The Finley Affair: A Signal Event in the History of Forecast Verification

ALLAN H. MURPHY

Prediction and Evaluation Systems, Corvallis, Oregon

(Manuscript received 23 May 1995, in final form 31 August 1995)

ABSTRACT

In 1884 a paper by J. P. Finley appeared in the *American Meteorological Journal* describing the results of an experimental tornado forecasting program in the central and eastern United States. Finley's paper reported ''percentages of verifications'' exceeding 95%, where this index of performance was defined as the percentage of correct tornado/no-tornado forecasts. Within six months, three papers had appeared that identified deficiencies in Finley's method of verification and/or proposed alternative measures of forecasting performance in the context of this 2×2 verification problem. During the period from 1885 to 1893, several other authors in the United States and Europe, in most cases stimulated either by Finley's paper or by the three early responses, made noteworthy contributions to methods-oriented and practices-oriented discussions of issues related to forecast verification in general and verification of tornado forecasts in particular.

The burst of verification-related activities during the period 1884–1893 is referred to here as the ''Finley affair.'' It marked the beginning of substantive conceptual and methodological developments and discussions in the important subdiscipline of forecast verification. This paper describes the events that constitute the Finley affair in some detail and attempts to place this affair in proper historical context from the perspective of the mid-1990s. Whatever their individual strengths and weaknesses, the measures introduced during the period from 1884 to 1893 have withstood important tests of time—for example, these measures have been rediscovered on one or more occasions and they are still widely used today (generally under names assigned since 1900). Moreover, many of the issues vis-à-vis forecast verification that were first raised during the Finley affair remain issues of considerable importance more than 100 years later.

1. Introduction

Operational weather forecasting based on real-time synoptic charts was initiated in the United States and in several countries in western Europe during the period 1850–1870, in conjunction with the creation of regional and national weather services (Hughes 1994; Whitnah 1961). Perhaps not surprisingly, questions—and, in some cases, controversies—soon arose about the quality of the forecasts produced by these services. Burton (1986), for example, contains a rather detailed discussion of various issues related to the accuracy of storm warnings produced in the early 1860s by the United Kingdom's Meteorological Department under the direction of Admiral Fitzroy. Nevertheless, the practice of forecast verification—and, in particular, concepts and methods underlying this practice—appear to have received relatively little attention prior to 1880.

In 1884, a Sergeant Finley of the U.S. Army Signal Corps published some results of an experimental tornado forecasting program (Finley 1884). Finley's results indicated that his tornado/no-tornado forecasts generally had achieved levels of accuracy—measured in terms of the percentage of correct forecasts—exceeding 95%. These results, and the underlying verification problem, attracted the immediate attention of individuals both inside *and* outside the meteorological community. Several of these individuals proposed alternative measures of forecasting performance and then used these measures to demonstrate (inter alia) that percentage correct is an inappropriate measure of performance in this context. These and other events, which constitute the so-called Finley affair, occurred during the decade from 1884 to 1893 and marked the first substantive discussions of basic issues underlying the development of verification methods and measures.

The primary purpose of this paper is to describe the Finley affair in some detail, with particular emphasis on the measures proposed to verify these tornado/no-tornado forecasts and the basic issues that arose in connection with the formulation and application of these measures. It is important to note that the verification measures proposed in connection with this affair are still widely used today and in some cases have been rediscovered one or more times in the intervening 100 years! Moreover, many of the conceptual and methodological issues that were raised during the Finley affair are still of interest and concern today.

Corresponding author address: Dr. Allan H. Murphy, Prediction and Evaluation Systems, 3115 NW McKinley Drive, Corvallis, OR 97330-1139.
E-mail: murphya@ucs.orst.edu

/ams v5077 0196 Mp 3 Friday Feb 23 07:21 PM AMS: Forecasting (March 96) 0196

Section 2 summarizes the contents of Finley's original paper. His results are presented, first in a form almost identical to that employed by Finley himself and then in a more general framework. Three early methodologically oriented responses to Finley's paper, responses that include formulations of alternative measures of forecasting performance, are reviewed in section 3. The contents of several papers constituting the so-called aftermath of the Finley affair are described in section 4. These papers cover a relatively wide variety of topics, ranging from substantive discussions of various methodological issues to practical suggestions related to the formulation and evaluation of tornado forecasts. Section 5 consists primarily of a discussion and interpretation of some specific features of the Finley affair, mainly from the perspective of the mid-1990s. This section includes a discussion of the similarities and differences among the verification measures proposed during the Finley affair, an overview of subsequent rediscoveries of these measures, and a comparative discussion of several basic concepts and issues in forecast verification as viewed in the late nineteenth and late twentieth centuries. Section 6 consists of a brief summary and some concluding remarks.

2. J. P. Finley and his tornado forecasts

a. Finley (1884)

J. P. Finley of the U.S. Army Signal Corps published a paper in the *American Meteorological Journal* in July 1884 (Finley 1884) in which he summarized some results of an experimental tornado forecasting program initiated earlier that same year. [For details of Finley's work and life, see Galway (1985).] In this experimental program tornado predictions were made for each of 18 districts in the central and eastern United States during March, April, and May. The forecasts in question were produced twice a day and generally were valid for 8-h periods beginning at 0700 and 1500 LT. During May, predictions were made for the 16-h period beginning at 0700 LT and the 8-h period beginning at 1500 LT.

The tornado forecasts were based on surface weather maps produced each morning and afternoon at 0700 and 1500 LT, respectively. These predictions specified whether conditions in a district were ''favorable for tornadoes'' or ''unfavorable for tornadoes.'' In verifying the forecasts, Finley considered the predictions as forecasts of ''tornadoes'' and ''no tornadoes,'' respectively, and we follow this practice here.

Some overall results of this tornado forecasting program for 1884, stratified by district and/or month, are summarized in Table 1. This table is identical in almost all respects to Table No. 1 in Finley's paper (p. 86). However, we have added a column on the right to the original table; this column contains the fraction (percentage) of correct forecasts by month and/or valid period. Finley, in his Table No. 2 (not reproduced here), reported this percentage, which he referred to as the ''percentage of verification,'' by district and month. Table No. 2 also contained the overall percentage correct by month, evidently computed as the average of the district percentages.

It is clear from Table 1, and from Finley's Table No. 2 (p. 87), that the fraction (percentage) of correct tornado/no-tornado forecasts is very high indeed. The text of the paper contains no specific comments or claims about the results; the author evidently was satisfied to let the results speak for themselves. To facilitate discussion of Finley's method of verification and the results, as well as subsequent reactions to his paper, these results are presented in the form of a contingency table in section 2b.

In concluding his short paper, Finley made several points that bear directly on his approach to the problem of verifying tornado forecasts as well as the interpretation of the results (p. 88). His major points included 1) predictions of conditions favorable and unfavorable to tornadoes *both* required careful study; 2) tornado/no-tornado forecasts were considered to have verified with the appearance/nonappearance of one or more funnel-shaped clouds; 3) predictions of tornadoes were not considered to have verified unless the paths of the funnel-shaped clouds were clearly within the district during the valid period of the forecasts; and 4) district tornado reports, based on over 800 reporting stations, were inadequate to cover the territory encompassed by the districts and (in any case) were incomplete at the time the paper was written.

b. Finley's results in a general framework

To facilitate subsequent discussions of verification methods, measures, and results, it is useful at this point to introduce general notation to identify the various joint and marginal frequencies that arise in the verification problems of interest here. This notation is presented in Table 2, in which the generic symbol ''n'' is used to denote these frequencies. It will be convenient here to refer to such a table as a 2×2 ''contingency table,'' where the designation ''2×2'' refers to two possible forecasts and two possible observations. In this general description of the 2×2 problem the weather events of interest are identified as event 1 and event 2. Thus, n_{11} represents the joint frequency with which event 1 is both forecast and observed, etc.; $n_{1\cdot}$ is the marginal frequency of forecasts of event 1, etc.; $n_{\cdot 1}$ is the marginal frequency of observations of event 1, etc.; and $n_{\cdot\cdot}$ is the total frequency or sample size.

The results of Finley's experimental tornado forecasting program, summarized in Table 1, are pooled for all months and valid periods and presented in the form of a 2×2 contingency table in Table 3. (Finley's paper contains no such contingency tables.) This table depicts the joint frequencies of the four possible com-

TABLE 1. Finley's tornado predictions and verifications [after Table No. 1, Finley (1884, p. 86)].

Month	Valid period (hours)	Favorable to tornadoes		Unfavorable to tornadoes		Total		Fraction (percentage) correct
		Number of forecasts	Number verified	Number of forecasts	Number verified	Number of forecasts	Number verified	
March	8	43	6	728	721	774	727	0.943 (94.3)
April	8	25	11	909	906	934	917	0.982 (98.2)
May	8	10	8	548	542	558	550	0.986 (98.6)
May	16	22	3	518	511	540	514	0.952 (95.2)

Note: An apparent error in Finley's original Table No. 1 has been corected. Finley reported the total number of forecasts for the 16-h valid period in May as 549 instead of 540.

binations of tornado forecasts and observations, the marginal frequencies of the tornado/no-tornado forecasts, the marginal frequencies of the tornado/no-tornado observations, and the total frequency (or sample size). Thus, tornadoes were both forecast and observed on 28 occasions, etc.; tornadoes were forecast on 100 occasions, etc.; tornadoes were observed on 51 occasions, etc.; and the sample size was 2803.

It is now possible to use the notation introduced here to present a general expression for the verification measure used by Finley to assess tornado forecasting performance. The basic measure, denoted here by i_F (''i'' for ''index'' and ''F'' for ''Finley''), can be defined in terms of joint and overall frequencies of forecasts and/or observations as follows:

$$i_F = (n_{11} + n_{22})/n_{..} \quad (1)$$

This measure determines the fraction of correct forecasts, and it is usually referred to simply as the ''fraction correct.'' Note that the range of values of i_F is the closed unit interval [0, 1], with 1 (0) representing the best (worst) possible result.

The overall value of i_F for Finley's tornado forecasts, as presented in Table 3, is 0.966 [=(28 + 2680)/2803]. Numerical values of measures such as i_F are frequently presented in percentage terms (e.g., Finley follows this practice). If we denote the percentage of correct forecasts—or ''percent correct''—by $i_F(\%)$, then $i_F(\%) = 100 i_F$. The overall value of $i_F(\%)$ for Finley's tornado forecasts is 96.6%.

3. The early responses

This section reviews three papers that were published or presented within six months of the appearance of Finley's paper in July 1884. In each case, the author formulates one or more measures of forecasting performance and uses the measure(s) to assess the quality of Finley's tornado forecasts. In addition to the measures themselves, these papers are notable for their distinct approaches to and insights into the problem of forecast verification.

a. Gilbert (1884)

Only two months after the appearance of Finley (1884), a paper by G. K. Gilbert (Gilbert 1884) was published in the same journal. In the opening paragraphs of his paper, Gilbert referred to a ''serious fallacy'' in Finley's paper, which he identified as the assumption that forecasts of tornadoes and no tornadoes were of equal difficulty in a situation in which tornado occurrences were very rare and tornado nonoccurrences were very frequent. He then went on to point out that predictions of no tornadoes on all 2803 occasions, a strategy that did not require ''any study of the meteorological record (p. 166),'' would yield a numerical value of the measure $i_F(\%)$ equal to 98.2% [=100(0 + 2752)/2803]!

Gilbert then proceeded to formulate two measures of forecasting performance as alternatives to the measure i_F. The first measure, which he referred to as the ''ratio of verification'' (v), was defined as follows:

TABLE 2. General description of 2 × 2 verification problem in terms of a contingency table consisting of joint and marginal frequencies.

Forecasts	Observations: Event 1	Observations: Event 2	
Event 1	n_{11}	n_{12}	$n_{1.}$
Event 2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

TABLE 3. Representation of pooled results of Finley's experimental tornado forecasting program in terms of 2 × 2 contingency table consisting of joint and marginal frequencies.

Forecasts	Observations: Tornado	Observations: No tornado	
Tornado	28	72	100
No tornado	23	2680	2703
	51	2752	2803

$$v = \frac{n_{11}}{(n_{1\cdot} + n_{\cdot 1} - n_{11})} = \frac{n_{11}}{(n_{11} + n_{12} + n_{21})}. \quad (2)$$

This measure is the ratio of the joint frequency of forecasts *and* observations of the event of interest (e.g., tornadoes) to the total frequency of forecasts *or* observations of this event. Gilbert noted that the values of v ranged from 0 when $n_{11} = 0$ to 1 when $n_{11} = n_{1\cdot} = n_{\cdot 1}$ (i.e., $0 \leqslant v \leqslant 1$).

The author was quick to point out that v ''falls far short of a just measure of success in scientific forecasting, for . . . this ratio may be larger or smaller according as the phenomena foretold are normally frequent or rare'' (p. 168). To illustrate this point, Gilbert computed v for Finley's tornado (T) *and* no-tornado (NT) forecasts separately and showed that v(T) = 0.228 [=28/(100 + 51 − 28)] and v(NT) = 0.966 [=2680/(2703 + 2752 − 2680)]. In the case of no-tornado forecasts, v(NT) is computed from (2) by interchanging the labels on tornado and no-tornado events.

In formulating the second measure, which was designed to take account of relative difficulty in assessing forecasting performance, Gilbert introduced the notion that ''a certain number of predictions . . . would fortuitously coincide with occurrences'' (p. 168). He estimated the number of such predictions on the basis of the assumption of ''random prognostication'' (equivalent to the assumption of independence of forecasts and observations). Under this assumption, Gilbert determined that the expected number of correct predictions of event 1 would be $(n_{1\cdot}n_{\cdot 1})/n_{\cdot\cdot}$. He then defined the second measure, to which he gave the name ''ratio of success'' (i_G), as

$$i_G = \frac{\left[n_{11} - \frac{(n_{1\cdot}n_{\cdot 1})}{n_{\cdot\cdot}}\right]}{\left[n_{1\cdot} + n_{\cdot 1} - n_{11} - \frac{(n_{1\cdot}n_{\cdot 1})}{n_{\cdot\cdot}}\right]}. \quad (3)$$

In effect, the measure i_G is an ''adjusted'' form of the measure v, where the adjustment—included in both numerator and denominator—consists of subtracting the number of correct forecasts that a random prognosticator could be expected to obtain. Gilbert refers to the difference between n_{11} and $(n_{1\cdot}n_{\cdot 1})/n_{\cdot\cdot}$ in the numerator of (3) as a ''product of . . . skill in inference'' (p. 168). Whether or not the author's use of the term ''skill'' in this context constitutes its first appearance in the verification literature, Gilbert appears to be using this term to characterize the basic inferential capabilities of forecasters. He uses the term ''success'' to describe the results of these capabilities, as reflected by the numerical value of i_G (or its numerator).

Gilbert then subjected the measure i_G to a series of tests. These tests included examining the values of i_G under various conditions, as well as assessing the sensitivity of i_G to changes in the values of the underlying joint, marginal, and total frequencies. He showed that i_G cannot exceed unity (which occurs when $n_{11} = n_{1\cdot} = n_{\cdot 1}$), and he identified the conditions under which i_G is either zero ($n_{11}n_{\cdot\cdot} = n_{1\cdot}n_{\cdot 1}$) or negative ($n_{11}n_{\cdot\cdot} < n_{1\cdot}n_{\cdot 1}$). With regard to sensitivity, Gilbert demonstrated that i_G ''varies directly and very rapidly with the number of coincidences between prediction and occurrence,'' ''varies inversely and less rapidly with the number of occurrences and with the number of predictions,'' and ''varies directly, and still more slowly, with the total number of cases under consideration'' (p. 170).

Finally, the author showed (by inversion) that the expression for i_G in (3) applies to the nonoccurrence as well as to the occurrence of the event of interest. He then illustrated this result by computing the values of i_G(T) and i_G(NT) for Finley's tornado forecasts. For these data, i_G(T) = i_G(NT) = 0.216.

In conclusion, Gilbert suggested that the measure i_G could be used to compare success in predicting tornadoes with success in predicting other events defined in terms of ''simple occurrence'' (i.e., binary outcomes). He also pointed out that this measure is applicable only in situations involving ''two alternatives'' (i.e., two events).

b. Peirce (1884)

In November 1884, C. S. Peirce, a well-known logician and philosopher, published a short paper in *Science* (Peirce 1884), in which he proposed an alternative to Gilbert's and Finley's measures of forecasting success. Peirce's approach was based on two interesting principles. First, he assumed that any two forecasting methods should be considered equally successful (specifically, ''equal approximations to complete knowledge'') if they yielded, in the long-run, the same numerical values of the joint frequencies n_{11}, n_{12}, n_{21}, and n_{22}.

The second principle consisted of assuming 1) that a known fraction of the predictions were made by an infallible forecasting method, 2) that the remaining fraction of the predictions were made by a random forecasting method, and 3) that the fraction of predictions made by the infallible method was a suitable measure of success in forecasting. On the basis of these principles, Peirce was able to write down a set of equations whose solution led to a measure of forecasting performance denoted here by i_P (Peirce simply used the notation ''i''), where

$$i_P = (n_{11}/n_{\cdot 1}) - (n_{12}/n_{\cdot 2}) = \frac{(n_{11}n_{22} - n_{12}n_{21})}{(n_{\cdot 1}n_{\cdot 2})}. \quad (4)$$

Although not explicitly shown by Peirce, the numerical values of i_P range from +1 when $n_{12} = n_{21} = 0$ to −1 when $n_{11} = n_{22} = 0$.

In his very brief discussion of the measure i_P, Peirce noted that it had the same numerator as i_G but a different

denominator (see section 5a). Applying i_P to Finley's tornado data, he obtained a value of 0.523 (recall that $i_F = 0.966$ and $i_G = 0.216$).

Peirce concluded by indicating that he had a solution to the problem of extending this approach to situations involving more than two alternatives. He also presented a simple expression for the average profit per prediction in the two-alternative case, thereby quite possibly providing the first analytical treatment of the problem of assessing the economic value of weather forecasts.

c. Doolittle (1885a,b)

Early in 1885 M. H. Doolittle, a mathematician known for his method of solving systems of simultaneous linear equations, published a paper in the *Bulletin of the Philosophical Society of Washington* (Doolittle 1885a), in which he proposed still another measure of forecasting performance for this 2×2 problem. (The paper was communicated to the Philosophical Society of Washington in December 1884.) An abbreviated version of the same paper subsequently appeared in the *American Meteorological Journal* (Doolittle 1885b). Reference here is made to the original full-length paper.

As in the case of Gilbert, Doolittle was concerned with giving the forecasting method or forecaster credit only for the coincidences of forecasts and observations over and above those that could be expected on the basis of chance alone. His estimate of the latter was identical to that of Gilbert (i.e., $n_{1\cdot}n_{\cdot 1}/n_{\cdot\cdot}$).

First, the author argued that ''since success is proportional to each of two fractions $n_{11}/n_{1\cdot}$ and $n_{11}/n_{\cdot 1}$, it may be represented by their product $n_{11}^2/n_{1\cdot}n_{\cdot 1}$'' (p. 123). His measure of the ''degree of logical connection'' (between event and prediction) was then formulated by subtracting the number of random successes from each component in this product of fractions. If this measure is denoted by i_D, then

$$i_D = \frac{(n_{\cdot\cdot}n_{11} - n_{1\cdot}n_{\cdot 1})^2}{(n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2})} = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{(n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2})}. \quad (5)$$

The author showed that i_D ranges from 0 to 1 and passes all of the tests proposed by Gilbert. Moreover, he demonstrated that $i_D = 1$ in the perverse situation in which all forecasts are unsuccessful (i.e., $n_{11} = n_{22} = 0$).

Doolittle calculated i_D for Finley's tornado data (Table 3) and found that $i_D = 0.142$. He then argued that since $n_{11}^2/n_{1\cdot}n_{\cdot 1} = 0.154$, 92.3% [$=100(0.142/0.154)$] of Finley's success was due to skill and only 7.7% [$=100(0.012/0.154)$] was due to chance. In the case of the no-tornado forecasts, the corresponding percentages were shown to be 14.7% (skill) and 85.3% (chance).

In conclusion, Doolittle briefly discussed the issue of extending his approach to problems involving three or more events. He indicated that ''it seems clear to me that no single numerical expression can be a proper solution of such a problem'' (p. 126).

4. The aftermath

As defined here, the aftermath of the Finley affair consists of a group of papers that appeared in the 10-year period from 1884 to 1893. In most cases, these papers were motivated either by Finley's paper and/or by the three methods-oriented papers that were published shortly after Finley's paper (see section 3). These so-called aftermath papers are considered in chronological order.

a. Curtis (1887)

Curtis first used the verification measures formulated by Gilbert to verify the experimental tornado/no-tornado forecasts made (by Finley) during June 1885. Evaluating the forecasts on a district-by-district basis, Curtis obtained relatively low values for these measures; namely, $v = 0.21$ and $i_G = 0.14$. Even when adjacent districts were incorporated into both the forecasts and the observations (based on the premise that the original districts might be too small), the value of the measure v increased only to 0.40. These verification figures prompted the author to raise several questions concerning the interpretation of the results and the appropriateness of this (i.e., Finley's) approach to the tornado forecasting problem.

As a result of these (and other) calculations, Curtis concluded that the ''whole system of fixed districts is inappropriate for making advantageous predictions'' (p. 72). In particular, Curtis indicated that the prediction itself, as well as the method of verification, should take account of what was currently known about the district, or area, over which tornadoes might be expected to occur on any particular day. In his terms, ''since the anticipated area favorable for tornadoes is, in general, an undivided district, it follows that the prediction should likewise be made for a single district covering all the territory within which tornadoes are expected to occur'' (pp. 72–73). Thus, he concluded that ''a district, movable from day to day, must take the place of rigid, fixed districts, having no relation to each individual case'' (p. 73).

The author then discussed the possible shape and size of such areas. With regard to shape, he suggested rectangular or elliptical areas. In the case of a rectangular areas, he proposed a rectangle with dimensions of 400 miles $\times$ 600 miles. Curtis noted that a larger area would presumably contain more of the tornado occurrences but might be less useful ''if the ratio of people benefitted to the number alarmed is smaller'' (p. 73) than the ratio of the sizes of the areas.

In describing the manner in which such tornado predictions should be made, Curtis indicated that ''a card representing the assumed area of prediction will be placed on the weather map in such position as best to cover the locality where tornadoes are anticipated'' (p. 73). As a method of verification, he suggested com-

puting the ''relative number of tornadoes occurring within and without the marked area'' (p. 74).

Finally, Curtis pointed out that ''the use of a single movable district corresponds to what we know of the occurrence of tornadoes, and to the principles that obtain in their prediction'' (p. 74). Moreover, he expressed the belief that ''the percentage of verification obtained by predicting in the way described above will be sufficiently high to warrant the public announcement of tornado predictions with the other predictions of the Signal Service'' (p. 74).

b. Hazen (1887, 1892)

Hazen (1887) raised several objections to the method of verification—in particular, the measure v—proposed by Gilbert. Hazen's objections derived from the following considerations: 1) the distance between a tornado occurring outside a district and the boundary of the district was not considered; 2) all tornado occurrences may not have been reported; 3) difficulties existed in distinguishing between tornadoes and other types of destructive storms; 4) the actual number of tornadoes that occur within a district was not considered; 5) no account was taken of differences among the meteorological conditions that lead to tornado occurrences in different parts of the country; and 6) no distinction was made between districts with relatively high and relatively low frequencies of tornado occurrence. In summarizing his point of view on verification systems, Hazen indicated that ''we cannot apply rigid mathematical analysis to the questions, but must seek for a rational system which will best treat the prediction as worded and an occurrence so indefinite'' (p. 129).

To overcome at least some of these difficulties, Hazen proposed a method of verification involving weights that would take into account the distance of tornado occurrences from the center of a district. Applying this method to the tornado/no-tornado forecasts for June 1885 yielded a percentage of verification [$i_F(\%)$] of 49%.

Hazen then discussed other issues related to the verification of Finley's tornado forecasts, including the issue of giving credit for correct forecasts of no tornadoes. Hazen also stressed the difference between the probability of tornado occurrence on any day of the year and the probability of tornado occurrence on any of the 50 ''special days when they are very likely to occur'' (p. 131).

In conclusion, Hazen (1887) suggested that ''the division of the country into districts . . . is hardly wise'' (p. 131). Specifically, he argued that ''it would be more satisfactory to predict, in a region where at least 25 or 30 destructive storms and tornadoes occur each year, a central point or locus of destructive storms, giving boundaries, more or less definite, to the limit of destruction, and in verifying to give weights to storms occurring at distances of 50, 100, etc., miles from that locus'' (p. 131).

Hazen (1892) first discussed the issue of what (in his opinion) constituted a proper weather forecast. He ruled out forecasts in which little or no uncertainty existed, as well as forecasts in which uncertainty predominated (referred to as ''certainties'' and ''guesses,'' respectively). For Hazen, a proper forecast was evidently a forecast that correctly anticipated changes in basic meteorological conditions (e.g., the position of a high or low pressure area), since ''in the long run, one who foresees the changes best will make the best forecasts'' (p. 393).

With regard to verification of forecasts, Hazen's point of view can be summarized by his statement ''that to make a proper verification of a weather forecast, . . . , it should be done by an expert or one thoroughly acquainted with the average conditions and he should verify from the map on which the prediction is based and not from subsequent maps'' (p. 394). The author then discussed the comparison of forecasting performance across periods with different rain frequencies and across periods with different definitions of temperature categories. To determine whether forecasts have improved over time, Hazen suggested conducting an experiment to determine whether a forecaster who made forecasts during some earlier period could improve upon these forecasts, using the same maps but having access to current forecasting methods. In conclusion, he indicated that factors such as the requirement to formulate long-range (48 h!) forecasts as well as short-range forecasts and the time taken to produce a forecast should be taken into account in the verification process.

c. Doolittle (1888)

This paper consists of two quite distinct parts. In the first part the author developed the measure i_D (see section 2c) under the more general heading of ''association ratios.'' In the second part he presented a critique of Gilbert's response (Gilbert 1884) to Finley's paper (Finley 1884).

Doolittle's general development of i_D began with a description of what he referred to as ''indiscriminate association ratios'' and ''discriminate association ratios.'' The former measured the degree of correspondence between forecasts and observations that resulted from both general and special causes, whereas the latter measured the degree of correspondence that resulted only from special causes. In effect, the phrase ''general causes'' referred to that part of the overall relationship between forecasts and observations that could be attributed to chance alone. On the other hand, the phrase ''special causes'' referred to that part of this overall relationship that could not be attributed to chance (i.e., the part over and above that due to the chance relationship). To set the stage for his development of i_D—a discriminate association ratio—Doolittle offered the following description of the 2×2 verification problem:

> Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things (p. 85).

After describing the development of the measure i_D (see section 3c), Doolittle discussed some of its properties. These properties included the following: 1) i_D is not affected by interchanging the roles of the forecasts and observations, and 2) i_D reduces to the underlying indiscriminate measure $n_{11}/n_{1\cdot}n_{\cdot 1}$ as $n_{\cdot\cdot}$ approaches infinity.

In the second part of the paper (pp. 94–96), Doolittle reviewed and criticized Gilbert's criticism of Finley's paper. First, Doolittle used Gilbert's method of taking account of chance coincidences (see section 3a) to ''correct'' the verification measure—namely, i_F—used by Finley. (In this case, chance coincidences for both events were taken into account.) This process led to a new measure, denoted here by i_D^* (Doolittle simply used the notation ''i''), where

$$i_D^* = \frac{2(n_{11}n_{\cdot\cdot} - n_{1\cdot}n_{\cdot 1})}{(n_{1\cdot}n_{\cdot 2} + n_{2\cdot}n_{\cdot 1})}. \qquad (6)$$

For Finley's tornado forecasts, Doolittle noted that $i_D^* = 0.355$.

Doolittle pointed out that i_D^* passed all of the tests set forth by Gilbert (for i_G) but stated ''it is not maintained that it has any scientific value'' (p. 96). Unlike Gilbert, Doolittle suggested that the fallacy in Finley's method ''consists . . . in the supposition that any valuable result can be obtained by averaging the percentages of verification of heterogeneous classes of predictions'' (p. 96). The author concluded as follows: ''Mr. Finley correctly computed his indiscriminate percentage of verification, and thereby furnished a striking and, perhaps, much-needed illustration of the worthlessness of such computations. The elimination of hypothetical chance from such mixed percentages merely renders their worthlessness less apparent'' (p. 96).

d. Clayton (1889, 1891)

In Clayton (1889), the author first discussed the degree of specificity of weather forecasts and its impact on the success of forecasts and their economic value. He pointed out that since increased specificity generally enhances value but adversely affects success, some latitude is usually given in judging success. Clayton then examined some of the procedures used by the U.S. Signal Service and the German Seewarte in verifying their forecasts. Although he accepted the need for some latitude in these procedures, he argued that ''the amount of latitude to be allowed ought to be definitely stated so as to eliminate entirely the influence of personal judgment or bias in any direction'' (p. 212). In a similar vein, Clayton indicated that information related to the success of forecasts is meaningful only if the meteorological phenomena in question possess clear, unambiguous definitions and these definitions are made known to the public.

In citing the efforts of Gilbert and Doolittle in the United States and Köppen in Europe (see below) to develop better verification methods, Clayton identified four desirable properties that such methods should possess. 1) The ability to ascertain at which point weather forecasts cease to have value. 2) The ability to determine what part of their success can be attributed to chance. 3) The ability to measure success in proportion to the degree to which the forecasts depart from chance and approach the observations. 4) The capability to arrange the verification material in such a way that the weak points are indicated and the road to improvement is opened. In commenting briefly on the first of these desiderata, Clayton pointed out that since ''every one ought to be familiar with the climate of the place in which he lives, . . . weather forecasts would only become of value as they exceeded the per cent(age) of success which might be gained by stating well known, or easily ascertained climatic facts'' (p. 214). He also noted parenthetically that the German Seewarte was already reporting verification results in a manner consistent with the fourth desideratum.

Clayton then used a sample of forecasts and observations made at Blue Hill Observatory (Milton, Massachusetts) during the first three months of 1889 to illustrate ''a method of verification which seems to most fully embody the ideas expressed in this paper'' (p. 214). These forecasts, which were made every day (except Sunday) at 1330 LT and were valid for the 24-h period beginning at 2400 LT that day, described the weather conditions in half-day intervals. The predictions included the terms ''fair,'' ''cloudy,'' ''light rain,'' ''rain,'' and ''snow,'' each of which was defined in as clear and unambiguous a manner as possible.

TABLE 4. Clayton's verification method applied to forecasts made at Blue Hill Observatory [after Clayton (1889, p. 216)].

	Occurrences						
Predictions	F	C	LR	R	HR	S	
F	85	6	0	3	0	5	99
C	9	6	0	1	0	1	17
LR	1	1	0	0	0	0	2
R	7	6	4	13	1	2	33
S	1	0	0	0	0	2	3
	103	19	4	17	1	10	154

Key: F is fair, C is cloudy, LR is light rain, R is rain, HR is heavy rain, and S is snow.

The author's method of verifying these forecasts is illustrated in Table 4. This table, which is identical in almost all respects to the table presented by Clayton (p. 216), is a contingency table in which the joint and marginal frequencies of the forecasts and observations for this data sample are displayed. It should be noted that the term ''contingency table'' was not used by Clayton anywhere in his paper. [Note: This term was evidently introduced by K. Pearson in 1904 (David 1995).] In discussing the advantages of this ''arrangement'' (i.e., method), the author made the following points (p. 216). 1) ''It shows at a glance the directions in which the predictions have been most and least successful,'' and ''attention is thus called to the direction in which improvement may be made.'' 2) ''Such tables will show the relative frequency of each phenomena and thus prove of much value . . . in guiding future weather forecasting.'' 3) ''The arrangement is such as to admit of the application of mathematical formulae for determining the relative amount of skill in forecasting.''

Clayton gave some examples of the types of information that could be derived from this table: 1) 86% [=100(85/99)] of the forecasts of fair weather and 35% [=100(6/17)] of the forecasts of cloudy weather verified; 2) 91% [=100(106/116)] of the forecasts indicating no measurable precipitation (i.e., fair or cloudy) verified; and 3) 53% [=100(22/38)] of the forecasts indicating measurable precipitation verified. He then compared the overall performance of the predictions as precipitation/no-precipitation forecasts under two conditions: 1) when they were evaluated at Blue Hill Observatory according to U.S. Signal Service rules, and 2) when they were evaluated at the U.S. Signal Service station in Boston. These comparisons gave some indication of the sensitivity of the results to different methods of verification and/or different interpretations of these methods. Clayton pointed out that even if the results were insensitive to such factors, ''the per cent(age) of predictions verified . . . gives only a poor estimate of the success of the predictions, without a consideration of the number of occurrences which were not predicted,'' and ''in order to form a true estimate of the value of the predictions, the number which might have been verified by chance coincidence ought to be taken into account'' (p. 217).

Clayton illustrated the computation of verification measures from the contingency table by calculating Gilbert's measure i_G for forecasts of different types of weather. In the case of fair weather, $i_G = 0.37$ [=(85 − 66.2)/(99 + 103 − 85 − 66.2)]. Clayton also reported that $i_G = 0.35$ [=(106 − 91.9)/(116 + 122 − 106 − 91.9)] for precipitation/no-precipitation forecasts.

In conclusion, Clayton (1889) pointed out what he considered to be two deficiencies in Gilbert's verification measure. First, skill was determined with respect to the frequencies of the event of interest in the period covered by the predictions rather than the frequencies in a prior historical period. Second, since the measure focused on 2 × 2 verification problems, it failed to account for the magnitudes of errors in cases in which the forecasts were concerned with several degrees of intensity of a phenomenon.

In the opening paragraph of Clayton (1891), the author identified three reasons for verifying forecasts: 1) to determine how near perfection they approach, 2) to compare forecasts made in different places, and 3) to compare forecasts made at the same location but at different times (to determine trends). He then argued that verification methods currently in use ''fail in all these particulars'' (p. 370). His arguments with respect to each reason can be summarized in turn as follows (pp. 370–371). 1) ''The vagueness or . . . definiteness of the predictions'' was not considered when forecasts were judged to be either 100% verified or 100% not verified. 2) ''No idea of the frequency of occurrence of the phenomenon at the various places is included.'' 3) ''When the conditions are settled . . . the percentage of success will be higher than when the weather changes are frequent.'' In making these points the author raised questions about the relationship between the percentage of success and both ''skill in forecasting'' and the ''relative value of the forecasts.''

Clayton then considered several ''elements'' that must be included when describing success in forecasting: 1) the kind of phenomenon (clouds, rain, etc.), 2) the time of occurrence, 3) the duration of occurrence, 4) the intensity, 5) the importance to the community, and 6) the lead time. In the case of forecasts made irregularly, he identified a seventh item—the periods during which ''weather was . . . settled and hence easily predicted'' (p. 371). Clayton prefaced this list by pointing out the need to take ''the chances of accidental coincidences into account'' (p. 371). In addition, with regard to item 5, he noted that ''this also involves the question as to how much information furnished by the prediction exceeds the information derived by the average man from the ordinary portents of the sky'' (p. 371).

In summarizing his overall point of view on measuring the success of predictions, Clayton indicated that ''each one of these elements would need to be considered separately, the probability of chance coincidences eliminated, and the whole then combined into one statement'' (p. 372). As to the realization of such a ''complete'' approach in practice, Clayton noted that ''if . . . the method of verification needs to be very simple it would have to be confined to the determination of the success in predicting the mere occurrence of different kinds of phenomen[a]'' (p. 372).

Clayton then proceeded to discuss several aspects (or elements) of a general approach, based either on his own opinions or on recent work by others. With regard to the latter, he cited Nichols's method of measuring the relative value of weather predictions (see section 4e) and Gilbert's method of accounting for

chance coincidences (see section 3a). In applying these methods Clayton noted the difficulties associated with determining the economic consequences (in dollars) of anticipating or not anticipating alternative events and the limitation of Gilbert's measure i_G to phenomena described in terms of occurrence and nonoccurrence. Clayton also discussed problems associated with comparing predictions made for different lead times, including the duration of phenomena that were already occurring at the time the prediction was made.

Next, the author presented and discussed a slightly modified version of the contingency table reproduced in Table 4, which he described as ''the simplest method of verification'' (p. 374). [In a subsequent footnote he indicated that ''this is the method used in Germany and may be called the Köppen method'' (p. 375).] As in his earlier paper, Clayton computed the ''per cent of skill'' (p. 374) for these data using Gilbert's measure i_G and obtained a value of 37% ($=100i_G$) when all of the phenomena were considered to be of equal importance and (individual values of i_G for each phenomenon) were weighted in proportion to their frequency of occurrence. However, he subsequently argued that such phenomena are *not* of equal importance, and ''hence in determining the average success in predicting it seems desirable that each phenomenon should be weighted in proportion to its importance even though the weighting be somewhat arbitrary'' (p. 375).

In conclusion, Clayton (1891) indicated that ''it would be of great advantage if our Weather Service fully verified forecasts by some such method as that described above because besides giving a much better idea of the relative value of the forecasts and furnishing valuable information for the guidance of forecasters, it would be a direct method of climatic research by furnishing the relative frequency of the kind, duration, and intensity of those phenomena in which the public are most directly interested'' (p. 375).

e. Nichols (1890)

Nichols (1890) was concerned primarily with questions related to the economic value of the weather forecasts issued by the U.S. Signal Service and the influence of forecast accuracy on forecast value. In his opening paragraph, he noted that ''the proper test of the service is its value to the community'' (p. 386).

Nichols focused on forecasts of rainfall (as an example) and developed simple mathematical expressions for both forecast accuracy and forecast value, in the case of categorical forecasts for a two-event (i.e., rain/no-rain) variable. His proposed measures of accuracy included the overall fraction correct, the fraction correct for each of the two possible forecasts, and the ratio of correct to incorrect forecasts for each possible forecast.

In the process of developing a method of measuring forecast value, Nichols introduced quantities representing the gains and losses associated with anticipating and not anticipating rainfall occurrence and nonoccurrence. An important feature of this development was Nichols's recognition that to determine the value of the U.S. Signal Service forecasts it was necessary to compare the overall gain or loss when users' decisions were based on the forecasts with the overall gain or loss that would have been realized if users' decisions had been based on ''popular information concerning the weather derived from experience, weather signs, etc.'' (p. 388). He presented expressions for overall gain or loss in both situations and then briefly described the conditions with respect to forecast accuracy that would or would not guarantee that decisions based on the forecasts would produce a gain. He concluded this development by indicating that ''the relative importance of the thing predicted is an important factor in determining whether a series of predictions have been of service to the community'' (p. 388).

Nichols then addressed the issue of forecast value in relation to factors such as forecast length (i.e., lead time), the amount and duration of rainfall, and the spatial extent of rainfall coverage. In brief, he presented an argument as to why forecasts with longer lead times should be more valuable, indicated that rainfall predictions should contain information on amount and ''avoid unnecessary vagueness,'' and suggested that—to enhance value—''predictions should be localized even at the sacrifice of accuracy'' (p. 390).

Another interesting and important feature of the author's paper is his discussion of issues related to the uncertainty in forecasts. It is useful to quote the author at length here. In the case of the influence of this uncertainty—and the various possible gains and losses experienced by users—Nichols writes: ''The greater the value resulting from fulfillment in proportion to the injury resulting from failure, the smaller is the degree of probability that will justify a prediction . . . . A probability which would justify a warning of tornadoes or destructive gales, might not warrant an ordinary weather prediction. It is not simply the likelihood of the event but its relative importance, that determines the wisdom of the warning'' (p. 390).

Two short paragraphs later Nichols wrote: ''The problematic character of the language employed such as rain will fall, or is probable, or is possible, is in a crude way a measure of the degree of probability, and is governed by similar principles. It defines more closely the character of the prediction and tends to render positive, values which might otherwise be negative. In so far as it can be successfully employed, it broadens the scope and usefulness of the Service. . . . A knowledge of the degree of certainty with which an event may be expected, increases the value of the information'' (p. 391).

In summary, Nichols states that ''weather predictions are not to be judged simply from their technical accuracy, but rather from their practical value to the

community. . . . Where the event is unimportant the probability should be correspondingly strong to justify predictions which may tend to discredit the reliability of the reports. But the greater the importance the smaller need be the probability involved, while to avoid the sacrifice of accuracy and confidence the problematic character of such predictions should, as far as possible, be indicated'' (p. 391).

f. Köppen (1884, 1893)

Although the authorship of the paper identified here as Köppen (1884) is in some doubt, it is generally credited to Köppen [e.g., see Goodman and Kruskal (1959), p. 132]. Since this paper, which is written in German, appears to be similar in many respects to Köppen (1893) [abstracts of both papers appear in Muller (1944)], attention is focused here on the latter paper. With regard to priority, however, it should be kept in mind that many of the concepts and methods set forth in Köppen (1893) may well have been enunciated at least as early as 1884.

In the opening paragraphs of Köppen (1893), the author outlined and illustrated his approach to verification problems. This approach was based on the notion that different kinds of weather should follow different forecasts. In a 2×2 problem, Köppen argued that chance predictions and real predictions should be readily distinguishable, since ''opposite'' forecasts should be followed by similar weather in the former case and by different weather in the latter case.

Köppen illustrated his approach by comparing two kinds of predictions made for Hamburg, Germany, during the summer of 1883. The results, as summarized in Table 5, indicate the percentages of cases in which the succeeding weather was observed to fall in each of three temperature categories (warm, normal, cold) given each of two possible forecast categories (warm, cold). Class A predictions (Table 5a) were forecasts made a month in advance, whereas class B predictions (Table 5b) were daily forecasts. Clearly, the weather following the class A forecasts, which Köppen characterized as true chance predictions (recall that the time is the late nineteenth century), is similar in aggregate whether the predicted category is ''warm'' or ''cold.'' In the case of the class B forecasts, on the other hand, marked differences can be seen in the aggregate weather following the predicted categories of warm and cold. Köppen remarked that ''in cases of this sort this method (of verification) is entirely conclusive and sufficient''(p. 30).

The author then addressed the question of whether the relationship between forecasts and observations (as reflected in the class B forecasts above) can be expressed by a single number. Citing ''detailed researches,'' the author concluded that ''an irreproachable derivation of such a simple number for the expression of the worth of a prediction is impossible—almost as impossible as to estimate the value of a person or a nation by a numerical expression'' (p. 30). He referred to the measures formulated by Gilbert and Peirce as ''ingenious,'' but suggested that they were ''one sided'' and did not have ''universal application'' (p. 30).

TABLE 5. Succeeding weather in percentages following two kinds of predictions [after Köppen (1893, p. 30)].

	Succeeding weather		
Predicted weather	Warm	Normal	Cold
	(a) Class A predictions		
Warm	17	33	50
Cold	14	32	54
	(b) Class B predictions		
Warm	77	15	8
Cold	0	5	95

Köppen then discussed the deficiencies in the percentage of success as a measure of forecasting performance, noting its failure to take into account the influence of chance and its lack of regard for the frequencies of the events. He used Finley's tornado data (see Table 3) to contrast his method of assessing forecasting performance in terms of the percentages of cases in which predictions of ''tornadoes'' and ''no tornadoes'' were followed by observations of these two conditions with the use of the percentage of success (i.e., the measure i_F) as a measure of overall performance. To support his position, the author presented Finley's results in a form analogous to that employed in Table 5—Köppen's table in the case of Finley's forecasts is essentially identical to Table 7b. In concluding this discussion, the author asks ''what good is it when . . . 96.61 is given as the percentage of success of this prediction, a seemingly high number, but which is, nevertheless, inferior to that which would be had if, without trouble, a daily prediction of 'no tornado' had been made, for then this number would be . . . 98.18 per cent'' (p. 31). Köppen concluded this discussion of the deficiencies in the percentage of success as measure of forecasting performance by mentioning problems related to the characterization of weather conditions over an area in terms of the occurrence/nonoccurrence of phenomena and difficulties related to the consistent application of rules—in the course of the verification process—regarding the interpretation of forecasts and observations.

Next, Köppen described in some detail the method of verification used by the German Seewarte since 1886. In essence, this method involved the preparation of comprehensive tables containing the forecasts and observations, recorded for a prescribed set of weather variables (i.e., temperature, wind speed and direction, and precipitation) in well-defined categories. [Exam-

TABLE 6. General description of 2×2 verfication problem in terms of contingency tables consisting of joint, conditional, and marginal relative frequencies.

Forecasts	Observations: Event 1	Observations: Event 2	
(a) Joint and marginal relative frequencies			
Event 1	p_{11}	p_{12}	$p_{1\cdot}$
Event 2	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	1
Key: $p_{ij} = n_{ij}/n_{\cdot\cdot}$ $(i,j = 1, 2)$, $p_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot}$ $(i = 1, 2)$, $p_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot}$ $(j = 1, 2)$			
(b) Conditional relative frequencies of observations given forecasts			
Event 1	q_{11}	q_{12}	1
Event 2	q_{21}	q_{22}	1
Key: $q_{11} = p_{11}/p_{1\cdot}$, $q_{12}/p_{1\cdot}$, $q_{21} = p_{21}/p_{2\cdot}$, $q_{22} = p_{22}/p_{2\cdot}$.			
(c) Conditional relative frequencies of forecasts given observations			
Event 1	r_{11}	r_{12}	
Event 2	r_{21}	r_{22}	
	1	1	
Key: $r_{11} = p_{11}/p_{\cdot 1}$, $r_{12} = p_{12}/p_{\cdot 2}$, $r_{21} = p_{21}/p_{\cdot 1}$, $r_{22} = p_{22}/p_{\cdot 2}$			

ples of these basic tables appeared in Köppen (1884, pp. 401–403).] Köppen noted that these tables ''present much material for study which is well consolidated and is thoroughly controllable and comparable'' (p. 33). Moreover, Köppen (1884) contains contingency tables (pp. 400 and 404) describing forecasting performance—in terms of joint and marginal frequencies—during the summer of 1883 for several different variables, each of which is defined in terms of two, three, or four categories.

Köppen then posed the question, ''What is the value of weather predictions?'' (p. 33). He argued that different individuals could find the same forecasts, with the same percentage of success, of quite different value. Among other things, he noted that an individual ''may declare the most successful warnings to be useless because he might have predicted just as well the bad weather'' (p. 33). Köppen concluded this discussion by indicating that ''a strict proof of the practical worth of weather predictions and storm warnings is impossible,'' and, in this forecast-value context, that ''fulfillment of the prophecy counts for only one, though perhaps the most important, of many circumstances'' (p. 33).

In conclusion, the author emphasized the difference between what should be done within a weather service to assess forecasting performance and the way in which the results of such an assessment should be reported to the public. With regard to the former, he recommended the use of methods at least as strict as those introduced at the German Seewarte. With respect to the public, he indicated that ''a statement, as direct and as free from excuses as possible, with concrete examples and the opinion of specialists, is the most convincing and suitable'' (p. 34).

5. Discussion and interpretation of the Finley affair

a. Similarities and differences among 1880s measures

The verification measures formulated and/or used by Finley, Gilbert, Peirce, and Doolittle—the so-called 1880s measures—possess certain noteworthy similarities and differences. To facilitate a discussion of these similarities and differences, it is useful to introduce general notation to describe joint, conditional, and marginal relative frequencies of forecasts and/or observations. Table 6a identifies the notation for the joint relative frequencies $p_{ij} = n_{ij}/n_{\cdot\cdot}$, the marginal relative frequencies of the forecasts $p_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot}$, and the marginal relative frequencies of the observations $p_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot}$ $(i, j = 1, 2)$.

Conditional relative frequencies of two types are of interest here—namely, relative frequencies conditional on the forecasts and relative frequencies conditional on the observations. These conditional relative frequencies can be derived from the joint and marginal relative frequencies. Table 6b identifies the conditional relative frequencies of the observations given the forecasts, $q_{ij} = p_{ij}/p_{i\cdot}$, and Table 6c identifies the conditional relative frequencies of the forecasts given the observations $r_{ij} = p_{ij}/p_{\cdot j}$ $(i, j = 1, 2)$.

The pooled results of Finley's experimental tornado forecasting program are presented in the form of joint, conditional, and marginal relative frequencies in Table 7. Interpretation of the joint and marginal relative frequencies in Table 7a is relatively straightforward. For

TABLE 7. Representation of pooled results of Finley's experimental tornado forecasting program in terms of contingency tables consisting of joint, conditional, and marginal relative frequencies.

Forecasts	Observations: Tornado	Observations: No tornado	
(a) Joint and marginal relative frequencies			
Tornado	0.010	0.026	0.036
No tornado	0.008	0.956	0.964
	0.018	0.982	1
(b) Conditional relative frequencies of observations given forecasts			
Tornado	0.280	0.720	1
No tornado	0.009	0.991	1
(c) Conditional relative frequencies of forecasts given observations			
Tornado	0.549	0.026	
No tornado	0.451	0.974	
	1	1	

example, tornadoes were both forecast and observed on 1.0% of the forecasting occasions, and tornadoes were observed on 1.8% of the forecasting occasions.

Conditionality is determined by the forecasts in Table 7b. This table indicates that the event ''tornado'' was observed on 28% of the occasions on which it was forecast and that the event ''no tornado'' was observed on more than 99% of the occasions on which it was forecast. Table 7b is essentially identical to a table presented by Köppen (1893, p. 31).

In Table 7c conditionality is determined by the observations. This table indicates that the forecast ''tornado'' was made on almost 55% of the occasions on which tornadoes were subsequently observed and that the forecast ''no tornado'' was made on more than 97% of the occasions on which no tornadoes were subsequently observed. These conditional relative frequencies (in percentages) are identical, or closely related, to terms that appear in Peirce's measure i_P [see (4) and Table 8].

For the purposes of comparing the 1880s measures, expressions defining these measures are presented in Table 8. In this table each measure is expressed in terms of joint and/or marginal relative frequencies. Some measures are also expressed in terms of conditional relative frequencies. In addition to the 1880s measures, a verification measure proposed by Clayton (1927, 1934) is included. This measure, denoted here by i_C, is latent in ideas set forth by Köppen (1893) (see section 4f). Moreover, it bears interesting relationships to the 1880s measures proposed by Peirce (1884) and Doolittle (1885a).

Many similarities and differences among these measures could be noted. For example, all of the measures included in Table 8, with the exception of i_F and v, can be expressed as ratios of joint and/or marginal relative frequencies with essentially identical numerators. This numerator—namely, $p_{11}p_{22}-p_{12}p_{21}$ (or its square)—is the determinant of the 2×2 matrix consisting of the joint relative frequencies of forecasts and observations (see Table 6a).

Here we focus on relationships between i_F and v, between i_G and i_D^*, and among i_P, i_C, and i_D. Note that both i_F and v represent measures of accuracy (i.e., the correspondence between individual pairs of forecasts and observations), with the former assessing accuracy over event 1 *and* event 2 and the latter assessing accuracy over only event 1. Since $v = (i_F - p_{22})/(1 - p_{22})$, it follows that $v \leqslant i_F$.

The measures i_G and i_D^* are quite similar. In today's terminology, these measures would be identified as skill scores (i.e., measures of relative accuracy). They are based on different measures of accuracy but the same standard of reference (i.e., chance). The measure i_G assesses skill with respect to forecasts of event 1, whereas the measure i_D^* assesses skill with respect to forecasts of both events. Note that $i_G \leqslant i_D^*$.

As indicated in Table 8, the measures i_P, i_C, and i_D can all be written solely in terms of conditional relative frequencies. Comparison of i_C and i_P reveals that they are symmetric in the sense that, in terms of conditionality, the roles of forecasts and observations are reversed in these two measures. The former is concerned with assessing performance with reference to the distribution of observed events (i.e., the $p_{\cdot j}$), whereas the latter is concerned with assessing performance with reference to the distribution of forecast events (i.e., the $p_{i\cdot}$). Each measure represents the algebraic difference between two independent conditional relative frequencies in the corresponding conditional contingency table (cf. Tables 6b and 6c). The measure i_D is simply the product of i_C and i_P; that is, $i_D = i_C i_P$.

b. Rediscoveries of 1880s measures

Since the late 1930s several reviews and bibliographies containing references to the early literature on forecast verification have appeared (e.g., Bleeker 1946; Daan 1984; Goodman and Kruskal 1959; Johnson 1957; Meglis 1960; Muller 1944a,b,c; Weightman et al. 1939). Collectively, these publications, together with the references cited therein, include most of the veri-

TABLE 8. The 1880s verification measures (see section 5a) expressed in terms of (a) joint and/or marginal relative frequencies and (b) conditional relative frequencies (see note).

Author (year)	Measure	(a) Joint and/or marginal relative frequencies	(b) Conditional relative frequencies	Range
Finley (1884)	i_F	$p_{11} + p_{22}$		[0, 1]
Gilbert (1884)	v	$p_{11}/(p_{11} + p_{12} + p_{21})$		[0, 1]
Gilbert (1884)	i_G	$(p_{11}p_{12} - p_{22}p_{21})/[(p_{11}p_{22} - p_{12}p_{21}) + (p_{12} + p_{21})]$		$(-\infty, 1]$
Peirce (1884)	i_P	$(p_{11}p_{22} - p_{12}p_{21})/(p_{\cdot 1}p_{\cdot 2})$	$r_{11} - r_{12}$	[−1, 1]
Doolittle (1885)	i_D	$(p_{11}p_{22} - p_{12}p_{21})^2/(p_{1\cdot}p_{2\cdot}p_{\cdot 1}p_{\cdot 2})$	$(r_{11} - r_{12})(q_{11} - q_{21})$	[0, 1]
Doolittle (1888)	i_D^*	$(p_{11}p_{22} - p_{12}p_{21})/[(p_{11}p_{22} - p_{12}p_{21}) + (1/2)(p_{12} + p_{21})]$		$(-\infty, 1]$
Clayton (1927)	i_C	$(p_{11}p_{22} - p_{12}p_{21})/(p_{1\cdot}p_{2\cdot})$	$q_{11} - q_{21}$	[−1, 1]

Note: Since conditional relative frequencies are defined as the ratio of joint relative frequencies to marginal relative frequencies (see Table 6), all of these verification measures can be expressed in terms of conditional relative frequencies *and/or* joint or marginal relative frequencies. Here, expressions are included under (b) only for those measures that can be written exclusively in terms of conditional relative frequencies.

fication papers published during the period 1884–1893 and summarized here. Evidently, relatively few individuals in the last 50–75 years have taken the time to acquaint themselves with the contents of these papers. This statement is supported by the fact that several of the 1880s measures have been ''rediscovered'' on one or more occasions since 1900. Moreover, in most cases, these 1880s measures are now known by names assigned to them by later authors. Some examples of these rediscoveries are discussed next.

The measure labeled v by Gilbert (1884) [see (2) and Table 8] has been rediscovered and renamed at least twice. Palmer and Allen (1949), in the context of verifying categorical precipitation forecasts, found that precipitation was neither forecast nor observed in a vast majority of the cases (i.e., p_{22} dominated the other elements in the 2×2 contingency table; see Table 6a). To avoid being overwhelmed by such cases, Palmer and Allen (1949) developed a measure called the threat score (TS). About 25 years later, Donaldson et al. (1975), working in the area of severe weather forecasting, proposed the critical success index (CSI) as an indicator of forecasting performance in this context. As noted recently by Schaefer (1990), both TS and CSI are identical to Gilbert's v.

Schaefer (1990) contains a noteworthy example of an exception to the tendency to rename the 1880s measures. In this paper the author formulates a skill-dependent CSI (or TS) following the approach taken by Gilbert (1884). Recognizing that this measure is identical to Gilbert's i_G [see (3) and Table 8], Schaefer labels it the Gilbert skill score. On the other hand, a measure identical to i_G has recently been referred to as the equitable threat score (Black 1994).

Measures similar or identical to the measure i_P [see (4) and Table 8], formulated by Peirce (1884), have been ''discovered'' at least twice. The first example is Kuipers's performance index (Hanssen and Kuipers 1965), which was introduced in the Netherlands in 1954 (see Daan 1984, p. 18). This measure is similar but not identical to i_P, at least in its original formulation. The difference is that the correction for chance in Kuipers's index is based on historical climatological relative frequencies, whereas this correction is based on sample relative frequencies in the case of i_P.

More recently, Flueck (1987) formulated a verification measure that he called the true skill statistic. This measure is identical to i_P, a fact noted by Flueck in a footnote. In the footnote, Flueck also indicated that Clayton's i_C was identical to i_P. Although i_C and i_P possess a certain symmetry (see section 5a), they are identical only when the forecasts of interest are unconditionally unbiased (i.e., $p_{1\cdot} = p_{\cdot 1}$), a fact that can be established by comparing the expressions for these measures in Table 8.

In the case of the measure i_D, formulated by Doolittle (1885), it is relatively easy to show that this measure is identical to the square of the binary-event version of the Pearson product-moment correlation coefficient (r) (e.g., Bishop et al. 1975, pp. 380–381). That is, $i_D = r^2$ [see (5) and Table 8]. This measure of association is generally referred to as the coefficient of determination. Moreover, the measure i_D ($= r^2$) is closely related to Pearson's chi-square statistic and Pearson's coefficient of mean square contingency (see Bishop et al. 1975, pp. 382–383).

The measure labeled here i_D^* was formulated by Doolittle (1888) [see (6) and Table 8]. In today's terminology, it was developed by Doolittle as a skill-dependent, or chance-corrected, version of Finley's i_F. As such, this measure is identical to the 2×2 version of the skill score proposed by Heidke (1926) and known today almost universally as the Heidke skill score. Thus, Doolittle can be credited with developing the Heidke skill score almost 40 years before Heidke himself!

As noted in section 5a, the measure i_C was proposed by Clayton (1927, 1934). Specifically, this measure was defined verbally in Clayton (1927) and numerically (i.e., in terms of a mathematical expression) in Clayton (1934). Despite its development more than 30 years after the Finley affair (as defined in this paper), the measure i_C is included here for three reasons: 1) it is very closely connected to ideas advanced by Köppen (1893); 2) it bears a symmetrical relationship to i_P (Peirce 1884; see Table 8); and 3) it is one of two factors—the other is i_P—that enters into the ''construction'' of i_D (Doolittle 1885a; see section 3c).

With regard to Köppen's anticipation of the measure i_C, he suggested comparing the conditional relative frequencies (or percentages) of the observed events when event 1 is forecast to occur with the conditional frequencies of these same observed events when event 2 is forecast to occur (see Table 5), as a basis for evaluating forecasting performance. In a 2×2 verification problem, only two of these (four) conditional relative frequencies are independent, and Clayton's measure i_C is the difference between the two independent conditional relative frequencies that relate to the occurrence of event 1. For Finley's data, $i_C = q_{11} - q_{21} = 0.280 - 0.009 = 0.271$ (see Table 7b). Thus, although Köppen did not suggest a specific scalar measure of forecasting performance in this conditional framework, Clayton's i_C accomplishes precisely the kind of comparison that Köppen obviously had in mind. Moreover, i_C is the only measure defined as a difference between conditional relative frequencies of observations given forecasts that makes any sense in such 2×2 problems. (Since $q_{11} + q_{12} = 1$ and $q_{21} + q_{22} = 1$, it follows that $q_{22} - q_{12} = q_{11} - q_{21}$.)

c. Issues in forecast verification: 1880s and 1990s

In addition to the noteworthy verification methods and measures formulated during the Finley affair, various basic issues underlying verification methods and

practices were discussed—in some cases, for the first time—during the 10-year period from 1884 to 1893. In this section some of these issues are examined briefly, first from the perspective of the participants in the Finley affair and then from the perspective of the mid-1990s. A list of the issues to be considered, together with the names of the participants in the Finley affair who contributed to the discussion of these issues, can be found in Table 9.

The concept of taking chance coincidences between forecasts and observations into account in assessing forecasting performance, evidently introduced by Gilbert (1884), is perhaps the most obvious example of such an issue. In effect, Gilbert and others who promoted this concept recognized the need to distinguish between absolute (or pure) accuracy and relative accuracy. In this case, relative accuracy was defined in terms of the difference between the accuracy of the forecasts of interest and the accuracy of forecasts based on chance (i.e., statistical independence between forecasts and observations). Gilbert also introduced the term ''skill'' (initially ''skill in inference'') to describe this aspect of relative forecasting performance. According to current usage, ''skill'' refers to accuracy relative to any of several common standards of reference (e.g., chance, climatology, persistence).

A related issue mentioned only by Clayton (1889) is the issue of using sample relative frequencies of events rather than historical climatological relative frequencies of events in defining the standard of reference and assessing skill. Clayton advocates the use of the latter. It is clear today (and possibly to Clayton and others at the time of the Finley affair) that use of the former, although convenient and internally consistent, underestimates forecast skill. Specifically, it fails to give any credit to the forecaster or forecasting method for recognizing that the relative frequencies of the events during the period of interest differ from those that prevailed during the prior historical period.

Practical issues related to the verification of Finley's tornado forecasts were mentioned briefly by Finley himself and discussed in greater detail by Curtis and Hazen. The suggestion by Curtis (1887) that tornado forecasts should be formulated for movable rather than fixed districts is particularly noteworthy. His specific recommendation that the forecasts apply to a rectangular area with a length 1.5 times its width anticipates the format adopted shortly after tornado and severe local storm forecasts were first issued officially by the National Weather Service in the early 1950s (Galway 1989). Some of the practical difficulties associated with verifying tornado forecasts that were identified by these early writers have not yet been satisfactorily resolved more than 100 years later (e.g., see Anthony and Leftwich 1992; Doswell et al. 1990; Mason 1989). To a considerable degree these problems exist for all forecasts that are not made on a regular basis and for which routine observations are not available.

TABLE 9. Some issues related to forecast verification discussed by individuals who contributed to the Finley affair.

Issue	Individual(s)
Chance coincidences	Gilbert, Doolittle, Clayton
Sample climatology vs historical climatology	Clayton
Problems in verifying tornado forecasts	Finley, Curtis, Hazen
Extensions to polychotomous events	Peirce, Doolittle, Clayton
Purposes of verification	Clayton
Metaverification	Gilbert, Doolittle
Qualitative vs quantitative verification	Hazen, Clayton, Köppen
Deficiencies in one-dimensional measures	Köppen
Joint distribution perspectives	Peirce, Clayton, Köppen
Forecast quality vs forecast value	Peirce, Nichols, Clayton, Köppen
Reporting forecasting performance	Köppen

Several authors mentioned the issue of extending the measures formulated for the 2×2 verification problem exemplified by Finley's tornado/no-tornado forecasts to the general $k \times k$ ($k \geq 2$) verification problem. Peirce (1884) claimed to have a solution for this general problem, evidently involving the assignment of weights to different kinds of errors, but he did not present it. Doolittle and Clayton both remarked on the difficulties associated with formulating an acceptable measure for the $k \times k$ problem. Over the last 100+ years, many attempts have been made to develop such measures, and a variety of solutions have been presented (e.g., Doswell et al. 1990; Gandin and Murphy 1992; Gringorten 1967). However, all of these solutions involve introducing simplifying assumptions, imposing restrictive conditions, and/or assigning arbitrary weights. It is now generally understood that no universally acceptable measure of performance for the $k \times k$ problem can be found.

Before embarking on studies involving the formulation and/or application of verification methods, it is essential to specify the purposes for which the verification is to be undertaken. Recognition of this fact is implicit if not explicit in several places in Clayton's papers (e.g., Clayton 1891, pp. 369–370). More recently, the importance of developing a clear understanding of the purposes for which verification is performed has been strongly emphasized in the review paper by Brier and Allen (1951) and in some of the other reviews or surveys of the subject that have appeared since 1939 (see section 5b).

As indicated in section 3a, Gilbert (1884) used several tests to investigate the properties of the measure i_G. These tests embody an early form of what might be referred to as ''metaverification''—namely, determining whether or not verification measures satisfy specific criteria and/or possess particular properties. Despite the limited nature of Gilbert's tests, the notion that verification measures themselves should be evaluated us-

ing metaverification methodology is clearly latent in his paper. Recent examples of the use of metaverification methods in conjunction with 2×2 (and other) verification measures can be found in Daan (1984) and Woodcock (1976).

Perhaps not surprisingly, not everyone supported the idea that forecast verification should consist of calculating measures of the degree of correspondence between forecasts and observations of weather conditions. Hazen (1892, p. 394) suggested that verification should be based on an expert's assessment of the extent to which the forecast was consistent with the weather map on which the forecast was based. However, the majority view—expressed by Clayton and Köppen among others—indicated that forecast verification served important purposes and that it should be performed in a quantitative and objective manner. Moreover, both of these latter authors mentioned the need to reduce ambiguity and subjectivity in the verification process.

Köppen (1893, p. 30) commented on the difficulties inherent in any attempt to describe the relationship between ''predictions . . . and the following weather'' in terms of a single number. By a ''single number,'' he meant a scalar, or one-dimensional, measure of overall forecasting performance. As noted in section 4f, Köppen clearly believed that it was impossible to derive a one-dimensional measure that could describe all aspects of forecast quality in a universally acceptable manner.

The fact that one-dimensional measures are unable to provide an entirely satisfactory solution to verification problems raises the question of what would constitute a more appropriate approach. Comments by Peirce, Clayton, and Köppen address this issue, directly or indirectly, and they are of particular interest in view of recent efforts (by the author of this paper among others) to develop a conceptually sound and methodologically insightful approach to verification problems. In brief, this recent effort is based on the concept that the empirical joint relative frequencies of forecasts and observations contain all of the information relevant to forecast verification (Murphy and Winkler 1987). In this regard, it is interesting to note that, as indicated in section 3b, Peirce (1884) assumed that two forecasting methods that yield the same joint frequencies (equivalently, the same joint relative frequencies) should be considered to be equally successful.

Moreover, both Clayton and Köppen recommended an approach to forecast verification that emphasized inspection of the joint frequencies of the various possible combinations of forecasts and observations, and they displayed these joint frequencies in tables that we would identify today as contingency tables. Since the contents of such tables represent the empirical joint distributions of forecasts and observations (where each joint relative frequency has been multiplied by the sample size), these authors appear to be very early advocates of a distributions-oriented approach to forecast verification (see Murphy and Winkler 1987).

Although the question of the economic value of forecasts is generally not considered to fall within the scope of forecast verification, a strong argument can be made that, in the final analysis, it is not possible—or necessarily even desirable—to separate issues of forecast quality from issues of forecast value. In this regard, issues related to assessing the economic benefits of forecasts and/or determining the relationship between forecast quality and forecast value were raised on several occasions during the period 1884–1893. For example, Peirce (1884) and Nichols (1890) developed expressions for determining the value of forecasts in relatively simple situations. Moreover, Nichols quite properly pointed out the need to measure the value of the forecasts as (in effect) the difference between the economic welfare of users when their decisions were made with and without the aid of the forecasts. This basic concept is essential to an understanding of the way in which weather forecasts acquire value. Nevertheless, evidence appears from time to time that suggests that even in the mid-1990s this concept is not well understood in the meteorological community at large.

Köppen's appreciation of the complex nature of the relationship between forecast quality (or specific aspects of quality) and forecast value is evident in his brief comments on this issue. Studies of the quality/value relationship in a relatively wide variety of prototypical and real-world decision-making problems over the last 15–20 years indicate that this relationship is inherently nonlinear (e.g., see Murphy 1994). Moreover, it is even possible for reversals in the usual monotonic quality/value relationship to occur, if quality is not measured in a way that respects the underlying dimensionality of the associated verification problem (e.g., Murphy and Ehrendorfer 1987).

Finally, Köppen (1893, pp. 33–34) briefly discussed issues related to reporting (i.e., summarizing, describing, and communicating) information regarding forecasting performance to members of the meteorological community as well as to members of the public at large. He advocated the provision of detailed information to the meteorological community and, as noted in section 3f, the communication of relatively simple reports to the public.

In view of the relatively wide range of verification-related issues discussed in these early papers, it might be appropriate to identify one or two issues that evidently did not arise at the time of the Finley affair. An example of such an issue is the effect that verification measures themselves may have on a forecaster's decision as to whether to forecast the occurrence or nonoccurrence of a particular weather event or condition (e.g., tornadoes). It is not clear when this issue first arose, but it was referred to by Brier (1950) in his discussion of the properties of the so-called Brier score (see also Brier and Allen 1951). Another example re-

lates to methods of verifying forecasts expressed in terms of probabilities. Since forecasts containing quantitative descriptions of uncertainty were evidently first reported by Cooke (1906) (see Liljas and Murphy 1994), verification methods for such forecasts were not an issue at the time of the Finley affair.

6. Conclusions

This paper has focused on a signal event—the Finley affair—that marked the beginning of substantive developments and discussions related to concepts, methods, and practices in forecast verification. As described here, the Finley affair covers the period from 1884 to 1893 and is conveniently, if somewhat arbitrarily, divided into three parts. The first part consists simply of Finley's paper, published in July 1884. Finley's results and method of verification stimulated considerable interest inside and outside the meteorological community. The second part constitutes three early responses to Finley's paper by Gilbert, Peirce, and Doolittle, published in a five-month period from September 1884 to January 1885. One or more measures of forecasting performance were formulated in each of these papers. The third part is identified here as the ''aftermath'' and includes papers published by several authors during the period from 1887 to 1893. These aftermath papers addressed a variety of issues related to verification methods and practices.

Undoubtedly, the most remarkable features of the Finley affair are the number of verification methods and measures introduced and the range of issues discussed. In the case of measures of forecasting performance designed for the 2×2 verification problem (the problem posed by Finley's data), it is particularly noteworthy from the perspective of the mid-1990s that the following measures were formulated by participants in the Finley affair: (a) TS or CSI (Gilbert's v), (b) a skill score based on TS or CSI (Gilbert's i_G), (c) Kuipers's performance index or the true skill statistic (Peirce's i_P), (d) the coefficient of determination or square of the correlation coefficient (Doolittle's i_D), and (e) the 2×2 version of Heidke's skill score (Doolittle's i_D^*). In addition, a measure of performance subsequently described by Clayton (1927, 1934) is clearly latent in ideas set forth by Köppen (1893).

From the perspective of the author of this paper, it is also noteworthy that both Clayton and Köppen appear to advocate an approach to forecast verification based on the joint distribution of forecasts and observations (although these authors do not use the terminology ''joint distribution''). Köppen (1893) in particular pointed out the deficiencies in individual measures of forecasting performance and recommended summarizing the relevant data in the form of verification tables—in essence, contingency tables displaying the joint frequencies of forecasts and observations. Moreover, he introduced the concept of conditional verification tables in which the conditional relative frequencies (or percentages) of the weather conditions following each possible forecast were summarized and compared. The joint and conditional distributions characterized by these tables are cornerstones of a general distributions-oriented approach to forecast verification recently advanced by Murphy and Winkler (1987).

Several issues related to verification methods and practices were discussed during the course of the Finley affair (see Table 9). These issues ranged from methodological topics such as the need to take chance coincidences between forecasts and observations into account in measuring forecasting performance and the need to determine the value of forecasts and its relationship to forecast quality, to practical topics related to problems inherent in verifying tornado forecasts. In some cases novel solutions for particular problems were proposed (e.g., Curtis's suggestion of replacing fixed, irregularly shaped geographical districts with movable rectangular areas). Despite the fact that many of the problems associated with these issues do not possess simple or straightforward solutions, the discussants' arguments and insights reflect favorably on their understanding of the underlying problems, especially in view of the fact that these issues were addressed in papers published more than 100 years ago.

The significance of the Finley affair extends even beyond the innovative verification methods and measures developed during the 10-year period from 1884 to 1893 and the insightful and wide-ranging discussions of issues that accompanied and enhanced these methodological developments. This affair clearly contributed to a greater recognition of the importance of ensuring that the practice of forecast verification was based on sound concepts and methods, and it underlined the key role that forecast verification could play in describing—and in potentially improving—forecasting performance. Thus, Finley's paper, which has been known heretofore primarily for its presentation of the results of the first serious attempt to forecast tornadoes in the United States, should also be recognized as the catalyst for the first substantive developments within the subdiscipline of forecast verification. The basic threads of many concepts and methods in forecast verification advanced during the last 50 years can be traced back to the Finley affair.

In a related vein, a reviewer raised the question of the credit that should be given to contributors to the Finley affair, in view of the fact that these developments appear to have had relatively little direct impact on the evolution of the subdiscipline of forecast verification over the last 100 years. This question admits no simple answer. Clearly, priority regarding the formulation of the verification measures identified in sections 3 and 4 belongs to those individuals who made methodological contributions to the Finley affair. Moreover, the fact that these contributions were largely ignored or overlooked for 100 years may serve only to

exemplify the benign neglect afforded the subdiscipline of forecast verification by the meteorological community as a whole during this period. In any case, it seems clear that attention today should be focused on building upon these very early conceptual and methodological developments rather than on rediscovering, or renaming, verification measures first formulated more than a century ago.

Finally, it is also possible to draw some general lessons concerning the identification and evaluation of early work in other subdisciplines from this study of verification-related activities in the period between 1884 and 1893. First, even individuals who are acknowledged experts in a particular subdiscipline and are generally familiar with its historical development may occasionally be surprised by the contents of early papers. For example, prior to undertaking this historical study the author of this paper was unaware that Doolittle had formulated the 2×2 version of the Heidke skill score in 1888. Second, although review papers and bibliographies represent reasonable starting points for most historical studies, the search for relevant material should not be based solely on such sources. It is almost certain that any collection of secondary sources will overlook some potentially important material. Third, it is frequently necessary to look beyond the literature in the particular discipline or subdiscipline of interest. Similar problems or issues can arise—and relevant results may appear—in the context of other disciplines or subdisciplines. Moreover, in the nineteenth and early twentieth centuries individuals often worked in several disciplines (during a career), and the disciplines themselves—and their associated publications—were defined more broadly. Consequently, historical research requires the same high level of scholarship and attention to detail as basic and applied research.

Acknowledgments. The author is pleased to acknowledge the help provided by J. G. Galway, J. T. Schaefer, and C.-H. Yang in locating and/or acquiring copies of some of the papers and reports on which this historical study is based. The assistance of the staff at the libraries of the following institutions is greatly appreciated: U.S. Air Force Phillips Laboratory (Bedford, Massachusetts), National Weather Service, NOAA (Silver Spring, Maryland), and Oregon State University (Corvallis, Oregon). Three anonymous reviewers provided useful comments on an earlier version of the manuscript.

REFERENCES

Anthony, R. W., and P. W. Leftwich, 1992: Trends in severe local storm watch verification at the National Severe Storms Forecast Center. *Wea. Forecasting,* **7,** 613–622.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland, 1975: *Discrete Multivariate Analysis: Theory and Practice.* Massachusetts Institute of Technology Press, 557 pp.

Black, T. L., 1994: The new NMC mesoscale eta model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

Bleeker, W., 1946: The verification of weather forecasts. *Mededeelingen en Verhandelingen,* Serie B, Deel 1, No. 2, Royal Netherlands Meteorological Institute, 23 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

——, and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology,* T. F. Malone, Ed., Amer. Meteor. Soc., 841–848.

Burton, J., 1986: Robert Fitzroy and the early history of the Meteorological Office. *Br. J. Hist. Sci.,* **19,** 147–176.

Clayton, H. H., 1889: Verification of weather forecasts. *Amer. Meteor. J.,* **6,** 211–219.

——, 1891: Verification of weather forecasts. *Amer. Meteor. J.,* **8,** 369–375.

——, 1927: A method of verifying weather forecasts. *Bull. Amer. Meteor. Soc.,* **8,** 144–146.

——, 1934: Rating weather forecasts. *Bull. Amer. Meteor. Soc.,* **15,** 279–283.

Cooke, W. E., 1906: Forecasts and verifications in Western Australia. *Mon. Wea. Rev.,* **34,** 23–24.

Curtis, G. E., 1887: Tornado predictions and their verification. *Amer. Meteor. J.,* **4,** 68–74.

Daan, H., 1984: Scoring rules in forecast verification. *PSMP Publication Series,* No. 4, World Meteorological Organization, 62 pp.

David, H. A., 1995: First(?) occurrence of common terms in mathematical statistics. *Amer. Stat.,* **49,** 121–133.

Donaldson, R. J., R. M. Dyer, and M. J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *Ninth Conf. on Severe Local Storms,* Norman, OK, Amer. Meteor. Soc., 321–326.

Doolittle, M. H., 1885a: The verification of predictions. *Bull. Philos. Soc. Washington,* **7,** 122–127.

——, 1885b: The verification of predictions. *Amer. Meteor. J.,* **2,** 327–329.

——, 1888: Association ratios. *Bull. Philos. Soc. Washington,* **10,** 83–87, 94–96.

Doswell, C. A., R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.,* **1,** 85–88.

Flueck, J. A., 1987: A study of some measures of forecast verification. Preprints, *10th Conf. on Probability and Statistics in Atmospheric Sciences,* Edmonton, AB, Canada, Amer. Meteor. Soc., 69–73.

Galway, J. G., 1985: J. P. Finley: The first severe storms forecaster. *Bull. Amer. Meteor. Soc.,* **66,** 1389–1395, 1506–1510.

——, 1989: The evolution of severe thunderstorm criteria within the Weather Service. *Wea. Forecasting,* **4,** 585–592.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.,* **120,** 361–370.

Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.,* **1,** 166–172.

Goodman, L. A., and W. H. Kruskal, 1959: Measures of association for cross classifications. Part II: Further discussion and references. *J. Amer. Stat. Assoc.,* **54,** 123–163.

Gringorten, I. I., 1967: Verification to determine and measure forecasting skill. *J. Appl. Meteor.,* **6,** 742–747.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Mededeelingen en Verhandelingen,* No. 81, Royal Netherlands Meteorological Institute, 65 pp.

Hazen, H. A., 1887: Verification of tornado predictions. *Amer. J. Sci.,* Series 3, **34,** 127–131.

——, 1892: The verification of weather forecasts. *Amer. Meteor. J.,* **8,** 392–396.

Heidke, P., 1926: Berechnung des erfolges und der güte der windstärkvorhersagen im sturmwarnungsdienst. *Geografika Annaler,* **8,** 301–349.

Hughes, P., 1994: The great leap forward. *Weatherwise,* **47,** 22–27.

Johnson, D. H., 1957: Forecast verification: A critical survey of the literature. Bracknell, United Kingdom, Air Ministry, Meteorological Research Committee, M.R.P. 1056, S.C. 11/237, 40 pp.

Köppen, W., 1884: Eine rationelle methode zur prüfung der wetterprognosen. *Meteor. Z.,* **1,** 397–404.

——, 1893: The best method of testing weather predictions. U.S. Weather Bureau Bull. 11, pp. 29–34.

Liljas, E., and A. H. Murphy, 1994: Anders Ångström and his early papers on probability forecasting and the use/value of weather forecasts. *Bull. Amer. Meteor. Soc.,* **75,** 1227–1236.

Mason, I. B., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

Meglis, A. J., 1960: Annotated bibliography on forecast verification. *Meteor. Geoastrophys. Abstr. Bibliogr.,* **11,** 1129–1174.

Muller, R. H., 1944a: Verification of short-range weather forecasts (A survey of the literature). *Bull. Amer. Meteor. Soc.,* **25,** 18–27.

——, 1944b: Verification of short-range weather forecasts (A survey of the literature) II. *Bull. Amer. Meteor. Soc.,* **25,** 47–53.

——, 1944c: Verification of short-range weather forecasts (A survey of the literature) III (Conclusion). *Bull. Amer. Meteor. Soc.,* **25,** 88–95.

Murphy, A. H., 1994: Assessing the economic value of weather forecasts: An overview of methods, results, and issues. *Meteor. Appl.,* **1,** 69–73.

——, and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Wea. Forecasting,* **2,** 243–251.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Nichols, W. S., 1890: The mathematical elements in the estimation of the Signal Service reports. *Amer. Meteor. J.,* **6,** 386–392.

Palmer, W. C., and R. A. Allen, 1949: Note on the accuracy of forecasts concerning the rain problem. U.S. Weather Bureau, 4 pp.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science,* **4,** 453–454.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Weightman, R. H., R. T. Zoch, and H. R. Byers, 1939: Forecast verification: Report to the chief of the Weather Bureau by the Special Committee on Forecast Verifications. U.S. Weather Bureau, 44 pp.

Whitnah, D. R., 1961: *A History of the United States Weather Bureau.* University of Illinois Press, 267 pp.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.,* **104,** 1209–1214.