A rose by any other name: On basic scores from the 2x2 table and the plethora of names attached to them

Harold E. Brooks, a,b Montgomery L. Flora, c,a Michael E. Baldwin c, d

a NOAA/National Severe Storms Laboratory, Norman, OK

b School of Meteorology, University of Oklahoma, Norman, OK

c Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, OK

d NOAA/National Weather Service/Storm Prediction Center, Norman, OK

Revised submission to Artificial Intelligence in the Earth Sciences

February 2024

Corresponding author: Harold E. Brooks, harold.brooks@noaa.gov

Early Online Release: This preliminary version has been accepted for publication in Artificial Intelligence for the Earth Systems, may be fully cited, and has been assigned DOI 10.1175/AIES-D-23-0104.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2024 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
ABSTRACT

Forecast evaluation metrics have been discovered and rediscovered in a variety of contexts, leading to confusion. We look at measures from the 2x2 contingency table and the history of their development and illustrate how different fields working on similar problems has led to different approaches and perspectives of the same mathematical concepts. For example, Probability of Detection is a quantity in meteorology that was also called Prefigurance in the field, while the same thing is named Recall in information science and machine learning, and Sensitivity and True Positive Rate in the medical literature. Many of the scores that combine three elements of the 2x2 table can be seen as either coming from a perspective of Venn diagrams or from the Pythagorean means, possibly weighted, of two ratios of performance measures. Although there are algebraic relationships between the two perspectives, the approaches taken by authors led them in different directions, making it unlikely that they would discover scores that naturally arose from the other approach.

We close by discussing the importance of understanding the implicit or explicit values expressed by the choice of scores. In addition, we make some simple recommendations about the appropriate nomenclature to use when publishing interdisciplinary work.
1. Introduction and background

The problems of forecast evaluation, classification, and comparative relationships between groups of concepts or things share some common characteristics. We can think of them as becoming concerned with the similarity or dissimilarity between two (or more) different descriptions of some “reality.” In weather forecasting, quantifying the similarity of the forecasts and weather associated with the forecasts (observations) describes the quality of the forecast (Murphy 1993). When looking at the cultural traits of different people groups (Driver and Kroeber 1932) or the plant species in different areas (Jaccard 1901), we can use the properties of one group and compare it to the other group to see how alike they are. With similar basic problems facing many fields, multiple metrics that quantify the same properties have been developed and, often, given new names. Identical metrics, but with a bewildering array of names for the same things can lead to confusion, especially in interdisciplinary work where researchers from different backgrounds can end up, essentially, speaking foreign languages to each other. It is useful, however, to see how different fields have thought about solving similar problems and dealing with issues that arise in their fields. In many cases, the applicability of terms developed in a field is obvious, but when that term is adopted by another field, the applicability is not so obvious. In this paper, we’ll focus on some illustrative examples from what is a seemingly simple problem, the so-called 2x2 problem, and make some recommendations for improved, if not best, practices to ease the communication challenges.

2. The 2x2 contingency table and scoring rules

As a starting point, we begin with the classic 2 x 2 contingency table, introduced by Pearson (1904). It is illustrative to note that this is an example of something that has been given a different name in a specific context. Miller and Nicely (1955) introduced the term “confusion matrix” for the same thing as a contingency table in their study of how listeners confused different consonant sounds in English. The name is evident in that context, as it is in the study of cattle feeding behavior of Ruuska et al (2018), but several other areas have adopted that term, such as psychology (Cheng et al. 2023), machine learning (Heydarian et al. 2022), statistical classification (Riehl et al. 2023), and environmental science (Phillips et al. 2024), where it is not so obviously connected to the original context. This does not mean the usage is wrong, but the additional terminology can confuse new users. For the simplest
possible forecast problem of a yes/no forecast and a yes/no observation, we get a 2x2 table (Table 1).

The first use of contingency table-style elements in meteorology was in the tornado forecast experiment of Finley (1884). The history of the early days after Finley’s experiment, the measures derived from the table that were discovered in the immediate aftermath, and the subsequent rediscoveries of those scores is described in Murphy (1996). Researchers discussed several fundamental issues in the first years after Finley, such as the quality of observations, the possible difference in importance of different kinds of forecast errors, and what was an appropriate baseline for how many forecasts should have been correct by chance.

One of the other issues that was discussed in some of the early papers was what we might now think of as the correct forecast of a “non-event” (e.g., no tornado warning was issued and no tornado occurred), the “d” element of Table 1. In some fields, such as geobotany, this term has no meaning. In comparing the species of plants in different Swiss valleys, Jaccard (1901) ignored the existence of “d” completely. This was a sensible thing to do, given that it would be “all of the plants that don’t exist in either of two valleys.” Putting bounds on what constitutes “all” is challenging. Palmer and Allen (1949), in the context of weather forecasts, excluded $d$ by framing the forecasts as being associated with threats, meaning that the $d$ term would be a forecast for a non-threat that didn’t happen. They assumed those cases were “less difficult, perhaps less significant forecasting problem(s).” As a result of this issue, our primary emphasis will be on scores derived from the $a$, $b$, and $c$ terms of the table. Despite the seeming simplicity, Brusco et al. (2021) discuss 18 different metrics relying on those 3 terms, which is not exhaustive. In total, Brusco et al. (2021) compiled 71 metrics that are derived from the 2x2 contingency table and they used synthetic data to examine which scores behave similarly under a variety of conditions, producing clusters of metrics that behave similarly to each other. Other lists of metrics derived from the 2x2 table have been compiled by Murphy (1996), Marzban (1998), and Warrens and van der Hoef (2022).
Observed

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Table 1: The 2x2 contingency table for forecast and observations.

Two philosophical areas have been approached by researchers leading to different metrics.

1. Should we measure similarity or dissimilarity between our two sets?

2. Should we consider the individual elements \((a,b,c)\) or composite statistics like the total number of each category or ratios \((a/(a+b))\)?

As a result of addressing (1), there are many scores that have complements of other scores, so that the two scores would sum to one. Many of the dissimilarity metrics are referred to as “distance”, so that a perfect score would be 0, while most similarity metrics are called “index” or “score” with a perfect value of 1. As for (2) although, obviously, all of the scores can be written in terms of the individual elements, many are written as combinations of the ratios, so the underlying approach makes some scores more accessible and others less so.

\textit{a. Gilbert’s ratio of verification and its copies}

Gilbert (1884) produced what he referred to as a “ratio of verification”

\[ v_g = \frac{a}{a+b+c} \]  \hspace{1cm} (1)

in terms of Table 1. Gilbert derived his ratio by considering the number of forecasts \((a+b)\), the number of events \((a+c)\), and the “coincidences”, \(a\), and wrote (1) as
\[
vg = \frac{a}{(a+b)+(a+c)-a}
\]  
(2)

which reduces to (1), emphasizing the forecasts and events as classes and subtracting out the coincidences. In meteorology, this score would be rediscovered by Palmer and Allen (1949) and given the name “% Success”, later to be called the “Threat Score” by Larue and Younkin (1961). The word “threat” came from Palmer and Allen describing their work as relating to the forecasting of threats. Palmer and Allen worked directly with the three elements of Table 1. They explicitly ignored the \(d\) term in the table because “when no precipitation was forecast and no precipitation was observed, it is assumed that the possible occurrence of precipitation was a less difficult, perhaps insignificant forecast problem.” Probabilistic approaches to forecasting can help quantify the difficulty, and thresholding probabilities to create yes/no forecasts can separate the probabilistic forecasts into a series of 2x2 tables.

Donaldson et al. (1975) also produced the same metric for severe thunderstorm forecasting by looking at the individual elements, \(a\), \(b\), and \(c\). They describe their choice of (1) as “arbitrary” and don’t reference Gilbert, Palmer and Allen, or Larue and Younkin. They gave the score the name “Critical Success Index” (CSI), which has become the most common name in use in meteorology and is the term used in the official verification of severe thunderstorm and tornado warnings by the US National Weather Service.

One of the most common names for \(v_g\) outside of meteorology is the Jaccard index, which came out of the field that might now be referred to as biogeography. It’s widely used in medical research (e.g., Scheller et al. 2023), computer imaging (Amirkahni and Bastanfard 2021), information science (Leydesdorff 2008), and machine learning as seen in its use in Scikit-learn in Python (Pedregosa et al. 2011), This score was developed by the botanist Jaccard (1901)\(^1\) as the *coefficient de communauté* (community coefficient) in his work comparing the distribution of plant species in valleys in Switzerland. Jaccard followed a

\(^1\) Although Jaccard (1901) is clearly the first use, some authors have credited Jaccard (1902, 1907, 1912) as the origin. The 1912 reference is an English translation of the 1907 reference.
similar approach to Gilbert, in that he compiled lists of species in each valley, then did pairwise comparisons between two valleys at a time. In Gilbert’s terms, the list for one valley could be considered as the “forecast” for the “event” species in another valley. From the 2x2 table perspective, the $d$ entry, which would be the plant species that didn’t appear in a pair of valleys, was problematic for Jaccard and, as a result, it was ignored.

The other primary development of the same score as $v_s$ comes from Tanimoto (1958). In contrast to Gilbert and Jaccard, Tanimoto developed his index from a set theory approach, with the score being the intersection of two data sets over their union; commonly known as intersection over union (IoU) today (Wilks et al. 1990). In a spatial perspective, this can be easily seen in the diagrams of Venn (1880). Larue and Younkin (1961) called it the threat score and Stensrud and Wandishin (2000) referred to it as the correspondence ratio for areal forecasts, particularly with application to ensemble forecasts. A timeline showing the introduction of these different names for the scores is given in Figure 1. The applicability to both identification of individual events and spatial fields was highlighted by Mason (1989). Marczewski and Steinhaus (1958) also used a set theory approach to define a dissimilarity metric between two sets that is $1-v_s$. They show that this metric has all of the properties of a mathematical distance.

---

Figure 1: Timeline of introduction of scores of the form of $v_g$.

As part of a larger effort to look at binary similarity coefficients, computational chemists Consonni and Todeschini (2012) built on $v_g$ by taking logarithmic transformations of the numerator and denominator, leading to a similarity coefficient they referred to as $T4$

$$S_{T4} = \frac{\log(1+a)}{\log(1+a+b+c)}$$  

This was one of a set of five scores they developed, using logarithmic transformations based off of previous scores from the literature. This was the only one that ignored the $d$ term in the 2x2 table. They tested their “new” scores in comparison to the original scores as well as some other traditional scores. Most of the logarithmic transformations appeared to provide little new information compared to the original scores. $T4$, however, gave a different ordering in their tests and, hence, provided a different view of performance in their problem. It has not been applied widely, so it’s not clear how it might behave in meteorological applications and whether it has significant value.

b. Ratios from the 2x2 table as the basis of scoring

Another approach to developing scores from the three elements is to consider ratios developed of terms from the row and the column of Table 1. Specifically, many researchers have looked at $a/(a+c)$ and $a/(a+b)$ and their complements, $c/(a+c)$ and $b/(a+b)$, creating
relationships between those ratios. Again, a plethora of names have been applied to those ratios.

In meteorology, the most common name for the $a/(a+c)$ is the Probability of Detection (POD), from Donaldson et al. (1975 a, b). Prefigurance (Brier and Allen 1951) also has been used, but POD dominates in modern literature. Its origins in meteorology appear to come from House (1960), who was studying the problem of what density of observational stations was needed to observe different atmospheric features and estimated the probability of detecting a feature of interest given station spacing. It is unclear what impact House had on the choice by Donaldson et al.

In the machine learning community, Recall is the most common descriptor of $a/(a+c)$. It comes out of information retrieval and was first used by Kent et al. (1955) in the context of literature searching with the name “recall factor.” They wrote “This fraction, which we shall term the “recall factor,” measures the proportion of pertinent documents to which the information retrieval system directed attention when a given search was conducted.” In effect, its name comes from the consideration of how many of the appropriate documents are recalled from the database. Sensitivity is another common name for this expression, coming out of the medical diagnostic community (Yerushalmy 1947), deriving from measuring how sensitive a diagnostic test was in responding to a particular condition. Dice (1948) simply called both $a/(a+b)$ and $a/(a+c)$ “association indices” with notations added to indicate which one was being referred to.

Although Post-agreement (US Army Air Forces 1944) was suggested in meteorology for the term $a/(a+b)$, it has not been widely used. Neither has the Probability of Hits, suggested by Doswell et al. (1990). At this time, the most common term is the Success Ratio (SR) (Schaefer 1990). For reasons that aren’t particularly clear, meteorology has tended to focus

---

3 Brier and Allen is the standard reference for this term, but they cite the unpublished United States Army Air Force (USAAF) (1944) technical memo as the source. It is likely, based on the terminology used in the abstract of USAAF, that one or both of them are responsible for USAAF.
on the complement of the SR, the fraction of forecasts of an event that did not have the event occur. \( b \) is often labelled as “false alarms”, so that the \( b/(a+b) \) term has been called the False Alarm Ratio (Donaldson et al. 1975) and the False Alarm Rate (Olson 1965), among others. Given that the term \( b/(b+d) \) has also been called False Alarm Rate, this has caused great confusion. Barnes et al. (2007) give an outstanding review of the conflicting usage.

As a companion to Recall, Kent et al. (1955) named \( a/(a+b) \) the Pertinency Ratio. In the context of information retrieval, the origin of this term seems clear, giving the fraction of retrieved pieces of information that are pertinent to the search request. The early 1960s provided a huge boost to the study of information retrieval, funded by the National Science Foundation, to help scientists find the most important references to speed research in the aftermath of the Soviet launch of Sputnik. One of the biggest recipients of that support was at the Cranfield College of Aeronautics in the UK. The leader of that project, Cyril Cleverdon, followed up on Kent et al., adopting the use of Recall, but preferring Relevancy over Pertinency (Cleverdon 1962). The other big project, Information Storage and Retrieval (ISR), led by Gerard Salton at Harvard, introduced the term Precision (Salton 1964) with no discussion of the reasons. Interestingly, in the same report that Salton used “Precision”, Rocchio (1964) used Relevancy, in keeping with Cleverdon, but added a parenthetical comment that it “was sometimes called Precision.” Eventually, at a closed-door meeting in Washington in December 1964, Cleverdon agreed to use Precision (Cleverdon 1965).

One of the biggest advances made by Cleverdon (1962) was the notion of plotting Recall and Relevancy on a two-dimensional plot. Although Kent et al. (1955) had plotted distributions of the individual terms, Cleverdon was the first to plot them jointly. It could be done for a single 2x2 table or for a series of tables where the events were the same, but different thresholds were used, in the original case, by the addition of search terms to make the search more restrictive. That addition led to an apparent trade-off between Recall and

\[ \frac{b}{b+d} \]

4 Multiple reports from both the Cleverdon and Salton groups are available online, along with a number of other documents from the history of information retrieval, from the Special Interest Group on Information Retrieval at https://www.sigir.org/resources/museum/.

File generated with AMS Word template 2.0
Relevancy, although a debate about whether that trade-off was real or the result of the way Cleverdon had done the analysis went on for some time, despite a connection that could be made to the model of Tanner and Swets (1954) that supported the trade-off as real. For clarity, we have replotted Cleverdon’s original data (Figure 2, his Tables 7.5 and 7.6). It is of particular interest, that in the first plot of the two quantities, recall was plotted on the vertical axis in the same convention as used in the performance diagram as used in the performance diagram of Roebber (2009). This also means that the vertical axis is the same as used in the receiver operating characteristics (ROC) plots developed by Peterson and Birdsall (1953). Swets (1963) uses that convention in a discussion of what the horizontal axis on a plot should be (Relevancy or $b/(b+d)$ that Tanner and Swets (1954) had advocated, or, simply, $b$ as used by Swanson in a 1962 conference presentation).

![Figure 2: Reproduction of Cleverdon (1962) Table 7.6. Legend is as used in Cleverdon where different “Relevance” lines are associated with different search efforts. The dashed lines are from Cleverdon.](image)

The flipping of the axes on this two-dimensional plot of Recall and what was referred to as Precision first appears in Salton (1964, Figure 12) with no reason given for the change. No discussion appears in the reports of the Cranfield and ISR groups as to why the choices are made, but it appears to be conventions used by the groups. Interestingly, Michael Keen writes

File generated with AMS Word template 2.0
sections in reports for both groups, implying he may have received support from each. Even though he always refers to the diagrams as Recall-Precision diagrams, when they are published in Cranfield reports, Recall is the vertical axis (e.g., Keen 1966) and when they are published in ISR reports, Precision is the vertical axis (e.g., Keen 1968).

The question of the “standard” orientation of the two axes was not resolved for a long time and it is not clear why Precision as the vertical axis became the default in the machine learning community. Outside of the two big groups’ choices (Cranfield and ISR), Lancaster (1968, 1979) and Sparck Jones (1972, 1979) were among notable people who preferred Recall on the vertical axis. Keen (1981) also used Recall on the vertical. Heine (1973) used Precision on the vertical axis and overlaid contours of the Marczewksi-Steinhaus distance, marking the first appearance of a combined measure on such a plot. van Rijsbergen (1974) and, later, van Rijsbergen and Croft (1975) also used Precision on the vertical axis.

Recently, de Elia (2022) proposed a model of forecast performance that looked at the trade-off between false alarms and missed events from the perspective of expected utility. The core idea came from consideration of a performance diagram and modelling $POD$ as a power of $(1-SR)$, so that

$$POD = (1 - SR)^r$$  \hspace{1cm} (4)

Using that framework, de Elia et al. (2023) have proposed a skill score,

$$Q = \frac{\ln(1-SR)}{\ln(POD)}$$  \hspace{1cm} (5)

which represents a trade-off between $POD$ and $SR$ such that the total losses associated with misclassifications are constant along the curves. Given its recent development, this score has not been widely used. A big unanswered question is how well the power law formulation in (4) applies to real forecasting or classification problems. The behavior of $Q$ is illustrated in Fig. 3 as a function of $SR$ for constant values of $v_g$, so that as $SR$ increases, $POD$ decreases. $POD$ and $SR$ cannot be less than $v_g$, by definition. As $SR$ or $POD$ approach $v_g$, $Q$ becomes very large $POD$. That means that $Q$ is very sensitive to small changes at the edge of the range of the values.
c. The Pythagorean Means scores

If one starts with the two quantities POD and SR (or Recall and Precision), combining the quantities in ways that differ from what might be considered a Venn diagram approach seem logical. It is often possible to use algebra to show that scores developed in the two approaches are mathematically related, but the paths taken to get there are very different. Scores have repeatedly been developed that are simple functions of the three Pythagorean Means.

Reviewing, the Arithmetic Mean (AM) for any two elements such as POD and SR is given by

\[ AM = \frac{x + y}{2}, \]  

the Geometric Mean (GM) is given by
\[ GM = \sqrt{xy}, \]  

(7)

and the Harmonic Mean (HM) by

\[ HM = \frac{2}{\frac{1}{x} + \frac{1}{y}}. \]  

(8)

The three means are equal if, for non-zero values of \( x \) and \( y \), \( x=y \). If they aren’t equal, then, in general, \( HM \leq GM \leq AM \) for a pair of \( x \) and \( y \). All three of these means have been the basis of scores based on combining \( POD \) and \( SR \), even though, in most cases, the authors have not acknowledged it. To illustrate the difference, we show \( AM, GM, \) and \( HM \) for curves passing through points where \( x=y \) on a performance diagram (Figure 4).

Figure 4: Curves of Pythagorean Mean scores equal to 0.25, 0.5, and 0.75 that pass through points where \( POD=SR \).
Kulczynski (1927) used the arithmetic mean of POD and SR as a measure to compare plant distributions in different parts of the Pieniny Mountains in Poland. McConnaughey (1964) developed a measure

\[ \text{McC} = (POD)(SR) - (1 - POD)(1 - SR) \]  

(9)

With some manipulation, it can be shown that this reduces to 2*(AM-0.5). In effect, it converts the AM from a scale of 0 to 1 to a measure that goes between -1 and 1 without providing additional information. Although the AM is available in many software packages, often under Kulczynski’s name, it doesn’t seem to have been used much in the meteorological literature.

The GM has also been found a number of times over the years in a variety of fields in a process that represents an excellent example of the curiosities of naming. It appears that the first use is from Thomson (1916), in the testing of psychological theories. It was cited by Driver and Kroeber (1932) in archaeology. Their application was similar to Jaccard’s, in that they were comparing the distribution of artifacts in an effort to measure similarity, or the lack thereof, of different cultural groups. Interestingly, one of the names that Brusco et al. (2021) give to the score is Driver and Kroeber, despite the fact that they reference Thomson. If any weight is given to the “reputation” of the authors in the adoption of names, it is curious to note that Thomson is regarded as one of the pioneers of intelligence research and who wrote his paper as part of a debate with Spearman, of Spearman’s rank correlation fame.

Another field which has given the GM a name that has passed to other fields has been marine biology. A common name for the GM is the Ochiai coefficient (Ochiai 1957). Ochiai credits Otuka (1936), but via an intermediate publication by Hamai (1955). To add to the

5 We use POD and SR in the discussion of the Pythagorean means instead of Recall and Precision, or any other pairs of names for the same quantities for simplicity.

6 The name is sometimes spelled Otsuka.

File generated with AMS Word template 2.0
naming confusion, Howarth (2017) credits Otuka, but attributes the score to another researcher with the same family name.

Doolittle (1885), in his response to Finley (1884), used the product of POD and SR, which is obviously the square of GM, in developing his inference ratio, which is intended to measure the skill of a forecaster by adjusting for the number of forecasts that would be correct by chance. In his work on the distribution of mollusks in the Miocene, Sorgenfrei (1959) also used $GM^2$. His problem was similar to that of Jaccard, in that he was looking at the probability of species occurring in each of two different areas and simply multiplying the two probabilities. Wagner (1993) referred to this quantity as the “unbiased hit rate” in a review paper about evaluation of behavioral psychology experiments. Armistead (2013) brought the metric back into meteorology, proposing it as a method for evaluating categorical forecasts. Armistead references Doolittle’s inference ratio without mentioning that Doolittle had used Wagner’s unbiased hit rate in its development, although in a later paper (Armistead 2016), he refers to it as Doolittle’s unattributed joint probability measure.

The third Pythagorean mean, $HM$, is widely used. With some algebraic manipulation, it can be shown that $v_g$ is related to $HM$ by

$$v_g = \frac{HM}{2-HM} \quad (10)$$

Scores that are identical to $HM$ have been used in many contexts and are known by a variety of names. In ecological studies, Gleason (1920) used it without naming it, Sørensen (1945) referred to it as the quotient of similarity and Dice (1948) referred to it as the coincidence index. Neither of the names Sørensen or Dice used seem to have been used in references since then, with the score being called the Sørensen index, Dice’s coefficient, or the Sørensen-Dice index. An identical score was presented at a conference by van Rijsbergen and is often referred to as the $F$-score or $F_1$-score in many places in the machine learning community. Sasaki (2007) indicates that the origin of the name was apparently accidental, being confused with another $F$ function that van Rijsbergen (1975) introduced as a “combination” function in his derivation of a measure of effectiveness, $E$.

van der Maarel (1969) created a score that begins with $HM$ as a measure of similarity, but then includes an additional term that expresses an estimate of dissimilarity. In terms of the 2x2 table, van der Maarel’s index can be written as

$$VDM = \frac{(2a-b-c)}{(2a+b+c)} = HM - \left(\frac{b+c}{2a+b+c}\right) \quad (11)$$
The use of this score has been limited.

d. Scores that weight error terms differently

Any score that can be expressed as a simple function that weights the components of a Pythagorean mean equally (including \( v_g \)) implicitly values errors of false alarms or missed events equally. For many users, this is an unrealistic assumption. For instance, it is highly likely that failure to diagnose a disease with a high fatality rate is a greater threat to health than treating for the disease when it isn’t present. In that case, a patient would be more interested in high POD than high SR. There are two popular scores that provide weights. Although in an important selection of weights discussed below, they can be shown to be algebraically related to each other, as before, they start from different perspectives.

The first of these scores comes from van Rijsbergen (1975) and is usually given the notation, \( F_\beta \), even though, as noted before, van Rijsbergen never used that notation. It is defined as

\[
F_\beta = (1 + \beta^2) \frac{(POD)(SR)}{\beta^2(SR) + (POD)} \tag{12}
\]

van Rijsbergen’s goal was to measure the “effectiveness” of a system where a user attaches \( \beta \) times as importance to POD as to SR and the derivation comes from determining the ratio of trade-offs between the two that the hypothetical user is willing to take. It can be interpreted as a variant of HM, with the two terms being weighted. A key part of the derivation is the definition of the effectiveness, \( E \), in terms of \( \alpha \),

\[
E = 1 - \left( \frac{\alpha}{SR} + \frac{(1-\alpha)}{POD} \right)^{-1} \tag{13}
\]

\( F_\beta = 1-E \) with

\[
\alpha = (1 + \beta^2)^{-1}. \tag{14}
\]

In terms of the three elements of the 2x2 table, (10) can be rewritten as
\[ F_\beta = \left(1 + \beta^2\right) \frac{a}{(1 + \beta^2)a + b + \beta^2c} \quad (15) \]

Tversky (1977) approached the problem of weighting by putting weights on the error terms, rather than ratios, in the 2x2 table in an effort to develop a score to compare objects or stimuli that contain (or don’t contain) common features. The attractiveness of Tversky’s approach is that it allows users who have knowledge of the relative costs of errors to have a measure when they may not be able to express it in terms of POD and SR. In general, his index (or coefficient) can be written as

\[ T = \frac{a}{a + \gamma b + \delta c} \quad (16) \]

For specific values of the coefficients on \( b \) and \( c \), \( T \) becomes other well-known scores. For instance, if \( \gamma = \delta = 1 \), \( T = v_g \). If \( \gamma = \delta = 0.5 \), it becomes \( T = HM \). The situation where \( \gamma + \delta = 1 \) is particularly interesting. In that case,

\[ T_\gamma = \frac{a}{a + \gamma b + (1-\gamma)c} \quad (17) \]

If we divide the numerator and denominator by \( \gamma \), this becomes

\[ T_\gamma = \frac{\frac{1}{\gamma}a}{\frac{1}{\gamma}a + \frac{1}{\gamma}b + \frac{1}{(1-\gamma)c} \gamma} \quad (18) \]

Eqs. (15) and (17) are similar and \( T_\gamma = F_\beta \) if \( \gamma = \alpha \) from Eq. (14). Thus, a value of \( \beta = 2 \), which implies \( POD \) is twice as important as \( SR \) is equivalent to \( \alpha = \gamma = 0.2 \), which implies that missed events (c) are weighted four times as much as false alarms (b). Assuming that there are no costs associated with correct forecasts, the \( \alpha = \gamma \) value is equivalent to the decision threshold for taking action for a user whose decision problem can be described by the Misclassification Cost Ratio discussed by Roebber and Bosart (1996) and Wandishin and Brooks (2002). Kumler-Bonfanti et al. (2021) use (17) with \( \gamma = 0.3 \), equivalent to \( \beta = \sqrt{7/3} \approx 1.53 \). In general, the impacts of weighting of the \( POD \) and \( SR \), via \( \beta \), or \( b \) and \( c \) (via \( \gamma \)) can be seen in Figure 5.
Figure 5: Same as Fig. 3 except for $F_\beta=0.5, 1.0, 2.0$ ($T_\gamma=0.8, 0.5, 0.2$, respectively). The $F_\beta=1.0$ curves for 0.25 and 0.75 are not shown since they are on Fig. 3.

d. Impacts of score selection on apparent forecast performance

Brooks and Correia (2018) looked at long-term performance metrics for National Weather Service tornado warnings. One of the points of interest was the change in performance that took place in 2012-2013 as a result of an apparent increase in emphasis on false alarms (equivalent to lowering $b$ in (15) or increasing $\gamma$ in (17).) $POD$ decreased and $SR$ increased in that interval compared to the previous years. If we wish to consider some combination of the two scores to estimate an overall impact of the change, the impression we get depends on what combination is chosen. Time series of the results from Brooks and
Correia, updated through 2022, of the three Pythagorean means and $F_\beta$ for $\beta=0.5$ and 2 provide a range of impressions (Figure 6). (Note that, as seen before, $HM$ is related directly to $v_s$.)

![Scores for NWS Tornado Warnings (1986-2022)](image)

**Figure 6**: Tornado warning performance by year (updated from Brooks and Correia (2018)), as estimated by the Pythagorean means and $F_\beta=0.5$, 2.0 (equivalent to $T_\gamma=0.8$, 0.2, respectively).

$HM$ shows less of an impact on apparent performance in 2012-2013 than the other Pythagorean means and, in fact, shows increases back to the highest values ever seen by the late 2010s/early 2020s. $GM$ and $AM$ show large decreases with the increased emphasis on false alarms and the measure hasn’t returned to the previous values yet. For the weighted scores, the one that values $POD$ more highly shows large decreases after 2011 while the score that values $POD$ less shows continuous increases through to the current. We offer no judgment as to which of the scores is most “correct.” Our point is that the choice of performance measure, even for relatively simple evaluation exercises, can have a huge impact.

File generated with AMS Word template 2.0
on how one judges performance. Optimizing to one metric may lead to much poorer performance by other metrics. This is likely to be particularly important in cases where the costs of different kinds of errors are very different.

Often, there is a disconnect between the loss function and verification metrics used for training and evaluating a machine learning model, respectively. In recent work by Lagerquist and Ebert-Uphoff (2022), researchers explored using verification metrics like fraction skill score (Roberts and Lean 2008) and CSI as loss functions for training neural networks. One result was that using CSI as the loss function resulted in an overprediction bias, which is unsurprising given that CSI is a biased metric (i.e., one can maximize CSI by either overpredicting or underpredicting, known as hedging). Loss functions often require mathematical properties like continuity, differentiability, convexity, etc., while verification metrics are often more interpretable. A machine learning practitioner should carefully select a loss function appropriate for the decision task, and more work is required to determine which loss functions are appropriate for certain tasks.

4. Concluding thoughts

First, we strongly recommend that authors show the contingency table they are using and where the basic terms they are using come from using that table. There are enough options available from the literature that it is very easy to confuse even experienced users. We recommend that interdisciplinary researchers primarily use terms that are commonly used in the field where they are publishing, such as $v_g$ or “CSI” when publishing in a meteorological journal, to assist the primary audience in quickly understanding. If there is another term for a metric that is commonly used in the other discipline from which the researchers come, and may be their standard choice for the term, it would be extremely useful to include that term in a parenthetical remark or footnote. Marzban (1998) is an example of this when he mentioned that a “contingency table” is sometimes referred to as a “confusion matrix.” This process helps the person in a discipline who is reading papers in a related, but not necessarily core to them, area to learn the language of that other discipline. It also introduces that term to them, so that they can find what this unfamiliar discipline has done in terms of analysis of their problem and they may learn new techniques from that discipline. As an example,

File generated with AMS Word template 2.0
biogeography, in its study of similarity and differences in the distribution of organisms in
different situations, has produced a wide body of literature on classification metrics.

Much of what we have discussed has involved the rediscovery of metrics. As part of that process, we have seen the fruits of what might be considered less-than-complete literature reviews. Although we believe we have found many of the most important early works, if nothing else, we have learned to not have hubris to think we have been complete. We apologize to anyone who may have a favorite “old paper” that found one of these metrics that we missed.
Acknowledgments.

This material is based upon work supported by the Joint Technology Transfer Initiative Program within the NOAA/OAR Weather Program Office under Award No. NA22OAR4590171. Funding was also provided by NOAA/Office of Oceanic and Atmospheric Research under the NOAA-University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce. Allan Murphy inspired interest in the history of scoring metrics and discussions between the lead author and him in the 1990s were useful in clarifying credit for discovery or rediscovery, even if Allan might not agree with all we’ve done here. More recently, exposure to the work done by AI/ML researchers seeking to improve forecasters and an awareness of the different language they use has highlighted opportunities for growth in multidisciplinary work. We thank Bob Glahn for providing a copy of Palmer and Allen (1949), which is in process of being cataloged at the National Centers for Environmental Information after being lost for decades. Since that provision, Bob has unfortunately passed away and we dedicate this paper to his memory, as well as that of Allan Murphy. We thank Naoko Sakaeda for translating a portion of the Japanese papers. We thank Amanda Schilling and Tracy Chapman, scientific librarians at the National Weather Center, for their invaluable assistance in tracking down some of the more obscure references, particularly Kulczynski (1927), Tanimoto (1958), and McConnaughey (1964). We hope that they enjoyed the hunt as much as we did.

Data Availability Statement.

The data and Python code used to generate the figures can be found at https://github.com/monte-flora/verification_diagrams. To encourage adoption and exploration of other 2 x 2 contingency table metrics, this package contains all 71 scores from Brusco et al. (2021).
REFERENCES


Accepted for publication in Artificial Intelligence for the Earth Systems. DOI 10.1175/AIES-D-23-0104.1.


File generated with AMS Word template 2.0


Venn, J., 1880: I. On the diagrammatic and mechanical representation of propositions and reasonings. Philosophical Magazine Series 5. 10, 1-18, DOI: 10.1080/14786448008626877

File generated with AMS Word template 2.0


