table 2 as $\frac{a}{a+c} - \frac{b}{b+d}$, and the Clayton Skill Score (Clayton, 1927, 1934, 1941), expressed in the same terms as $\frac{a}{a+b} - \frac{c}{c+d}$. Even before publication, Doolittle was criticised by Henry Farquhar (1884): Doolittle had, said Farquhar, combined a measure that tests occurrences for successful prediction (the Peirce Skill Score), with a measure that tests predictions for fulfilment (the Clayton Skill Score). Now, Doolittle's measure is indeed the product of the indicated measures and Farquhar is correct in his claim regarding what those measures measure. But why is this ground for *complaint*? Doolittle saw none. Apparently taking on board Farquhar's observation, he says,

> Prof. C. S. Peirce (in *Science*. Nov. 14, 1884, Vol. IV., page 453), deduces the value
>
> $$i = \frac{a(a+b+c+d)-(a+c)(a+b)}{(a+c)(b+d)} \quad \left[= \frac{a}{a+c} - \frac{b}{b+d}\right],$$
>
> by a method which refers principally to the proportion of occurrences predicted, and attaches very little importance to the proportion of predictions fulfilled. (Doolittle, 1885b, p. 328, with a change of notation)

Farquhar allows that 'either of these differences [i.e., the Peirce Skill Score and the Clayton Skill Score] may be taken alone, with perfect propriety.' By multiplying the Peirce Skill Score and the Clayton Skill Score, one is multiplying a measure that tests occurrences for successful prediction by a measure that tests predictions for fulfilment. The resulting quantity is neither of these things—but that, in itself, does not formally prevent it being, as Doolittle took it to be, a measure of the skill exhibited in prediction. Why, then, should one not multiply them, or put differently: what is *wrong* with Doolittle's measure?

The answer, we suggest, lies in a notion known to philosophers as *direction of fit*. The idea, but not the term, is usually credited to Elizabeth Anscombe who introduced it thus:

> Let us consider a man going round a town with a shopping list in his hand. Now it is clear that the relation of this list to the things he actually buys is one and the same whether his wife gave him the list or it is his own list; and that there is a different relation where a list is made by a detective following him about. If he made the list itself, it was an expression of intention; if his wife gave it him, it has the role of an order. What then is the identical relation to what happens, in the order and the intention, which is not shared by the record? It is precisely this: if the list and the things that the man actually buys do not agree, and if this and this alone constitutes a mistake, then the mistake is not in the list but in the man's performance (if his wife were to say: "Look, it says butter and you have bought margarine", he would hardly reply: "What a mistake! we must put that right" and alter the word on the list

> to “margarine”); whereas if the detective’s record and what the man actually buys do not agree, then the mistake is in the record. (Anscombe, 1963, §32)

As Anscombe’s observation regarding butter and margarine makes clear, the ideal performance for the husband is to have the contents of his shopping basket match his shopping list; the ideal performance for the detective is for his list to match the contents of the shopping basket. The difference lies in whether list or basket sets the standard against which the other is evaluated—this is the difference in *direction of fit*. Put bluntly, Peirce has the sequence of weather events set the standard and evaluates sequences of predictions against that standard; Clayton has the sequence of actual predictions set the standard and evaluates sequences of weather events against that standard. This difference in direction of fit is beautifully pointed out *by Doolittle himself*. Contrasting Peirce’s measure with the other component of his own, the Clayton Skill Score as we now know it, he says,

> Prof. C. S. Peirce (in *Science*, Nov. 14, 1884, Vol. IV., page 453), deduces the first of these factors as the unqualified value of i [the *inference-ratio* or that part of the success which is due to skill and not to chance] .... He obtains his result by the aid of the supposition that part of the predictions are made by an infallible prophet, and the others by a man ignorant of the future. If Prof. Peirce had called on omnipotence instead of omniscience, and supposed the predictions to have been obtained from a Djinn careful to fulfill a portion of them corresponding to the data, the remainder of the occurrences being produced by an unknown Djinn at random, he would have obtained by parallel reasoning the second factor. (Doolittle, 1885a, p. 124)

When measuring occurrences for successful prediction, the aim is to *match predictions to the world*, something which an *omniscient* being succeeds in doing; in measuring predictions for fulfilment, the ideal is *to have the world match the predictions made*, something which an *omnipotent* being can arrange to be the case.

In considering improvements on the forecasting performance recorded in Table 2, what are kept fixed are the numbers of actual occurrences and non-occurrences, the marginal totals $a+c$ and $b+d$, not the numbers of actual predictions of occurrence and predictions of non-occurrence, the marginal totals $a+b$ and $c+d$. It is, after all, only a poor joke to say, ‘I would have had a higher skill score if more tornadoes had occurred,’ even though it may well be true. Doolittle has, despite himself, made clear for us that forecasters are like Anscombe’s detective and not like the husband with the shopping list. Forecasters aim to fit their predictions to the world, not the world to their predictions.

~~Let’s~~ go back to this form for a skill score:

$$\frac{actual\ score - score\ attainable\ by\ chance\ (relative\ to\ actual\ performance)}{best\ possible\ score - score\ attainable\ by\ chance\ (relative\ to\ best\ possible\ performance)}.$$

Peirce’s conception of the best possible performance, presented above in Table 3, keeps the marginal totals for actual observed occurrences and non-occurrences from Table 2, $a+c$ and $b+d$, respectively. The actual numbers of occurrence and non-occurrence set the standard against which performances are measured; so-constrained, the best possible performance is that of the as-it-were omniscient being who correctly predicts all occurrences and all non-occurrences (Table 3).

Clayton’s conception of the best possible performance, presented below in Table 4, keeps the marginal totals for actual predictions of occurrence and predictions of non-occurrence from Table 2, $a+b$ and $c+d$, respectively. The actual numbers of predictions of occurrence and predictions of non-occurrence provide the standard against which performances are measured; so-constrained, the best possible performance is that of the omnipotent being who fashions occurrences and non-occurrences to fit ~~her~~ predictions (Table 4). And so, returning to our original question, it should now be clear what is wrong with Doolittle's measure: it incorporates Clayton's measure which has the wrong direction of fit for a measure of skill in prediction.

		Observed		
		+	$-$	*totals*
Predicted	+	$a+b$	0	$a+b$
	-	0	$c+d$	$c+d$
	totals	$a+b$	$c+d$	$a+b+c+d$

Table 4. Omnipotent forecaster's score relative to Table 2's data on prediction.

What of the Heidke Skill Score? How does it fare with respect to direction of fit? What conception of best performance does it employ? In its denominator, the Heidke score takes the best possible performance to be one in which all $a+b+c+d$ predictions are correct but corrects that number for chance using Table 2's marginal totals for both predictions *and* occurrences. This is, quite simply, *incoherent*—unless, fortuitously, we are in the special case when the numbers of Type I and Type II errors are equal. Keeping Table 2's marginal totals, the highest attainable number of correct predictions is $a+d+2\times\min\{b,c\}$ (Table 5), *not* $a+b+c+d$.

		Observed		
		+	$-$	*totals*
Predicted	+	$a+\min\{b,c\}$	$b-\min\{b,c\}$	$a+b$
	-	$c-\min\{b,c\}$	$d+\min\{b,c\}$	$c+d$
	totals	$a+c$	$b+d$	$a+b+c+d$

Table 5. Highest number of correct predictions relative to Table 2's marginal totals

Using the marginal totals in Table 5, which are, by design, those of Table 2, to correct $a+d+2\times\min\{b,c\}$ for chance, we obtain this skill score:

$$\frac{ad-bc}{(a+\min\{b,c\})(\min\{b,c\}+d)}.$$

This measure, which had previously occurred in other literature (Benini, 1901; Forbes, 1925; Johnson, 1945; Cole, 1949), has been used to assess forecasting performances not in tornado forecasting nor in avalanche forecasting but in assessing predictions of juvenile delinquency and the like in criminology where it is known as *RIOC*, Relative Improvement Over Chance (Loeber and Dishion, 1983; Loeber and Stouthamer-Loeber, 1986; Farrington, 1987; Farrington and Loeber, 1989; Copas and Loeber, 1990).

The *RIOC* measure has the following feature: when there are successes in predicting occurrences and non-occurrence, i.e., $a > 0$ and $d > 0$, it awards a maximum score of 1 to any forecasting performance in which there are *either* no Type I errors ($b = 0$) *or* no Type II errors ($c = 0$) or both. This is a feature it shares with Stephenson (2000)'s *Odds Ratio Skill Score* (ORSS) (for which see Appendix D). In agreement with Woodcock (1976), we hold that a maximal score should be attained when, *and only when*, b and c are *both* zero.

That~~'s~~ one problem with the *RIOC* measure. The other is this. Like Anscombe's detective, the scientific forecaster's aim is to match ~~her~~ predictions to what actually happens. This is why we keep the column totals fixed when considering the best possible performance. Why on earth should we also keep the row totals, the numbers of predictions of occurrence and non-occurrence fixed? There is, we submit, no good reason to do so. The Heidke Skill Score embodies no coherent conception of best possible performance. Loeber *et al.*'s *RIOC* does at least embody a coherent notion of best possible performance but it is a needlessly hamstrung one, restricting the range of possible performances to those that make the same number of predictions of occurrence and of non-occurrence as the actual performance. On the one hand, this makes a "best possible performance" too easy to achieve and, on the other, sets our sights so low as to only compare a forecaster with others who make the same *number* of forecasts of occurrence and of non-occurrence—but forecasting is a scientific activity, not a handicap sport.

Finally, for completeness, ~~let's~~ consider the measure we started out with, proportion correct, $\dfrac{a+d}{a+b+c+d}$. How does it fare with respect to the second adequacy constraint? While it may be true to say that it doesn't evaluate a performance in relation to the *wrong* direction of fit, this is the case only because the measure doesn't properly engage with the issue of fit. Here, the evaluation is in relation to $a+b+c+d$ and so the performance is not evaluated in relation to any relevant proportion (neither of occurrences nor of predictions). A similar consideration applies to Gilbert's original measure $\dfrac{a}{a+b+c}$. More generally, though, the measures we considered at the end of the previous section, Gilbert's original, F_1/the Dice coefficient and the adjusted F-measures, the F_β family, and the Fowlkes–Mallows index/cosine similarity are, as we said, functions of $\dfrac{a}{a+b}$ and $\dfrac{a}{a+c}$. Uncorrected for chance though they be, these are, as Farquhar put it, a measure that tests predictions for fulfilment and a measure that tests occurrences for successful prediction, respectively. Our complaint against the measures is, then, the same as our complaint against Doolittle's measure: each incorporates a component which has the wrong direction of fit for a measure of skill in prediction.

The notion of direction of fit has much wider application than just forecasting: it applies in any setting in which we can see "the world" or a "gold standard test" or, more prosaically, some aspect of the set-up in question as setting the standard against which a "performance" is judged. Diagnostic testing is one obvious case—and there we have the Peirce Skill Score but under the name of the Youden Index (Youden's J) (Youden, 1950); in medical/epidemiological terms, it is commonly expressed as $sensitivity + specificity - 1$. More widely used to summarise diagnostic accuracy is the (positive) likelihood ratio $\dfrac{sensitivity}{1-specificity}$ (Deeks and Altman, 2004; Šimundić, 2009/2012b) but this is a function of the same proportions so fares equally well in terms of direction of fit. By way of contrast, in information retrieval, following the lead of early work by M.

Lesk and G. Salton (1969), M. H. Heine (1973), and N. Jardine and C. J. van Rijsbergen (1971; see also van Rijsbergen, 1974) it has become common practice to measure the effectiveness of a retrieval system in terms of precision, i.e., proportion of retrieved items that are relevant, and recall, i.e., the proportion of relevant items retrieved, committing exactly the direction-of-fit error we have diagnosed in the applications of the Doolittle, Gilbert, Dice, F_β, and cosine similarity measures with 'relevant' playing the role of 'observed' and 'retrieved' that of 'predicted' in Table 2. (It should be noted that the analogue of the Peirce Skill Score, known as $recall - fallout$, has had its advocates in information retrieval (see Goffman and Newill, 1966; Robertson, 1969a, b) but in practice the F-measures, and F_1 in particular, have come to dominate, so much so that 'it is now virtually impossible to publish work in Information Retrieval or Natural Language Processing without including it [F_1]' (Powers, 2015/2019).)

So, in summary, we can say that *in any context in which direction of fit plays a role* the Peirce score, alone of scores in the literature, evaluates performances in relation to the correct proportions (occurrences and non-occurrences predicted).

3.3 Third adequacy constraint: Weighting errors

We think there is a third feature of skill that is specific to *severe* event forecasting that a skill measure ought to take into account. Broadly speaking, it consists in being sensitive, in the right kind of way, to one's own fallibility. While the omniscient forecaster need not worry about mistakes, actual forecasters need to be aware of the different kinds of consequences of an imperfect forecast. To motivate our third constraint, consider the two forecasts in Table 6. While forecasts A and B issue the same total number of forecasts and both score an excellent 98.8% accuracy, they disagree on the *kinds* of errors they make. Forecast A makes fewer Type II errors (1) than Type I errors (5), while in forecast B this error distribution is reversed. Is there a reason to think that one forecast is *more skillful* than the other?

		Forecast A		Forecast B	
		Observed		Observed	
		+	−	+	−
Predicted	+	5	5	5	1
	−	1	500	5	500

Table 6. Example of two forecasts (A, B) that agree on the correct predictions and the total number of false predictions, but differ in the *kinds* of false predictions (Type I vs Type II).

Given the context of our discussion, i.e. rare *and severe* event forecasting, we believe there is. We saw Allan Murphy saying that "it is widely perceived that type 2 errors [erroneous predictions of non-occurrence] are more serious than type 1 errors [unfulfilled predictions of occurrence]". A skillful RSE forecaster should take this observation into account and consider, as it were, the *effects of their mistakes*. As a result a skill measure should incorporate—in a principled way—the different effects of Type I and Type II errors and judge forecast A as *more skillful* than forecast B, at least when the forecast is evaluated in the context of RSE forecasting.

Importantly, the Peirce Skill Score does just that. We can re-write it as

$$1-\frac{c}{a+c}-\frac{b}{b+d},$$

and read it as making a deduction from 1, the score for a perfect omniscient performance, for each Type II and each Type I error, respectively. Now, when we are concerned with rare events, i.e., when $a+c \ll b+d$, the "deduction per unit" is greater for Type II errors than for Type I errors. As a result, it is built into the Peirce Skill Score, in a principled way, that Type II errors count for more than Type I errors when we are dealing with rare events. This is borne out in the Peirce Skill Score for our two forecasts above: forecast A receives a score of .823 while forecast B receives a score of .498. Note that this *feature* of the Peirce score would turn into a *liability* if we were to consider very common but nevertheless severe events. (Put contrapositively, the Peirce Skill Score increases the *relative* score of correct predictions of occurrence as the base rate of occurrences decreases (Gandin and Murphy, 1992; Manzato, 2005).)

Now, when d is large, as it often is in the case of rare events forecasts, it is likely to be the case that $a+b \ll c+d$. When this is the case the Clayton Skill Score, which we may write as

$$1-\frac{b}{a+b}-\frac{c}{c+d},$$

turns the good behaviour of the Peirce Skill Score on its head, giving a greater "deduction per unit" for Type I errors than for Type II errors. According to the Clayton Skill Score, we should regard forecast B (.823) as more skillful than forecast A (.498). So, not only does the Clayton Skill Score fail to meet the direction of fit requirement, it also fails—in quite a spectacular way—our third requirement of weighting errors.

Formally, the Heidke Skill Score treats Type I and Type II errors equally in that interchanging b and c, i.e., Type I and Type II errors, in

$$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$$

leaves the measure unchanged. To wit, according to the Heidke Skill Score, forecasts A and B are equally skillful with a measure of .619.

In fact *all* measures known to us in the forecasting literature other than the Peirce Skill Score, the Clayton Skill Score and the F_β measure (which places more weight on minimizing Type II errors when $\beta > 1$) behave like the Heidke Skill Score on this issue: interchanging b and c leaves the measure unchanged and Type I and Type II errors are on a par (see Appendix D for a list of other skill scores for each of which this claim can easily be confirmed). Nonetheless, we can say in favour of the Heidke score that it provides the right incentive: in an application in which $b+d > a+c$, as it is in the case of rare event forecasting, an increase in Type II errors would lower the actually attained score by more than the same number of Type I errors and a decrease in Type II errors would increase the actually attained score by more than the same number fewer Type I errors (see Appendix B for the formal details).

	Better than chance	Direction of Fit	Weighting errors
Accuracy Score	No	no preferred direction	no weights
Heidke Skill Score	Yes	incoherent	no weights
RIOC	Yes	no preferred direction	no weights
Peirce Skill Score	**Yes**	**correct direction**	**correct weighting**
Clayton Skill Score	Yes	incorrect direction	incorrect weighting
Dice coefficient (F_1-score)	Not built in	correct and incorrect direction combined	no weights
F_β-measure	Not built in	correct and incorrect direction combined	adjustable weighting

Table 7. Summary comparison of skill measures in relation to the three adequacy constraints for RSE forecasting.

Table 7 collates our main claims and presents each of the measures discussed so far in relation to the three adequacy constraints. It is worth noting that while the first requirement arises in particular in the case of *rare* event forecasting, both the first and second are, arguably, relevant also in common event forecasting. The third constraint, however, is distinctive of *severe* event forecasting.

Finally, should we consider these three constraints as jointly sufficient? Of course, further debate may generate other adequacy constraints on skill measures, and we are open to such a development at this stage of the discussion. However, we take ourselves to have shown that there really is only one skill measure in the forecasting literature that meets the three constraints and so there is only one genuine *candidate* for a measure of skill in RSE forecasting.

4 Application: the relevance of skill scores in avalanche forecast-verification

In this section, we will show how our theoretical discussion concerning skill measures has consequences for the practice of avalanche forecast-verification. We focus on the use of the "nearest neighbour" (NN) method of avalanche forecasting as discussed in Heierli et al. (2004). The idea of NN forecasting for avalanches dates back to the 1980's (Buser, 1983; Buser et.al., 1987; Buser, 1989) and has been widely used for avalanche forecasting in Canada, Switzerland, Scotland, India, and the US (e.g. Brabec and Meister, 2001; Gassner et.al., 2001; Gassner and Brabec, 2002; Purves et.al., 2003; Heierli et al., 2004; Roeger et.al, 2004; Singh and Ganju, 2004; Singh et.al., 2005; Purves and Heierli, 2006; Singh et.al., 2015). ~~In order~~ to evaluate the quality of this forecasting techniques, forecast-verification is an indispensable tool and this process can take on different forms (Purves and Heierli, 2006). In the case of binary NN forecasts, a variety of measures have been used in the verification process, though there is currently no consensus about how to adjudicate between them (Heierli et al., 2004). Most studies simply present a list of different measures without providing principled reasons as to which measure is the most relevant one (an exception is Singh et.al. (2015) who opt for the Heidke score). This section offers a discussion as to how the many different measures should be used and ranked in their relevance for avalanche forecast-verification in the context of binary NN forecasting. ~~It's~~ worth noting, however, that broadly similar considerations will be applicable to the verification used in other avalanche forecasting techniques, or indeed to other kinds of binary RSE forecasts and their verification, such as landslides (e.g. Leonarduzzi and Molnar, 2020; Hirschberg et al., 2021).

The basic assumption of the NN forecasting approach is that similar ~~initial~~ conditions ~~with respect to external conditions~~, such as the ~~snow pack~~, temperature, weather, etc., will likely lead to similar outcomes and so historical data—weighted by relevance and ordered by similarity—is used to inform forecasting. More specifically, NN forecasting is a non-parametric pattern classification technique where data is arranged in a multi-dimensional space and a distance measure (usually the Brier score) is used to identify the most similar neighbours. NN forecasting can be used for categorical or probabilistic forecasts. In the case of the former, which is relevant to our current discussion, a decision boundary k is set and an avalanche is forecast, i.e. a positive prediction is issued, when the number of positive neighbours (i.e. nearest neighbours on which an avalanche was recorded) is greater than or equal to that decision boundary k.

Heierli et al.'s study on avalanche forecast-verification uses two data sets, one focused on Switzerland and the other on Scotland. Figure 1 summarises their results and ~~it~~ nicely shows how changes in the decision boundary k affect a variety of measures, such as accuracy and other skill measures. In what follows, we will follow their lead and investigate their finding through a more "methodological" lens. Using their study will help us explain differences in behaviour of the skill measures given variations in the decision boundary, and highlight how our discussion has practical consequences. As Heierli et al. emphasise, one core issue for NN forecasting is which decision-boundary k should be chosen, i.e. for which k do we get the "best" forecast. Naturally, this choice should depend, crucially, on how we assess the goodness of the different forecasts given variations in k. Our proposal is that the choice of k should be settled by establishing which value of k issues in the most skillful forecast. (Manipulating the value of k in NN forecasting is a special case of using different values of a parameter to categorise items—future events, say—in two classes. One can plot $\frac{a}{a+c}$, which, in the case of forecasting, is the probability of detection, against $\frac{b}{b+d}$, *the probability of false detection*, as it is sometimes called, resulting from different values of the parameter to yield a receiver operating characteristic (ROC) curve (Swets, 1973; Metz, 1978; Altman and Bland, 1994; Šimundić, 2009/2012a; Manzato, 2005, 2007). The Peirce Skill Score at a point on the curve, $\frac{a}{a+c} - \frac{b}{b+d}$, is a measure of the distance of the point from the line $x = y$ which represents a forecast of zero value. It seems built into this approach that the Peirce Skill Score is a measure of forecast quality.)

~~Let's~~ start our discussion by noting two immediate consequences of NN avalanche forecasting. Remember that a positive prediction is issued when the number of positive neighbours is equal or greater than k. From this follows that:

(i) the number of positive predictions $(a+b)$ is greater the lower k.

(ii) the number of *correct* predictions made, a, is greater, the lower k.

Given that $a+c$ and $b+d$ are fixed, and given (i) and (ii), we can also note that $\frac{a}{a+c}$, i.e. the *probability of detection* (*POD*) varies inversely with k, as is evident in the graphs in Heierli et al.'s Figure 1. With these observations in place, ~~let's~~ look at how the measures we discussed earlier fare with respect to variations in k.

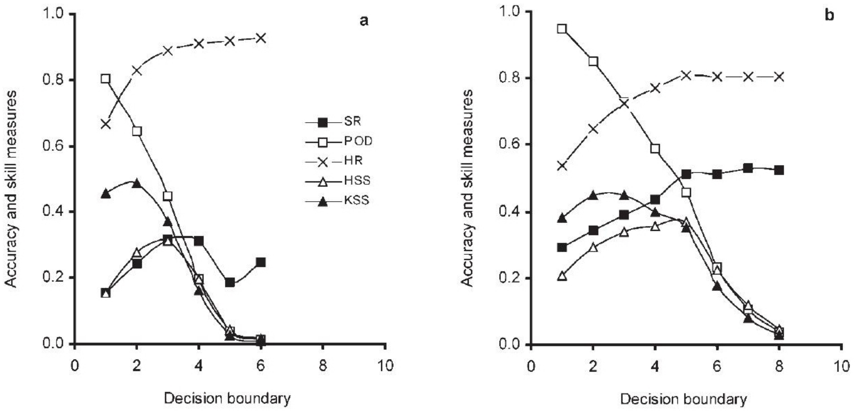

Figure 1. Dependence of accuracy and skill measures on the choice of decision boundary (number of positive neighbours of the forecast day). (a) Swiss dataset; (b) Scottish dataset. From (Heierli et al., 2004) and reprinted from the *Annals of Glaciology* with permission of the International Glaciological Society and the lead author.

4.1 Accuracy measure: its shortcomings exemplified

Our earlier discussion about the disadvantages of using accuracy as a measure are nicely borne out in Heierli et al.'s study, which renders the measure unsuitable as the main criterion when deciding on the decision-boundary k.

Accuracy or proportion correct, $\frac{a+d}{a+b+c+d}$ is what Heierli et al. call the *hit rate*, HR in Figure 1. From Heierli et al. (2004)'s graphs, we see that it increases as k increases in both datasets but also tends to level off: at 90$\%$ and over for $k \geq 4$ in the Swiss dataset, at about 80$\%$ for $k \geq 5$ in the Scottish dataset.

Given (ii), we know that a decreases as k increases, so this improvement in accuracy is entirely due to an increase in the number of correct negative predictions d achieved at the expense of a drop in the number of correct positive predictions. Of course that drop in the number of correct positive predictions goes hand-in-hand with a proportionally greater drop in the number of mistaken positive predictions (Type I errors). But that is accompanied by an increase in Type II errors, mistaken negative predictions, i.e. avalanches that were not predicted.

In short, as k increases more Type II errors are committed than Type I errors. However this "trading off" of errors is, as we discussed in section 3.3, a seriously bad trade in the context of RSE forecasting. Now, maybe to some extent the absolute numbers should matter here, but generally in the context of RSE forecasting, we do want to minimise Type II errors and have Type II errors weigh more than Type I errors. As we showed earlier, the accuracy measure fails to do that.

Moreover, and as to be anticipated given our discussion in section 2.2, if really all we want to achieve is to improve accuracy then we also have to consider the "ignoramus in avalanche studies" who uniformly makes negative predictions, i.e., uniformly forecasts non-occurrence. They have an accuracy score of $\frac{b+d}{a+b+c+d}$. This is exceeded by the accuracy score of the skilled employer of the nearest-neighbour method only when $a > b$, i.e., just when the success rate (SR) $\frac{a}{a+b} > 0.5$. But as we can see, in the Swiss dataset SR never gets above 0.3 and in the Scottish dataset it rises to about 0.5 and more or less plateaus.

Hence, if all that mattered was accuracy—and, to be sure, this is not what Heierli et al. are suggesting—then the lessons from this study for forecasting in Switzerland is to set the decision-boundary k to ∞, making it impossible to issue any positive predictions and in doing so increase accuracy. Hence accuracy really ~~isn't~~ a good measure to assess a professional avalanche forecaster's performance. We hope they agree not merely due to concerns about job security.

To be clear, these considerations do not imply that there's *no* role for accuracy. Accuracy is not an end in itself, that much we take as established. Nevertheless, we think accuracy may well play a secondary role in "forecast-choice": if two sets of predictions are graded equally with respect to genuine *skill*, we should prefer or rate more highly the one which has the greater accuracy. After all, it is making a greater proportion of correct predictions. So a view we are inclined to adopt is one where *all things considered*, accuracy can be a tie-breaker between sets of predictions that exhibit the same degree of skill according to the Peirce skill measure. Technically, our view amounts to a *lexicographic* all-things-considered ordering for forecast-verification: first rank by skill using the Peirce score, next rank performances that match in skill by accuracy. ~~Let's~~ next have a look at the behaviour of our favourite skill score.

4.2 The Peirce Skill Score and NN avalanche forecasting

The Peirce Skill Score is also known as the Kuipers Skill Score, *KSS*, the name used by Heierli et al.. Notice that it initially increases as k increases but then falls away, quite dramatically so, as k increases. The fall-off starts when k exceeds 2 and is immediately dramatic in Heierli et al.'s Swiss dataset; it starts when k exceeds 3 and is initially quite gentle in their Scottish dataset. We think that the most skilled forecasts are issued when the decision-boundary is set at 2 (Switzerland) and 3 (Scotland).

~~Let's~~ investigate a little further the behaviour of KSS. As said, $a+c$ and $b+d$ are fixed, hence the base rate BR is fixed. As k increases, a and b both decrease (or, strictly speaking, at least fail to increase but in practice decrease). Obviously, as a decreases, $\dfrac{a}{a+c}$ decreases; but as b decreases, $-\dfrac{b}{b+d}$ *increases*. $\dfrac{b}{b+d}$ is sometimes called the *false alarm rate* and sometimes the *probability of false detection*, i.e. PFD. Now, why does KSS so dramatically decrease? The answer should be clear given our discussion of how Type I and II errors are weighted: as k increases, a and b both decrease and c and d both increase. Given that $a+c$ and $b+d$ are fixed the number of Type II errors increases when k increases. As discussed in section 3.3, the KSS score penalises Type II errors more heavily than Type I errors when $a+c \ll b+d$. Hence a decrease in the latter is unable to outweigh the increase in the former. In addition, given that the KSS measure penalises Type II errors more heavily the rarer the to-be-forecasted event, the lower base rate in the Swiss data set—7% compared to 20% in the Scottish data set—explains the more dramatic fall in the KSS value in the Swiss data set compared to the Scottish one.

4.3 The Heidke Skill Score and NN forecasting

We previously noted our reservations concerning the Heidke Score; it is, however, an often used skill score in forecast-verification (compare Singh et.al. (2015) who uses it in their evaluation of nearest neighbour models for operational avalanche forecasts in India). Interestingly, the Heidke score arrives at a different choice of k for the two data sets, yet the behaviour of the Heidke Skill Score, *HSS* (Heidke, 1926; Doolittle, 1888),

$$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)},$$

is broadly similar to that of *KSS* in that it initially rises and then falls off. For the Swiss data set, *HSS* provides the highest skill rating for a decision boundary $k = 3$ and for the Scottish data $k = 5$. So, it really does matter which skill measure we choose when making NN forecast evaluations with important practical consequences. Why do we get such different assessments of the forecast performances?

In both graphs, $KSS > HSS$ for low values of k but not for larger values of k. This is intriguing. When the forecasting performance is better than chance, i.e., when $ad > bc$ in Table 2, and occurrence of the positive event is rarer than its non-occurrence, $a+c < b+d$, KSS exceeds HSS if, and only if, Type I errors, mistaken predictions of occurrence of the positive event, exceed Type II errors, failure to predict occurrence of the positive event—see Appendix C. In other words, in the stated circumstances, KSS exceeds HSS if, and only if, $b > c$, hence $a+b > a+c$ and there is "over-forecasting" which, as noted earlier, is penalised less heavily in the case of KSS than "under-forecasting". Now, we quoted Murphy earlier noting that given the seriousness of Type II errors overforecasting is, as it were, a general feature of RSE forecasting. However, Murphy goes on to say,

> The amount of overforecasting associated with forecasts of some RSEs is quite substantial, and efforts to reduce this overforecasting—as well as attempts to prescribe an appropriate or acceptable amount of overforecasting—have received considerable attention. (Murphy, 1991, p. 304)

Now, how "bad" too much overforecasting is and when it is too much is a separate issue that depends on the kind of event that is to be forecast and may also depend on the behavioural effects overforecasting has on individual decision and the public's *trust* in forecasting agencies. But this much is clear: we have to acknowledge that KSS encourages more overforecasting when compared to HSS. Naturally, this phenomenon just is the other side of the coin to penalising Type II errors more heavily, which we argued previously is a *feature* and not a *defect* of KSS. This is also something we identify in the graphs: with larger values of k HSS starts to exceed KSS, as Type II errors begin to exceed Type I errors.

4.4 The (ir)relevance of the Success Rate for NN forecasting

Heierli et al. also provide the *success rate*, *SR*, $\frac{a}{a+b}$ in Table 2, which is also known as the *positive predictive value*. What, however, is its relevance for RSE forecasting and should it have any influence on our choice of k?

~~Let's~~ first look at its behaviour. In the case of the Scottish dataset, SR more or less plateaus from $k = 5$ onwards. As a, hence the POD, is decreasing, b must be decreasing too and "in step". In the Swiss dataset, something else is going on. After $k = 4$, SR falls dramatically, indicating that while the number of verified positive predictions drop, the number of mistaken positive predictions does *not* drop in step. Moreover, in neither dataset does SR tend to 1 as k increases, meaning that a sizeable *proportion* of positive predictions are mistaken even when a comparatively high decision boundary is employed. In the Scottish case, SR plateaus at 0.5, meaning that while the number of positive predictions decreases as k increases, the proportion of such predictions that are mistaken falls to 50% and stays there. In the Swiss case, after improving up to $k = 5$, the SR drops dramatically, meaning that while the number of positive predictions has decreased between $k = 5$ and $k = 6$, the *proportion* of predictions that are mistaken has increased. Notice too that in the Swiss case, the SR never gets above 0.3, so a full 70% of positive predictions are mistaken, no matter the value of k—at least two out of three predictions of avalanches are mistaken.

So in both data sets SR might seem initially quite low. But as we know, forecasting rare events is difficult, and we should not be too surprised that the success rate of predicting rare events is less than 50%. In fact, given that rare event forecasting involves, by definition, low base rates of occurrence, and given our limited abilities in forecasting natural disasters such as avalanches, we should expect a low success rate (see Ebert, 2019; Techel et.al., 2020). But there are stronger reasons not to consider SR when assessing the "goodness" of an RSE forecast. SR fails all three adequacy constraints: it does not correct for chance, it has the wrong direction of fit since it is a ratio with denominator $a + b$, and it in effect only takes into account Type I errors. Given this comprehensive failure to meet our criteria of adequacy, we think that, in contrast to accuracy, SR is not even a suitable candidate to break a tie between two equally skillful forecasts.

So, then what are the main lessons from this practical interlude? Simply put: having the appropriate skill measure really does matter and has consequences for high-stakes practical decisions. Forecasters have to make an informed choice in the context of NN forecasting about which decision boundary to adopt. That choice has to be informed by an assessment of which decision boundary issues in the *best* forecast. Our discussion highlighted that the best forecast cannot simply be the most accurate one, rather it has to be the most skillful one. The Peirce skill measure (KSS) is, as we argued earlier, the only commonly used measure that captures the skill involved in RSE forecasting. Finally, if different k's are scored equally on the Peirce score, then we think that accuracy considerations should be used to break the tie: amongst the most skillful.

5 Conceptual challenges for RSE forecasting: the scope problem.

In this last section, we discuss a conceptual challenge for the viability of RSE forecasting (for a general overview of the other conceptual, physical, and human challenges in avalanche forecasting specifically, see (McClung, 2002, a)). Once again, we can draw on insights from the early pioneers of RSE forecast-verification to guide our discussion. In his annual report for 1887, the Chief Signal Officer, Brigadier General Adolphus Greely, noted a practical difficulty facing the forecasting of tornadoes; more specifically:

> So almost infinitesimal is the area covered by a line of tornado in comparison with the area of the state in which it occurs, that even could the Indications Officer say with absolute certainty that a tornado would occur in any particular state or even county, it is believed that the harm done by such a prediction would eventually be greater than that which results from the tornado itself. (Greely, 1887, pp. 21-2)

Now, there are two issues to be distinguished. First, there is the behavioural issue of how the public reacts to forecasts of tornadoes or other rare and severe events. In particular, there is a potential for overreaction which, in turn, led for many years in the United States to the word ‘tornado’ not being used when issuing forecasts (*cf.* Abbe, 1899; Bradford, 1999)! This policy option, to decide not to forecast rare events, is quite radical and no longer reflects current practice.

The other issue is the “almost infinitesimal” track of a tornado compared to the area for which warning of a tornado is given. A broadly similar issue faces regional avalanche forecasting: currently such forecasts are given for a wide region of at least 100km^2, yet avalanches usually occur on fairly localised slopes of which there are many in each region. And, while avalanches are different to many other natural disasters in that they are usually triggered by humans (Schweizer and Lütschg (2001) suggest that roughly 9 out of 10 avalanche fatalities involve a human trigger), RSE forecasts quite generally face what we call the *scope challenge*: The greater the area covered by the binary RSE forecast the less informative it is. Conversely, the smaller the size of the forecast region, the rarer the associated event and the more over-forecasting we can expect.

This type of trade-off applies equally to probabilistic and binary categorical forecasts. One consequence of the scope challenge, alluded to in the above quote, is that once the region is sufficiently large, forecasters may rightly be highly confident that one such event will occur. This means that on a large-scale level, we are not—technically speaking—dealing with *rare event* forecasts anymore, while on a more local level, the risk of such an event is still very low.

Now, ~~in a recent discussion,~~ Statham et.al. (2018) in effect appeal to a version of the scope problem—with an added twist of how to interpret verbal probabilities given variations in scope—as one reason why probabilistic (or indeed binary) forecasts are rarely used in avalanche forecasting. They write:

> The probability of an avalanche on a single slope of 0.01 could be considered likely, while the probability of an avalanche across an entire region of 0.1 could be considered unlikely. This dichotomy, combined with a lack of valid data and the impracticality of calculating probabilities during real-time operations, is the main reasons forecasters do not usually work with probabilities, but instead rely on inference and judgment to estimate likelihood. Numeric probabilities can be assigned when the spatial and temporal scales are fixed and the data are available, but given the time constraints and variable scales of avalanche forecasting, probability values are not commonly used. (Statham et.al., 2018, p. 682)

It might well be these kinds of problems that led in 1993 to the introduction of the European Avalanche Danger scale which involves a multi-categorical five point danger rating: low, moderate, considerable, high, very high. The danger scale itself is a function of ~~snow pack~~ stability, its spatial distribution, and potential avalanche size and it applies to a region of at least 100 km^2. The danger scale, at least on the face of it, focuses more on the conditions (~~snow pack~~ and spatial variation) that render avalanches more or less likely than on issuing specific probabilistic forecasts or predicting actual occurrences.

Given this development, verification of such avalanche forecasts has become more challenging. What makes it even more difficult is that each individual danger level involves varied and complex descriptors that are commonly used to communicate and interpret the danger levels. For example, the danger level *high* is defined as:

> Triggering is *likely*, even from *low additional loads* [i.e. a single skier, in contrast to high additional load, i.e. group of skiers], on *many* steep slopes. In some cases, numerous large and often very large natural avalanches can be expected. (EAWS, 2018)

The descriptor involves verbal probability terms—such as *likely*—that are left undefined, it contains conditional probabilities with nested modal claims [*given a low load trigger, ~~it's~~ likely there will be an avalanche on many slopes*]. And finally, it involves a hedged expectation statement of natural avalanches (i.e. those that are not human triggered) and their predicted size—in *some cases*, *numerous* large or very large natural avalanches can be expected. Noteworthy here is that while the forecasts are *intended* for large forecast areas only, the actual descriptors aim to make the regional rating relevant to local decisions. The side effect of making regional forecasts more locally relevant is that it makes verifying them a hugely complex, if not impossible, task. Naturally, the verification of avalanche forecasts using the five point danger scale is an important and thriving research field and numerous inventive ways to verify multi-categorical avalanche forecasts have since been proposed (Föhn and Schweizer, 1995; Cagnati et.al., 1998; McClung, 2000; Schweizer et.al., 2003; Jamieson et al., 2008; Sharp , 2014; Techel and Schweizer, 2017; Techel et.al., 2018; Statham et.al., 2018a; Schweizer et.al., 2020; Techel et.al., 2020; Techel, 2020). Here, we have to leave a more detailed discussion of which measure to use for multi-categorical forecasts for another occasion. Nonetheless, the now widespread use of multi-categorical forecasts may instead raise the question whether, and if so how, our assessment of skill scores suitable for binary RSE forecast-verification is of more than just historical interest.

There are numerous reasons why we think our discussion is still important with potentially significant practical implications. First, while regional forecasts are usually multi-categorical, there are many avalanche forecasting services that, in effect, have to provide localised binary RSE forecasts. Consider, for example, avalanche forecasting to protect large scale infrastructure such as the Trans Canada Highway along Rogers Pass where ~~over a 40km stretch~~ more than 130 avalanche paths threaten the highway. Ultimately, a binary decision has to be made whether to open or to close the pass and a wrong decision has huge economic impact in the case of both Type I and Type II errors; in the case of Type II errors there is in addition potential loss of life. Similarly so on a smaller scale: while regional multi-categorical forecasts usually inform and influence local decision-making, ultimately operational decisions in ski resorts or other ski operations are binary decisions—whether to open or to close a slope—that are structurally similar to binary RSE forecasts. These kinds of binary forecasting decisions will benefit from using forecast-verification methods that adopt the right skill measure.

Second, our discussion is relevant to the assessment of different localised slope specific stability tests widely used by professional forecasters, mountain guides, operational avalanche risk managers, and recreational skiers, mountaineers, and snowmobilers. A recent large scale study by Techel et.al. (2020) compared two different slope specific stability tests—the Extended Column Test and the so-called Rutschblock Test—and assessed their accuracy and success rate. Our discussion suggests that when assessing the "goodness" of what are in effect local *diagnostic* stability tests, or indeed when assessing the performance of individuals who use such tests, we should treat them as binary RSE forecasts. Using the correct skill score will be crucial to settle which type of stability test is the *better* test from a forecasting perspective.

Lastly, there are, as we noted above, numerous research projects to design manageable forecast-verification procedures for multi-categorical regional forecast. Assuming that the methodological and conceptual challenges we raised earlier can be overcome, we still require the right kind of measures to assess the "goodness" of multi-categorical forecasts. The Heidke, Peirce, and the other measures we discussed can be adapted for these kinds of forecasts. Moreover, given that the danger rating of *high* and *very high* are rarely used, and involve high stakes with often major economic consequences, our discussion may once again help to inform future discussions about how best to verify regional multi-categorical forecast. However, an in-depth discussion of multi-categorical skill measures for regional avalanche forecasts has to wait for another occasion as it will crucially depend on the details of the verification procedure.

6 Conclusions

In his classic 1993 article "What is a good forecast?" Murphy distinguished three types of goodness in relation to weather forecasts generally; all three apply to evaluations of RSE forecasts.

Type 1 goodness: consistency a good fit between the forecast and the forecasters best judgement given their evidence.

Type 2 goodness: quality a good fit between forecast and the matching observations.

Type 3 goodness: value the relative benefits for end-user's decision-making.

Our discussion has focused exclusively on what Murphy labelled the issue of *quality* and how to identify a good fit between binary forecasts and observations, though the *quality* of a forecast has, obviously, knock-on effects on the *value* of a forecast (Murphy, 1993, p. 289). Historically, a number of different measures have been used to assess the quality—the goodness of fit—of individual RSE forecasts and to justify comparative judgements about different RSE forecasts (such as in the case of NN-forecasting). However, there has not been any consensus about which measure is the most relevant in the context of binary RSE forecasts. In this article, we presented three adequacy constraints that any measure has to meet to properly be used in an assessment of the *quality* of a binary RSE forecast. We offered a comprehensive survey of the most widely used measures and argued that there is really only one skill measure that meets all three constraints. Our main conclusion is that goodness (i.e. quality) of a binary RSE forecast should be assessed using the Peirce skill measure, possibly augmented with consideration of accuracy. Moreover, we argued that the same considerations apply to the assessment of slope specific stability tests and other forecasting tools used in avalanche management. Finally, our discussion raises important theoretical questions for the thriving research project of verifying regional multi-categorical avalanche forecasts that we plan to tackle in future work.

Appendix A: Numbers of predictions correct and incorrect "by chance"

We model the actual presences and absences (e.g., occurrence and non-occurrence of avalanches) as constituting the sequence of outcomes produced by $n = a+b+c+d$ independent, identically-distributed random variables X_1, X_2 $\ldots$, X_n; each X_i takes two possible values, 1 (= presence) and 0 (= absence); each random variable takes value 1 with (unknown) probability p. The probability of producing the actual sequence of $a+c$ presences and $b+d$ absences is

$$p^{a+c}(1-p)^{b+d}.$$

The value of p which maximises this is $\dfrac{a+c}{a+b+c+d}$. We take this as the probability of presence on any forecasting occasion. Call it $\hat{p}$. $\hat{p}$ is the *maximum likelihood estimate* of the (unknown) probability of presence.

Putting the "random" into random prognostication, Step 1 We assume the actual forecasting performance to be produced by $n = a+b+c+d$ independent, identically-distributed random variables $Y_1, Y_2 \ldots, Y_n$; each Y_i takes two possible values, 1 (= prediction of presence) and 0 (= prediction of absence); each random variable takes value 1 with (unknown) probability q. The probability of the actual sequence of $a+b$ predictions of presence and $c+d$ predictions of absence is

$$q^{a+b}(1-q)^{c+d}.$$

$\hat{q}$, the maximum likelihood estimate of the unknown value q, is $\dfrac{a+b}{a+b+c+d}$. We take this as the probability of prediction of presence on any forecasting occasion.

Putting the "random" into random prognostication, Step 2 The probability of successful prediction of presence on the i^{th} trial is $prob(X_i = 1 \text{ and } Y_i = 1)$. We suppose that X_i and Y_j are independent, $1 \le i, j \le a+b+c+d$. In particular, then, the probability of successful prediction of presence on the i^{th} trial is $\hat{p}\hat{q}$.

Let $n = a+b+c+d$. Let $Z_i = 1$ if $X_i = 1$ and $Y_i = 1$, $Z_i = 0$ otherwise. The expected value of $Z_1 + Z_2 + \ldots + Z_n$ is

$$\sum_{i=0}^{n} i \frac{n!}{i!(n-i)!} (\hat{p}\hat{q})^i (1-\hat{p}\hat{q})^{n-i}.$$

This is

$$n\hat{p}\hat{q}, \quad i.e., \quad \frac{(a+c)(a+b)}{a+b+c+d}.$$

This is a_r, the "number" of successful predictions of presence we attribute to chance.

In similar fashion, we obtain b_r, c_r and d_r.

Appendix B: The effect of increases and decreases in errors on the Heidke Skill Score

Keeping the marginal totals $a+c$ and $b+d$ fixed, let us consider the score with an additional k Type I errors and again with an additional k Type II errors, $0 < k \le \min\{a, d\}$ (Table B1). We have, by hypothesis, that $0 < a+c < b+d$.

		Observed			Observed		
		+	-	*totals*	+	-	*totals*
Predicted	+	*a*	*b + k*	*a + b + k*	*a - k*	*b*	*a + b - k*
	-	*c*	*d - k*	*c + d - k*	*c + k*	*d*	*c + d + k*
	totals	*a + c*	*b + d*	*a+b+c+d*	*a + c*	*b + d*	*a+b+c+d*

Table B1. An increase of k Type I errors (left) and k Type II errors (right)

With an additional k Type I errors, the Doolittle–Heidke Skill Score is:

$$\frac{2(a(d-k)-(b+k)c)}{(a+b+k)(b+d)+(a+c)(c+d-k)}$$
$$=\frac{2(ad-bc)-2(a+c)k}{(a+b)(b+d)+(a+c)(c+d)+k[(b+d)-(a+c)]}.$$

With an additional k Type II errors, the Doolittle–Heidke Skill Score is:

$$\frac{2((a-k)d-b(c+k))}{(a+b-k)(b+d)+(a+c)(c+d+k)}$$
$$=\frac{2(ad-bc)-2(b+d)k}{(a+b)(b+d)+(a+c)(c+d)-k[(b+d)-(a+c)]}.$$

For k in the range 0 to $\min\{a,d\}$, the denominators are positive.

Let $x=ad-bc$, $y_1=a+c$, $y_2=b+d$, $z=(a+b)(b+d)+(a+c)(c+d)$ and $w=(b+d)-(a+c)$. The score for an additional k Type II errors is no less than the score for an additional k Type I errors, if and only if,

$$\frac{2x-2ky_1}{z+kw}\leq\frac{2x-2ky_2}{z-kw}\text{ iff }(x-ky_1)(z-kw)\leq(x-ky_2)(z+kw)$$
$$\text{iff }k^2w(y_1+y_2)\leq 2xkw-kz(y_2-y_1)$$
$$\text{iff }k(y_1+y_2)\leq 2x-z\text{ as }y_2-y_1=w>0\text{ and }k>0$$
$$\text{iff }k(a+b+c+d)\leq 2(ad-bc)-[(a+b)(b+d)+(a+c)(c+d)]$$
$$=-(b+c)(a+b+c+d)\leq 0,$$

which is impossible.

Let us consider next the score after a reduction of k Type I errors and after a reduction k Type II errors, $0<k\leq\min\{b,c\}$ (Table B2).

		Observed			Observed		
		+	-	*totals*	+	-	*totals*
Predicted	+	*a*	*b - k*	*a + b - k*	*a + k*	*b*	*a + b + k*
	-	*c*	*d + k*	*c + d + k*	*c - k*	*d*	*c + d - k*
	totals	*a + c*	*b + d*	*a+b+c+d*	*a + c*	*b + d*	*a+b+c+d*

Table B2. A decrease of k Type I errors (left) and k Type II errors (right)

By hypothesis, we have that $0<a+c<b+d$.

With a reduction of k Type I errors, the Doolittle–Heidke Skill Score is:

$$\frac{2(a(d+k)-(b-k)c)}{(a+b-k)(b+d)+(a+c)(c+d+k)}$$
$$=\frac{2(ad-bc)+2(a+c)k}{(a+b)(b+d)+(a+c)(c+d)-k[(b+d)-(a+c)]}.$$

With a reduction of k Type II errors, the Doolittle–Heidke Skill Score is:

$$\frac{2((a+k)d-b(c-k))}{(a+b+k)(b+d)+(a+c)(c+d-k)}$$
$$=\frac{2(ad-bc)+2(b+d)k}{(a+b)(b+d)+(a+c)(c+d)+k[(b+d)-(a+c)]}.$$

For k in the range 0 to $\min\{b,c\}$, the denominators are positive.

In the notation introduced above, the score for a reduction of k Type II errors is no greater than the score for a reduction of k Type I errors, if and only if,

$$\frac{2x+2ky_1}{z-kw} \geq \frac{2x+2ky_2}{z+kw} \text{ iff } (x+ky_1)(z+kw) \geq (x+ky_2)(z-kw)$$
$$\text{iff } k^2w(y_1+y_2)+2xkw \geq kz(y_2-y_1)$$
$$\text{iff } k(y_1+y_2) \geq z-2x \text{ as } y_2-y_1=w \geq 0 \text{ and } k \geq 0$$
$$\text{iff } k(a+b+c+d) \geq 2(ad-bc)-[(a+b)(b+d)+(a+c)(c+d)]$$
$$=(b+c)(a+b+c+d)$$
$$\text{iff } k \geq b+c,$$

which is impossible, since $0 \leq k \leq \min\{b,c\}$.

~~It's~~ clear that these results reverse when the forecasted events are common, *i.e.*, when $a+c > b+d$.

Appendix C: *KSS* and *HSS*

We assume that $ad > bc$ and that $b+d > a+c > 0$. Then

$$\frac{a}{a+c}-\frac{b}{b+d}=KSS \gtreqless HSS=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$$
$$\text{iff } \frac{ad-bc}{(a+c)(b+d)} \gtreqless \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$$
$$\text{iff } (a+b)(b+d)+(a+c)(c+d) \gtreqless 2(a+c)(b+d) \text{ as } ad > bc$$
$$\text{iff } (b+d)[(a+b)-(a+c)] \gtreqless (a+c)[(b+d)-(c+d)]$$
$$\text{iff } (b-c)[(b+d)-(a+c)] \gtreqless 0$$
$$\text{iff } b \gtreqless c \text{ as } b+d > a+c.$$

Appendix D: Skill scores in the literature

All scores are to be understood relative to Table 2. The *root mean square contingency* is the geometric mean of Clayton and Peirce Skill Scores. Its square is the measure proposed by Doolittle that attracted Farquhar's censure as discussed in section 3.2. See also (Wilks, 2019, Ch. 8) for a general overview of skill scores for binary forecast verification. Note that we disagree with some aspects of his assessment.

name	*definition*
accuracy / proportion correct / hit score	$\frac{a+d}{a+b+c+d}$
Threat Score/Critical Success Index Jaccard coefficient (Jolliffe, 2016)	$\frac{a}{a+b+c}$
Fowlkes–Mallow index cosine similarity	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Dice co-efficient (Jolliffe, 2016) F-score, F_1	$\frac{2a}{2a+b+c}$
adjusted F-measure, F_β	$\frac{(1+\beta^2)a}{(1+\beta^2)a+b+\beta^2 c}$
Equitable Threat Score Gilbert Skill Score	$\frac{ad-bc}{(ad-bc)+(a+b+c+d)(b+c)}$
Clayton Skill Score	$\frac{ad-bc}{(a+b)(c+d)}$
Heidke Skill Score / Cohen's κ	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$
Peirce Skill Score / Youden index Hanssen-Kuipers discriminant / Skill Score True Skill Statistic	$\frac{ad-bc}{(a+c)(b+d)}$
root mean square contingency, ϕ Matthews correlation coefficient	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
R(elative) I(mprovement) O(ver) C(hance)	$\frac{ad-bc}{(a+\min\{b,c\})(\min\{b,c\}+d)}$
The Skill Test (Woodcock, 1976)	$\frac{4(ad-bc)}{(a+b+c+d)^2}$
Odds Ratio Skill Score (Stephenson, 2000)	$\frac{ad-bc}{ad+bc}$

Table D1. Skill scores for binary, categorical forecasting

Author contributions. These authors contributed equally to this work.

Competing interests. None

Acknowledgements. Philip Ebert's research was supported by the Arts and Humanities Research Council AH/T002638/1 "Varieties of Risk". Publication costs were covered by the University of Stirling APC fund. We are grateful to the International Glaciological Society and Joachim Heierli for the permission to reuse Figure 1.

References

Abbe, C.: Unnecessary tornado alarms, Mon. Weather Rev., 27, 255, https://doi.org/10.1175/1520-0493(1899)27[255c:UTA]2.0.CO;2, 1899.

Akosa, J. S.: Predictive accuracy: A misleading performance measure for highly imbalanced data, in: SAS Global Forum 2017, April 2–5, Orlando FL, USA, Paper 924, http://support.sas.com/resources/papers/proceedings17/0942-2017.pdf, last access: 6 August 2021, 2017.

Altman, D. G. and J. M. Bland: Diagnostic tests 3: receiver operating characteristic plots, BMJ, 309, 188, https://doi.org/10.1136/bmj.309.6948.188, 1994.

Anscombe, G. E. M.: Intention, second edition, Basil Blackwell, Oxford,1963.

Benini, R.: Principii di Demografia, vol. 29 of Manuali Barbèra di Scienze Giuridiche, Sociali e Politiche, G. Barbèra, Florence, 1901.

Brabec, B. and Meister, R.: A nearest-neighbor model for regional avalanche forecasting, Ann. Glaciol., 32, 130-134, https://doi.org/10.3189/172756401781819247, 2001.

Bradford, M.: Historical roots of modern tornado forecasts and warnings, Weather Forecast., 14, 484–91, https://doi.org/10.1175/1520-0434(1999)014<0484:HROMTF>2.0.CO;2, 1999.

Brier, G. W. and Allen, R. A.: Verification of weather forecasts, in: Compendium of Meteorology, edited by: Malone, T. F., American Meteorological Society, Boston, USA, 841-8, 1951.

Brownlee, J.: Classification Accuracy is Not Enough: More Performance Measures You Can Use, Machine Learning Mastery, https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/, March 21, 2014.

Brownlee, J.: Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning, independently published, https://machinelearningmastery.com/ imbalanced-classification-with-python/, 2020.

Bruckhaus, T.: The business impact of predictive analytics, in: Knowledge Discovery and Data Mining: Challenges and Realities, edited by Zhu, X. and Davidson, I, Information Science Reference/IGI Global, Hershey and London, 114–138, 2007.

Buser, O.: Avalanche forecast with the method of nearest neighbours: An interactive approach, Cold Reg. Sci. and Tech., 8(2), 155–163, https://doi.org/10.1016/0165-232X(83)90006-X, 1983.

Buser, O.: Two years experience of operational avalanche forecasting using the nearest neighbour method, Ann. Glaciol., 13, 31-34, https://doi.org/10.3189/S026030550000759X, 1989.

Buser, O., Bütler, M., and Good, W.: Avalanche forecast by the nearest neighbor method, in: Avalanche Formation, Movement and Effects: Proceedings of a Conference Held at Davos, September 1986, International Association of Hydrological Sciences Publications, vol. 162, edited by: Salm, B. and Gubler, H., IAHS Press, Wallingford, UK, 557-570, 1987.

Cagnati, A., Valt, M., Soratroi, G., Gavaldà, J., and Sellés, C. G.: A field method for avalanche danger-level verification, Ann. Glaciol., 26, 343–346, https://doi.org/10.3189/1998aog26-1-343-346, 1998.

Clayton, H. H.: A method of verifying weather forecasts, B. Am. Meteorol. Soc., 8, 144–6, https://doi.org/10.1175/1520-0477-8.10.144, 1927.

Clayton, H. H.: Rating weather forecasts [with discussion], B. Am. Meteorol. Soc., 15, 279–82, 114–138, 1934. https://doi.org/10.1175/1520-0477-15.12.279

Clayton, H. H.: Verifying weather forecasts, B. Am. Meteorol. Soc., 22, 314–5, 10.1175/1520-0477-22.8.314, 1941.

Cole, L. C,, The measurement of interspecific association, Ecol., 30, 411–424, https://doi.org/10.2307/1932444, 1949.

Copas, J. B. and Loeber, R.: Relative improvement over chance (RIOC) for 2×2 tables, Brit. J. Math. Stat. Psy., 43, 293–307, https://doi.org/10.1111/j.2044-8317.1990.tb00942.x, 1990.

Davis, K. and R. Maiden: The importance of understanding false discoveries and the accuracy paradox when evaluating quantitative studies, Studies in Social Science Research, 2 (2): 1–8, https://doi.org/10.22158/sssr.v2n2p1, 2021.

Deeks, J. J. and D. G. Altman: Statistics notes: Diagnostic tests 4: likelihood ratios, BMJ, 329, 168–369, https://doi.org/10.1136/bmj.329.7458.168, 2004.

Doolittle, M. H.: The verification of predictions, Bull. Philosoph. Soc. Washington, 7, 122–7, 1885a.

Doolittle, M. H.: The verification of predictions [Abstract], Am. Meteorol. J., 2, 327–29, 1885b.

Doolittle, M. H.: Association ratios, Bull. Philosoph. Soc. Washington, 10, 83–7, 1988.

Ebert, P. A.: Bayesian reasoning in avalanche terrain: a theoretical investigation, Journal of Adventure Education and Outdoor Learning, 19, 84–95, https://doi.org/10.1080/14729679.2018.1508356, 2019.

European Avalanche Warning Services (EAWS), European Avalanche Danger Scale, https://www.avalanches.org/wp-content/uploads/2019/05/European_Avalanche_Danger_Scale-EAWS.pdf, last access: 24 June 2021, 2018.

Farquhar, H.: Verification of predictions, Science, 4, 540, https://doi.org/10.1126/science.ns-4.98.540, 1884.

Farrington, D. P.: Predicting Individual Crime Rates, in: Prediction and Classification: Criminal Justice Decision Making, edited by: Gottfredson, D. M., and Tonry, M., Crime and Justice, vol. 9, University of Chicago Press, Chicago IL, 53–101, 1987.

Farrington, D. P. and Loeber, R.: Relative improvement over chance (RIOC) and phi as measures of predictive efficiency and strength of association in 2×2 tables, J. Quant. Criminol., 5, 201–13, https://doi.org/10.1007/BF01062737, 1989.

Fernandes, J. A., Irigoien, X., Goikoetxea, N, Lozano, J. A., Inza, I., Pérez, A., and Bode, A.: Fish recruitment prediction, using robust supervised classification methods, Ecol. Model., 221, 338–52, https://doi.org/10.1016/j.ecolmodel.2009.09.020, 2010.

Finley, J. P.: Tornado predictions, Am. Meteorol. J., 1, 85–88, 1884.

Flueck, J. A.: A study of some measures of forecast verification, in: Preprints. 10th Conference on Probability and Statistics in Atmospheric Sciences, Edmonton, AB, Canada, 69–73, American Meteorological Society, Boston MA, 1987.

Föhn, P. M. B. and Schweizer, J.: Verification of avalanche hazard with respect to avalanche forecasting, in: Les apports de la recherche scientifique À la sécurité neige, glace et avalanche. Actes de colloque, Chamonix, 30 mai – 3 juin 1995, ANENA, Grenoble, France, 151-156, 1995.

Forbes, S. A.: Method of determining and measuring the associative relations of species, Science, 61, 524, https://doi.org/10.1126/science.61.1585.518.b, 1925.

G.: Letter to the editor: Tornado predictions, Science, 4, 126–7, https://doi.org/10.1126/science.ns-4.80.126, 1884.

Gandin, L. S. and A. H. Murphy: Equitable skill scores for categorical forecasts, Monthly Weather Review, 120, 361–370, https://dio.org/10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2, 1992.

Gassner, M., Birkeland, K., Etter, H. J., and Leonard, T.: NXD2000: An improved avalanche forecasting program based upon the nearest neighbour method, in: Proceedings of the International Snow Science Workshop, 2000, Big Sky, Montana, USA, 52-59, http://arc.lib.montana.edu/snow-science/item/706, last access: 6 August 2021, 2001.

Gassner, M., and Brabec, B.: Nearest neighbour models for local and regional avalanche forecasting, Nat. Hazards Earth Syst. Sci., 2, 247–253, https://doi.org/10.5194/nhess-2-247-2002, 2002.

Gilbert, G. K.: Finley's tornado predictions, Am. Meteorol. J., 1, 166–172, 1884.

Goffman, William and Newill, Vaun A.: A methodology for test and evaluation of information retrieval systems, Information Storage and Retrieval, 3, 19-25, https://doi.org/10.1016/0020-0271(66)90006-4, 1966.

Greely, A. W.: Annual Report of the Chief Signal Officer of the Army to the Secretary of War for the Year 1887. In Two Parts. Part I. Government Printing Office, Washington, with contributory reports by other authors, 1887.

Hanssen, A. W. and Kuipers, W. J. A.: On the relationship between the frequency of rain and various meteorological parameters (With reference to the problem of objective forecasting), Koninklijk Nederlands Meteorologisch Instituut Mededelingen en Verhandelingen, vol. 81. Staatsdrukkerij- en Uitgeverijbedrijf, 's-Gravenhage, 1965.

Heidke, P.: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst, Geogr. Ann., 8, 301–349, https://doi.org/10.2307/519729, 1926.

Heine, M. H.: Distance between sets as an objective measure of retrieval effectiveness, Information Storage and Retrieval, 9, 181–198, https://doi.org/10.1016/0020-0271(73)90066-1, 1973.

Heierli, J., Purves, R. S., Felber, A., and Kowalski, J.: Verification of nearest-neighbours interpretations in avalanche forecasting, Ann. Glaciol., 38, 84–8, https://doi.org/10.3189/172756404781815095, 2004.

Hirschberg, J., Badoux, A., McArdell, B. W., Leonarduzzi, E., and Molnar, P.: Evaluating methods for debris-flow prediction based on rainfall in an Alpine catchment, Nat. Hazards Earth Syst. Sci., 21, 2773–2789, https://doi.org/10.5194/nhess-21-2773-2021, 2021.

Jamieson, B., Campbell, C., and Jones, A.: Verification of Canadian avalanche bulletins including spatial and temporal scale effects, Cold Reg. Sci. Technol., 51, 204–213, https://doi.org/10.1016/j.coldregions.2007.03.012, 2007.

Jardine, N. and C. J. van Rijsbergen, The use of hierarchic clustering in information retrieval, Information Retrieval and Storage, 7, 217–240, https://doi.org/10.1016/0020-0271(71)90051-9, 1971.

Jolliffe, I. T.: The Dice co-efficient: a neglected verification performance measure for deterministic forecasts of binary events, Meteorol. Appl., 23, 89–90, https://doi.org/10.1002/met.1532, 2016.

Johnson, H. M.: Maximal selectivity, correctivity and correlation obtainable in 2 × 2 contingency-tables, Am. J. Psychol., 58, 65–8. https://doi.org/10.2307/1417575, 1945.

Leonarduzzi, E., and Molnar, P.: Deriving rainfall thresholds for landsliding at the regional scale: daily and hourly resolutions, normalisation, and antecedent rainfall, Nat. Hazards Earth Syst. Sci., 20, 2905–2919. https://doi.org/10.5194/nhess-20-2905-2020, 2020.

Lesk, M. E. and G. Salton: Relevance assessments and retrieval system evaluation, Information Storage and Retrieval, 4, 343–359, https://doi.org/10.1016/0020-0271(68)90029-6, 1969.

Loeber, R. and Dishion, T.: Early predictors of male delinquency: A review, Psychol. Bull., 94, 68–99, 1983. https://doi.org/10.1037/0033-2909.94.1.68

Loeber, R. and Stouthamer-Loeber, M.: Family factors as correlates and predictors of juvenile conduct problems and delinquency, Crime Justice, 7, 29–149, https://doi.org/10.1086/449112, 1986.

Manzato, A.: An Odds Ratio Parametrization for ROC Diagram and Skill Score Indices, Weather and Forecasting, 20, 918-930, https://doi.org/10.1175/WAF899.1, 2005.

Manzato, A.: A Note on the Maximum Peirce Skill Score, Weather and Forecasting, 22, 1148-1154, https://doi.org/10.1175/WAF1041.1, 2007.

McClung, D. M.: Predictions in avalanche forecasting, Ann. Glaciol., 31, 377–381, https://doi.org/10.3189/172756400781820507, 2000.

McClung, D. M.: The elements of applied avalanche forecasting, Part I: The human issues, Nat. Hazards, 26, 111–129, http://link.springer.com/article/10.1023/A:1015665432221, 2002.

McClung, D. M.: The elements of applied avalanche forecasting, Part II: the physical issues and the rules of applied avalanche forecasting, Nat. Hazards, 26, 131–146, http://link.springer.com/article/10.1023/A:1015604600361, 2002a.

Metz, C. E.: Basic principles of ROC analysis, Seminars in Nuclear Medicine, 8, 283–298, https://doi.org/10.1016/S0001-2998(78)80014-2, 1978.

Milne, P.: Rereading Peirce: The inevitability of the Peirce Skill Score as a measure of skill in binary, categorical forecasting, manuscript, 2021.

Murphy, A. H.: Probabilities, odds, and forecasts of rare events, Weather Forecast., 6, 302–7, https://doi.org/10.1175/1520-0434(1991)006<0302:POAFOR>2.0.CO;2, 1991.

Murphy, A. H.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather Forecast., 8, 281-293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.

Peirce, C. S.: The numerical measure of the success of predictions, Science, 4, 453–4, https://doi.org/10.1126/science.ns-4.93.453-a, 1884.

Powers, D. M. W.: What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes, Technical Report KIT-14-001 Computer Science, Engineering & Mathematics, Flinders University, https://arxiv.org/pdf/1503.06410v2, 2015, revised 2019.

Purves, R. S., Morrison, K. W., Moss, G., and Wright, D. S. B.: Nearest neighbours for avalanche forecasting in Scotland—development, verification and optimisation of a model, Cold Reg. Sci. Technol., 37, 343-355, https://doi.org/10.1016/S0165-232X(03)00075-2, 2003.

Purves, R. S. and Heierli, J.: Evaluating nearest neighbours in avalanche forecasting—a qualitative approach to assessing information content, in: Proceedings of the International Snow Science Workshop, Telluride, Colorado, USA, 701-708, http://arc.lib.montana.edu/snow-science/item/1004, last access: 6 August 2021, 2006.

Robertson, S. E.: Parametric description of retrieval tests: Part I: The basic parameters, Journal of Documentation, 25, 1–27, https://doi.org/10.1108/eb026462, 1969.

Robertson, S. E.: Parametric description of retrieval tests: Part II: Overall measures, Journal of Documentation, 25, 93-107, https://doi.org/10.1108/eb026466, 1969a.

Roeger, C., McClung, D., Stull, R., Hacker, J., and Modzelewski, H.: A verification of numerical weather forecasts for avalanche prediction, Cold Reg. Sci. Technol., 33, 189–205, https://doi.org/10.1016/S0165-232X(01)00059-3, 2004.

Schweizer, J., and Lütschg, M.: Characteristics of human-triggered avalanches, Cold Reg. Sci. Technol., 33, 147–162, http://www.sciencedirect.com/science/article/pii/S0165232X01000374, 2001.

Schweizer, J., K. Kronholm, and T. Wiesinger.: Verification of regional snowpack stability and avalanche danger, Cold Reg. Sci. Technol., 37, 277–288, https://doi.org/10.1016/S0165-232X(03)00070-3, 2003.

Schweizer, J., Mitterer, C., Techel, F., Stoffel, A., and Reuter, B.: On the relation between avalanche occurrence and avalanche danger level, The Cryosphere, 14, 737-750, https://doi.org/10.5194/tc-2019-218, 2020.

Sharp, E.: Avalanche Forecast Verification Through a Comparison of Local Nowcasts with Regional Forecasts, in: Proceedings of the International Snow Science Workshop, Banff, Canada, 475–480, http://arc.lib.montana.edu/snow-science/item/2098, last access: 6 August 2021, 2014.

Šimundić, A.-M.: Diagnostic accuracy – Part 1 Basic concepts: sensitivity and specificity, ROC analysis, STARD statement, acutecaretesting.org, Radiometer Medical ApS, June 2009. Reprinted in Point of Care, 11, 6–8, https://doi.org/10.1097/POC.0b013e318246a5d6, 2012.

Šimundić, A.-M.: Measures of diagnostic accuracy: basic definitions. Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine, 19, 203–211, www.ifcc.org/media/476873/ejifcc2008vol19no4pp203-211.pdf, 2009, and as Diagnostic Accuracy—Part 2 Predictive Value and Likelihood Ratio, Radiometer Medical ApS, October 2009. Reprinted as: Diagnostic Accuracy—Part 2 Predictive Value and Likelihood Ratio, Point of Care, 11, 9–11, https://doi.org/10.1097/POC.0b013e318246a5f9, 2012.

Singh, A., and Ganju, A.: A supplement to nearest-neighbour method for avalanche forecasting, Cold Reg. Sci. Technol., 39(2–3), 105–113, https://doi.org/10.1016/j.coldregions.2004.03.005, 2004.

Singh, A., Srinivasan, K., and Ganju, A.: Avalanche Forecast Using Numerical Weather Prediction in Indian Himalaya, Cold Reg. Sci. Technol., 43, 83-92, https://doi.org/10.1016/j.coldregions.2005.05.009, 2005.

Singh, A., Damir, B., Deep, K., and Ganju, A.: Calibration of nearest neighbors model for avalanche forecasting, Cold Reg. Sci. Technol., 109, 33–42, https://doi.org/10.1016/j.coldregions.2014.09.009, 2015.

Statham, G., Haegeli, P., Greene, E., Birkeland, K., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J.: A conceptual model of avalanche hazard, Nat. Hazards, 90, 663–691, https://doi.org/10.1007/s11069-017-3070-5, 2018.

Statham, G., Holeczi, S., and Shandro, B.: Consistency and Accuracy of Public Avalanche Forecasts in Western Canada, in: Proceedings of the 2018 International Snow Science Workshop, Innsbruck, Austria, 1491–1495, http://arc.lib.montana.edu/snow-science/item/2806, last access: 6 August 2021, 2018a.

Stephenson, D. B.: Use of the "odds ratio" for diagnosing forecast skill, Weather Forecast., 15, 221-32, https://doi.org/10.1175/1520-0434(2000)015<0221:UOTORF>2.0.CO;2, 2000.

Swets, John A.: The Relative Operating Characteristic in Psychology, Science, 182, 990-1000, https://dio.org/10.1126/science.182.4116.990, 1973.

Techel, F., and Schweizer, J.: On using local avalanche danger level estimates for regional forecast verification, Cold Reg. Sci. Technol., 144, 52–62, https://doi.org/10.1016/j.coldregions.2017.07.012, 2017.

Techel, F., Mitterer, C., Ceaglio, E., Coléou, C., Morin, S., Rastelli, F., and Purves, R. S.: Spatial consistency and bias in avalanche forecasts—a case study in the European Alps, Nat. Hazards Earth Syst. Sci., 18, 2697–2716, https://doi.org/10.5194/nhess-18-2697-2018, 2018.

Techel, F., Müller, K., and Schweizer, J.: On the importance of snowpack stability, its frequency distribution, and avalanche size in assessing the avalanche danger level: a data-driven approach, The Cryosphere, 14, 3503–352, https://doi.org/10.5194/tc-14-3503-2020, 2020.

Techel, F., Winkler, K., Walcher, M., van Herwijnen, A., and Schweizer, J.: On snow stability interpretation of extended column test results, Nat. Hazards Earth Syst. Sci., 20, 1941–1953, https://doi.org/10.5194/nhess-20-1941-2020, 2020.

Techel, F.: On Consistency and Quality in Public Avalanche Forecasting - a Data-Driven Approach to Forecast Verification and to Refining Definitions of Avalanche Danger. PhD thesis. University of Zürich, Switzerland, pp.227, 2020.

Thomas, C., and Balakrishnan, N.: Improvement in minority attack detection with skewness in network traffic, in: Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, edited by Dasarathy, B. V., SPIE, Bellingham, USA, https://doi.org/10.1117/12.785623, 2008.

Uddin, M. F.: Addressing accuracy paradox using enhanced weighted performance metric in machine learning, in: 2019 Sixth HCT Information Technology Trends (ITT), IEEE/Curran Associates, Red Hook NY, https://doi.org/10.1109/ITT48889.2019, 319–24. 2019.

Valverde-Albacete, F. J., and Peláez-Moreno, C.: 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox, PLOS One, 9, e84217, https://doi.org/10.1371/journal.pone.0084217, 2014.

van Rijsbergen, C. J.: Foundation of evaluation, Journal of Documentation, 30, 365-373, https://doi.org/10.1108/eb026584, 1974.

Wilks, D.S.: Statistical methods in the atmospheric sciences, fourth edition, Academic Press, Oxford, 2019.

Woodcock, F.: The evaluation of yes/no forecasts for scientific and administrative purposes, Mon. Weather Rev., 104, 1209-1214, https://doi.org/10.1175/1520-0493(1976)104<1209:TEOYFF>2.0.CO;2, 1976.

Youden, W. J.: Index for rating diagnostic tests, Cancer, 3, 32-35, https://doi.org/10.1002/4801097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3, 1950.