Verification of The Weather Channel Probability of Precipitation Forecasts

J. ERIC BICKEL* AND SEONG DAE KIM

Department of Industrial and Systems Engineering, Texas A&M University, College Station, Texas

(Manuscript received 30 January 2008, in final form 6 May 2008)

ABSTRACT

The Weather Channel (TWC) is a leading provider of weather information to the general public. In this paper the reliability of their probability of precipitation (PoP) forecasts over a 14-month period at 42 locations across the United States is verified. It is found that PoPs between 0.4 and 0.9 are well calibrated for near-term forecasts. However, overall TWC PoPs are biased toward precipitation, significantly so during the warm season (April–September). PoPs lower than 0.3 and above 0.9 are not well calibrated, a fact that can be explained by TWC's forecasting procedure. In addition, PoPs beyond a 6-day lead time are miscalibrated and artificially avoid 0.5. These findings should help the general public to better understand TWC's PoP forecasts and provide important feedback to the TWC so that they may improve future performance.

1. Introduction

The Weather Channel (TWC) is a leading provider of weather information to the general public via its cable television network and interactive Web site (see http://www.weather.com/). TWC's cable network is available in 95% of cable TV homes in the United States and reaches more than 87 million households. Their Internet site, providing weather forecasts for 98 000 locations worldwide, averages over 20 million unique users per month and is among the top 15 news and information Web sites, according to Nielsen/NetRatings (more information is available online at http://press.weather.com/company.asp).

The public uses TWC's forecasts to make decisions as mundane as whether to carry an umbrella or as significant as whether to seek shelter from an approaching storm. How accurate are these forecasts? Are they free from bias? Should the public accept TWC forecasts at face value or do they need to be adjusted to arrive at a better forecast?

In this paper, we analyze the reliability of probability of precipitation (PoP) forecasts provided by TWC (via weather.com) over a 14-month period (2 November 2004–16 January 2006), at 42 locations across the United States. Specifically we compare n-day-ahead PoP forecasts, where n ranges from 0 (same day) to 9, with actual precipitation observations.

This paper is organized as follows. In the next section, we describe our verification approach and review the associated literature. In section 3 we summarize our data collection procedure. In section 4 we present the reliability results and discuss the implications. In section 5 we present our conclusions.

2. Verification of probability forecasts

The literature dealing with forecast verification and value is extensive (e.g., for an overview see Katz and Murphy 1997; Jolliffe and Stephenson 2003). In this paper, we adopt the distribution-oriented framework proposed by Murphy and Winkler (1987, 1992).

a. Distributional measures

Let F be the finite set of possible PoP forecasts $f_i \in [0, 1]$, $i = 1$ to m. Here X is the set of precipitation observations, which we assume may obtain only the value $x = 1$ in the event of precipitation and $x = 0$ otherwise. The empirical relative frequency distribution of forecasts and observations given a particular lead time l is $p(f, x|l)$ and completely describes the performance of the forecasting system. A perfect

* Current affiliation: Operations Research/Industrial Engineering, The University of Texas at Austin, Austin, Texas.

Corresponding author address: J. Eric Bickel, Operations Research/Industrial Engineering Program, The University of Texas at Austin, Austin, TX 78712-0292.
E-mail: ebickel@mail.utexas.edu

DOI: 10.1175/2008MWR2547.1

forecasting system would ensure that $p(f, x|l) = 0$ when $f \neq x$. In the case of TWC, l may obtain integer values ranging from 0 (same day) to 9 (the last day in a 10-day forecast).

Since

$$p(f, x|l) = p(f|l)p(x|f, l) = p(x|l)p(f|x, l), \quad (1)$$

two different factorizations of $p(f, x|l)$ are possible and each facilitates the analysis of forecasting performance.

The first factorization, $p(f, x|l) = p(f|l)p(x|f, l)$ is known as the calibration-refinement (CR) factorization. Its first term, $p(f|l)$, is the marginal or predictive distribution of forecasts and its second term, $p(x|f, l)$, is the conditional distribution of the observation given the forecast. For example, $p(1|f, l)$ is the relative frequency of precipitation when the forecast was f and the lead time was l. The forecasts and observations are independent if and only if $p(x|f, l) = p(x|l)$. A set of forecasts is well *calibrated* (or reliable) if $p(1|f, l) = f$ for all f. A set of forecasts is perfectly *refined* (or sharp) if $p(f) = 0$ when f is not equal to 0 or 1 (i.e., the forecasts are categorical). Forecasting the climatological average or base rate will be well calibrated, but not sharp. Likewise, perfectly sharp forecasts generally will not be well calibrated.

The second factorization, $p(f, x|l) = p(x|l)p(f|x, l)$, is the likelihood-base rate (LBR) factorization. Its first term, $p(x|l)$, is the climatological precipitation frequency. Its second term, $p(f|x, l)$, is the likelihood function (also referred to as discrimination). For example, $p(f|1, l)$ is the relative frequency of forecasts when precipitation occurred, and $p(f|0, l)$ is the forecast frequency when precipitation did not occur. The likelihood functions should be quite different in a good forecasting system. If the forecasts and observations are independent, then $p(f|x, l) = p(f|l)$.

b. Summary measures

In addition to the distributional comparison discussed above, we will use several summary measures of forecast performance. The mean forecast given a particular lead time is

$$\bar{f}_l = \sum_F \sum_X fp(f, x|l) = E_{F,X|l}[f], \quad (2)$$

where $E[]$ is the expectation operator. Likewise, the climatological frequency of precipitation, indexed by lead time, is

$$\bar{x}_l = E_{F,X|l}[x]. \quad (3)$$

The mean error (ME) is

$$\mathrm{ME}(f, x|l) = \bar{f}_l - \bar{x}_l, \quad (4)$$

and is a measure of unconditional forecast bias. The mean-square error (MSE) or the Brier score (Brier 1950) is

$$\mathrm{MSE}(f, x|l) = E_{F,X|l}[(f - x)^2]. \quad (5)$$

The climatological skill score (SS) is

$$\mathrm{SS}(f, x|l) = 1 - \mathrm{MSE}(f, x|l)/\mathrm{MSE}(\bar{x}_l, x|l). \quad (6)$$

Note that

$$\mathrm{MSE}(\bar{x}_l, x|l) = E_{F,X|l}[(\bar{x} - x)^2] = \sigma_x^2,$$

where σ_x^2 is the variance of the observations. Therefore,

$$\mathrm{SS}(f, x|l) = \frac{\sigma_x^2 - \mathrm{MSE}(f, x|l)}{\sigma_x^2}, \quad (7)$$

and we see that SS measures the proportional amount by which the forecast reduces our uncertainty regarding precipitation, as measured by variance.

In addition to these scoring measures, we will also investigate the correlation between the forecasts and the observations, which is given by

$$\rho(f, x|l) = \frac{\mathrm{cov}(f, x|l)}{(\sigma_x^2 \sigma_f^2)^{1/2}}, \quad (8)$$

where cov is the covariance and σ_f^2 is the variance of the forecasts.

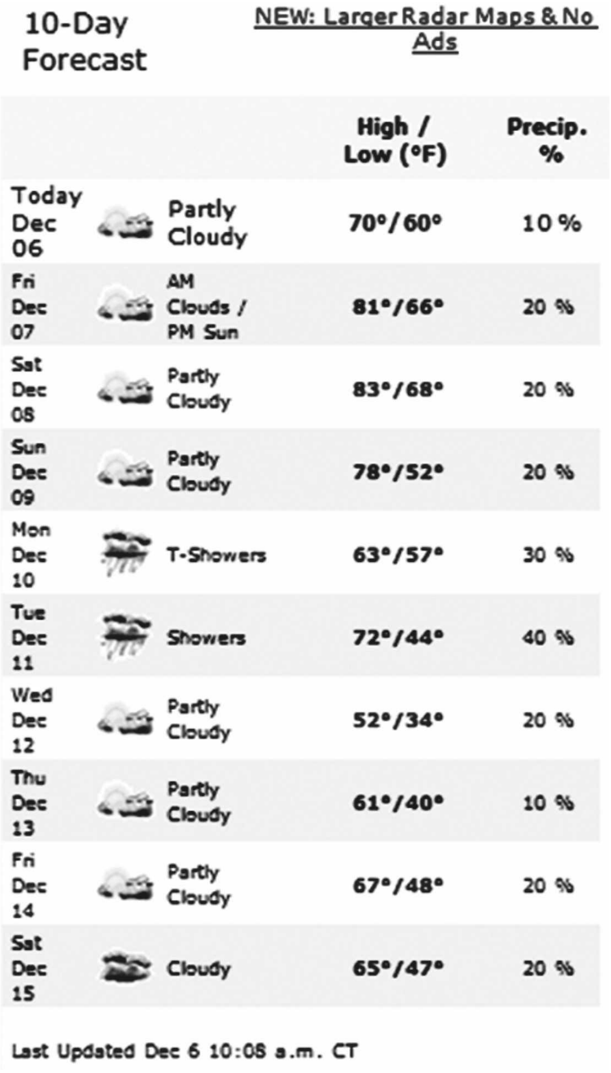
3. Data gathering procedure

a. PoP forecasts

We collected TWC forecasts from 2 November 2004 to 16 January 2006. [These data were collected from http://www.weather.com/, which provides a 10-day forecast that includes forecasts from the same day (0-day forecast) to 9 days ahead.] Figure 1 displays a representative 10-day forecast from 2007. These forecasts are available for any zip code or city and include probability of precipitation, high/low temperature, and verbal descriptions or weather outcomes such as "partly cloudy." The forecasts are updated on a regular basis and are freely available to the public.

TWC's PoP forecasts cover a 12-h window during the daytime (0700–1900 local time), rather than a complete 24-h day. The 12-h PoP is the maximum hourly PoP estimated by TWC during the forecast window. PoPs are rounded and must adhere to local rules relating PoPs to weather outcomes (B. Rose 2007, personal communication).[1]

[1] Bruce Rose is a meteorologist and software designer for TWC based in Atlanta, Georgia. The authors worked closely with Dr. Rose to understand TWC's forecasting process.



10-Day Forecast

NEW: Larger Radar Maps & No Ads

		High / Low (°F)	Precip. %
Today Dec 06	Partly Cloudy	70°/60°	10 %
Fri Dec 07	AM Clouds / PM Sun	81°/66°	20 %
Sat Dec 08	Partly Cloudy	83°/68°	20 %
Sun Dec 09	Partly Cloudy	78°/52°	20 %
Mon Dec 10	T-Showers	63°/57°	30 %
Tue Dec 11	Showers	72°/44°	40 %
Wed Dec 12	Partly Cloudy	52°/34°	20 %
Thu Dec 13	Partly Cloudy	61°/40°	10 %
Fri Dec 14	Partly Cloudy	67°/48°	20 %
Sat Dec 15	Cloudy	65°/47°	20 %

Last Updated Dec 6 10:08 a.m. CT

Fig. 1. Example of 10-day forecast available at the TWC Web site.

We selected 50 locations in the United States, one in each state. Within each state we selected a major city. Within each city we selected the lowest zip code, excluding P.O. boxes. See Table 1 for a list of the cities and zip codes included in this study.

Since TWC's forecasts are not archived, we recorded the forecasts daily. We automated this collection using a Web query and a macro in Microsoft Excel. The macro gathered forecast data directly from Web pages, such as that shown in Fig. 1. This process worked well, but was not completely automatic. In some cases, we experienced temporary problems with certain zip codes (e.g., http://www.weather.com/ data being unavailable) or faced Internet outages. These errors were generally discovered at a point at which forecast data could still be acquired. However, on some days (though fewer than 5%), we were unable to retrieve the PoP forecasts, and these data have been excluded from the analysis. While we did record high and low temperature in addition to PoP, we do not analyze temperature forecasts in this paper.

Because the archival process required intervention, we decided upon a single collection time. This timing is important because the forecasts are updated frequently and not archived. To ensure that we did not collect same-day forecasts before they were posted in our westernmost zip codes (Hawaii and Alaska) we established a collection time of 1130 central standard time (CST), which corresponds to 0730 Hawaii–Aleutian standard time, 0830 Alaska standard time, 0930 Pacific standard time, 1030 mountain standard time, and 1230 eastern standard time (EST). During daylight saving time (DST), we archived forecasts at 1130 central daylight time (CDT; 1030 CST). TWC builds their forecasts at 0100, 0300, 0900, 1100, 1810, and 2300 EST [or eastern daylight time (EDT); B. Rose 2007, personal communication]. These forecasts reach TWC's Web site approximately 15 min later. Therefore, our forecasts represent TWC's view at 1000 CST (or CDT). On rare occasions, TWC amends forecasts during the day, but we do not try to account for this.

b. Precipitation observations

The observed records of daily precipitation and high/low temperature of the current and previous month are available online at the TWC Web site. However, the Web site only archives daily precipitation observations, whereas we require hourly observations because the PoP forecast is for the 12-h window during the daytime. Therefore, we obtained hourly precipitation observation data from the National Climatic Data Center (NCDC; available online at www.ncdc.noaa.gov). Using NCDC's database, we selected the observation station that was closest to our forecast zip code.[2] Table 1 lists the observation stations used in this study and both the distance and elevation difference between the forecast zip code and the observation station. Most stations were within 20 km of the forecast zip code. However, eight stations were more than 20 km from the forecast area (i.e., Alaska, California, Colorado, Idaho, New Mexico, Oklahoma, Pennsylvania, and Vermont). In addition, one forecast–observation pair was separated by more than 500 m in elevation (i.e., Alaska). We have therefore removed these eight locations from our

[2] We considered an NCDC observation of less than 0.01 in. of precipitation as an observation of no precipitation.

TABLE 1. Forecast zip codes and observation stations.

State	City	Forecast zip code	Observation station (call sign)	Distance between forecast and observation (km)	Elev diff between forecast and observation (m)
Alabama	Montgomery	36104	Montgomery Regional Airport (MGM)	13	16
Alaska	Valdez	99686	M. K. Smith Airport (CDV)	72	1571
Arizona	Phoenix	85003	Phoenix Sky Harbor International Airport (PHX)	5	7
Arkansas	Little Rock	72201	Adams Field Airport (LIT)	5	15
California	Stanford	94305	Hayward Executive Airport (HWD)	24	15
Colorado	Denver	80002	Denver International Airport (DEN)	29	11
Connecticut	Hartford	06101	Hartford–Brainard Airport (HFD)	5	7
Delaware	Newark	19702	New Castle County Airport (ILG)	11	2
Florida	Tallahassee	32306	Tallahassee Regional Airport (TLH)	5	2
Georgia	Atlanta	30303	Hartsfield–Jackson Atlanta Intl AP (ATL)	11	12
Hawaii	Honolulu	96813	Honolulu International Airport (HNL)	8	17
Idaho	Idaho Falls	83401	Idaho Falls Regional ARPT (IDA)	32	246
Illinois	Chicago	60601	Chicago Midway International ARPT (MDW)	11	5
Indiana	Indianapolis	46201	Indianapolis International Airport (IND)	16	7
Iowa	Des Moines	50307	Des Moines International Airport (DSM)	6	21
Kansas	Wichita	67202	Colonel James Jabara Airport (AAO)	8	34
Kentucky	Frankfort	40601	Capital City Airport (FFT)	8	10
Louisiana	New Orleans	70112	Louis Armstrong New Orleans Intl AP (MSY)	16	1
Maine	Augusta	04330	Augusta State Airport (AUG)	5	66
Maryland	Baltimore	21201	Baltimore–Washington International Airport (BWI)	13	19
Massachusetts	Cambridge	02139	Logan International Airport (BOS)	8	1
Michigan	Detroit	48201	Detroit City Airport (DET)	6	4
Minnesota	Minneapolis	55401	Minneapolis–St. Paul International AP (MSP)	11	13
Mississippi	Jackson	39201	Jackson International Airport (JAN)	10	17
Missouri	Springfield	65802	Springfield–Branson Regional Airport (SGF)	5	6
Montana	Helena	59601	Helena Regional Airport (HLN)	14	14
Nebraska	Lincoln	68502	Lincoln Municipal Airport (LNK)	8	6
Nevada	Reno	89501	Reno–Tahoe International Airport (RNO)	3	24
New Hampshire	Manchester	03101	Manchester Airport (MHT)	6	14
New Jersey	Trenton	08608	Trenton Mercer Airport (KTTN)	6	45
New Mexico	Santa Fe	87501	Santa Fe Municipal Airport (SAF)	32	215
New York	New York	10001	Central Park (NYC)	5	30
North Carolina	Raleigh	27601	Raleigh–Durham International AP (RDU)	16	39
North Dakota	Fargo	58102	Hector International Airport (FAR)	2	0
Ohio	Columbus	43085	Port Columbus International Airport (CMH)	16	29
Oklahoma	Oklahoma City	73102	Wiley Post Airport (PWA)	21	23
Oregon	Portland	97201	Portland International Airport (PDX)	10	182
Pennsylvania	Pittsburgh	15201	Pittsburgh International Airport (PIT)	24	73
Rhode Island	Providence	02903	T. F. Green State Airport (PVD)	10	9
South Carolina	Charleston	29401	Charleston AFB/International Airport (CHS)	14	9
South Dakota	Sioux Falls	57103	Joe Foss Field Airport (FSD)	3	30
Tennessee	Memphis	38103	Memphis International Airport (MEM)	11	5
Texas	College Station	77843	Easterwood Airport (KCLL)	2	8
Utah	Salt Lake City	84101	Salt Lake City International Airport (SLC)	6	2
Vermont	Newport	05855	Morrisville–Stone St. ARPT (MVL)	48	3
Virginia	Richmond	23219	Richmond International Airport (RIC)	10	3
Washington	Seattle	98101	Seattle–Tacoma International Airport (SEA)	14	68
West Virginia	Charleston	25301	Yeager Airport (CRW)	3	95
Wisconsin	Madison	53703	Dane County Regional–Truax Field Airport (MSN)	6	4
Wyoming	Cheyenne	82001	Cheyenne Airport (CYS)	3	11

TABLE 2. Summary of forecast and observation data.

Lead time (days)	No. of forecasts	Precipitation observations ($x = 1$)	Avg PoP forecast	Frequency of precipitation	ME
0	17 338	3121	0.232	0.180	0.052
1	17 231	3651	0.245	0.212	0.034
2	17 161	3636	0.243	0.212	0.031
3	17 075	3610	0.242	0.211	0.031
4	16 975	3605	0.237	0.212	0.025
5	16 914	3550	0.231	0.210	0.021
6	16 909	3588	0.231	0.212	0.019
7	16 849	3580	0.198	0.212	−0.015
8	16 815	3577	0.265	0.213	0.052
9	15 742	3283	0.230	0.209	0.021

analysis, leaving 42 locations.[3] The average distance and elevation between observation stations and our zip codes for these 42 locations are approximately 7 km and 18 m, respectively. The maximum distance and elevation difference between forecast–observation pairs are 16 km and 181 m, respectively. We also verified that the surface conditions between the observation–forecast pairs for the 42 remaining stations are similar.

The hourly data for each observation station is archived according to local standard time (LST). We used a 12-h observation window from 0700 to 1900 LST for each location to calibrate the PoP forecast data, which corresponds to the TWC's PoP definition. Because the observations are always archived according to LST, during DST we slide our observation window up 1 h (0600–1800 LST) except in Arizona and Hawaii.

The verification of the same day PoP forecasts is more complicated than other PoP forecasts because the timing of the forecast collection determines which hours of observation data should be included. For example, in the eastern time zone, we only want to include precipitation observations between 1100 and 1900 EST (or between 1000 and 1800 EST during DST). Therefore, we removed hourly precipitation observations that occurred before the forecast time for the same-day forecasts at each location.

c. Data summary

Before beginning our analysis, we summarize our forecast and observation data in Table 2. We collected between 15 742 and 17 338 PoP forecasts, depending on the lead time (169 163 PoPs in total). Precipitation was observed approximately 21% of the time. The frequency of precipitation for same-day forecasts is lower (18%) because these forecasts span less than a 12-h window for some time zones. TWC's average PoP forecast varied over the lead times, ranging from a low of 0.198 (7 day) to a high of 0.265 (8 day). All but one lead time exhibits a positive mean error between the forecast and the observation, suggesting some degree of positive bias in TWC's PoP forecasts. The same-day bias is 0.052.

Table 3 details the number of forecasts by PoP and lead time. TWC forecast a 0.2 PoP 4930 times for their same-day forecast. Overall, a 0.0 PoP was forecast 24 382 times, while a PoP of 1.0 was forecast 410 times. The italic values identify forecasts that were made fewer than 40 times, which we exclude from further analysis.[4]

4. Forecast verification

a. Calibration-refinement factorization

Figure 2 displays a calibration or attributes diagram (Hsu and Murphy 1986) for TWC's 0-day PoP forecasts. The line at 45°, labeled "perfect," identifies PoPs that are perfectly calibrated [i.e., $p(1|f, l) = f$]. The horizontal line labeled "no resolution" identifies the case where the frequency of precipitation is independent of the forecast. The line halfway between no resolution and perfect is labeled "no skill." Along this line the skill score is equal to zero and according to Eq. (7), the forecast does not reduce uncertainty in the observation. Points above (below) this line exhibited positive (negative) skill.

[3] In hindsight, we should have selected forecasts that correspond to observation stations. However, we initially thought we would be able to use TWC's observation data, only later realizing that these observations do not cover the same length of time as the forecasts.

[4] A cutoff of 40 is common in hypothesis testing. The variance of a binomial distribution is $Np(1 - p)$. The normal approximation to the binomial is very good when this variance is greater than 10. Thus, if $p = ½$ then N should be greater than 40.

TABLE 3. Number of probability of precipitation forecasts by lead time.

Lead	PoP											
time	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	Subtotal
0	4316	2469	4930	2065	799	909	602	234	606	175	233	17 338
1	4169	2312	4537	2215	907	833	877	272	794	193	122	17 231
2	2285	2989	5435	3366	900	457	936	389	246	113	45	17 161
3	1084	2103	7212	4076	1486	231	720	93	70	*30*	*10*	17 115
4	1047	2164	7215	4116	1570	244	545	74	*29*	*9*	*0*	17 013
5	1053	2395	7106	4152	1541	232	435	*33*	*17*	*5*	*0*	16 969
6	1142	2465	6768	4220	1618	228	468	*13*	*1*	*2*	*0*	16 925
7	2737	3390	5344	3485	1266	63	564	*3*	*0*	*2*	*0*	16 854
8	3395	3271	2907	1810	1255	95	4082	*0*	*0*	*0*	*0*	16 815
9	3154	3456	3155	2218	1348	105	2306	*0*	*0*	*0*	*0*	15 742
Subtotal	24 382	27 014	54 609	31 723	12 690	3397	11 535	1111	1763	529	410	169 163

The gray area in Fig. 2 presents the frequency with which different PoPs are forecast [i.e., $p(f)$]. We notice peaks at PoPs of 0.0 and 0.2, each being forecast more than 20% of the time.

We identified a probability interval around the line of perfect calibration, based on the number of forecasts, which determines whether we identify a PoP as being not well calibrated. Based on the normal approximation to the binomial distribution, we establish a 99% credible interval, in which case there is a 1% chance a forecast–observation pair would be outside this interval (0.5% chance of being above and 0.5% chance of being below). For example, if the PoP was truly f, then there is a 99% chance that the actual relative frequency of precipitation would be within

$$f \pm \Phi^{-1}(0.995)\left[\frac{f(1-f)}{N}\right]^{1/2}, \qquad (9)$$

where Φ^{-1} is the inverse of the standard normal cumulative [$\Phi^{-1}(0.995) = 2.576$] and N is the number of forecasts. This range forms an envelope around the line of perfect calibration, the width of which is determined by Eq. (9). If a forecast–observation pair lies outside this range, then the forecast is not well calibrated.[5] PoPs of 0.0, 0.1, 0.2, 0.3, and 1.0 are not well calibrated. PoPs of 0.0 and 1.0 will not be well calibrated if even a single contrary event occurs, which is a good reason to restrict PoP forecasts to the open interval (0, 1).

The 0.3 PoP is not well calibrated and exhibits no skill. PoPs below 0.3 are quite poor: they are miscalibrated, exhibit negative skill, and are biased. For example, when TWC forecast a 0.2 chance of precipita-

[5] This is identical to a two-tailed t test with a 1% level of significance.

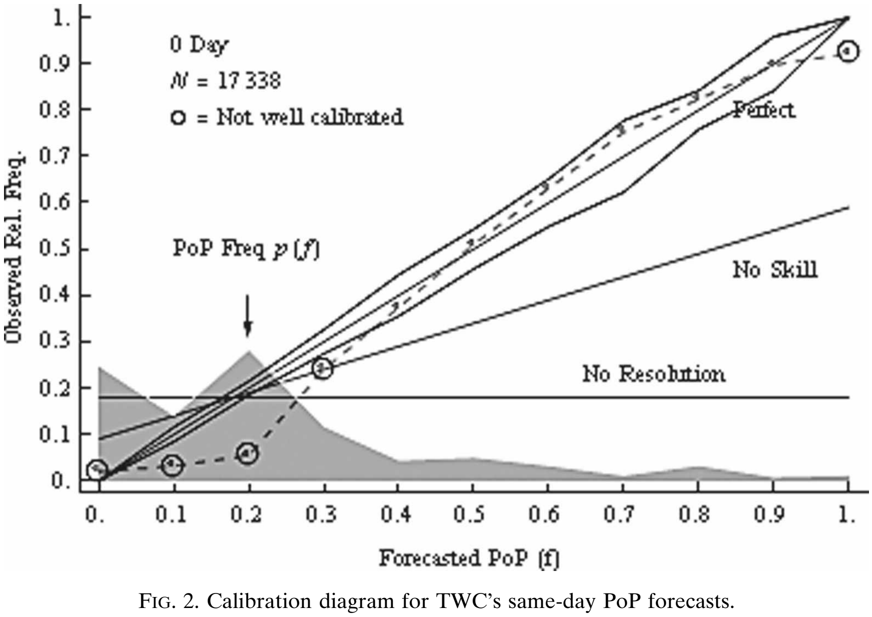


FIG. 2. Calibration diagram for TWC's same-day PoP forecasts.

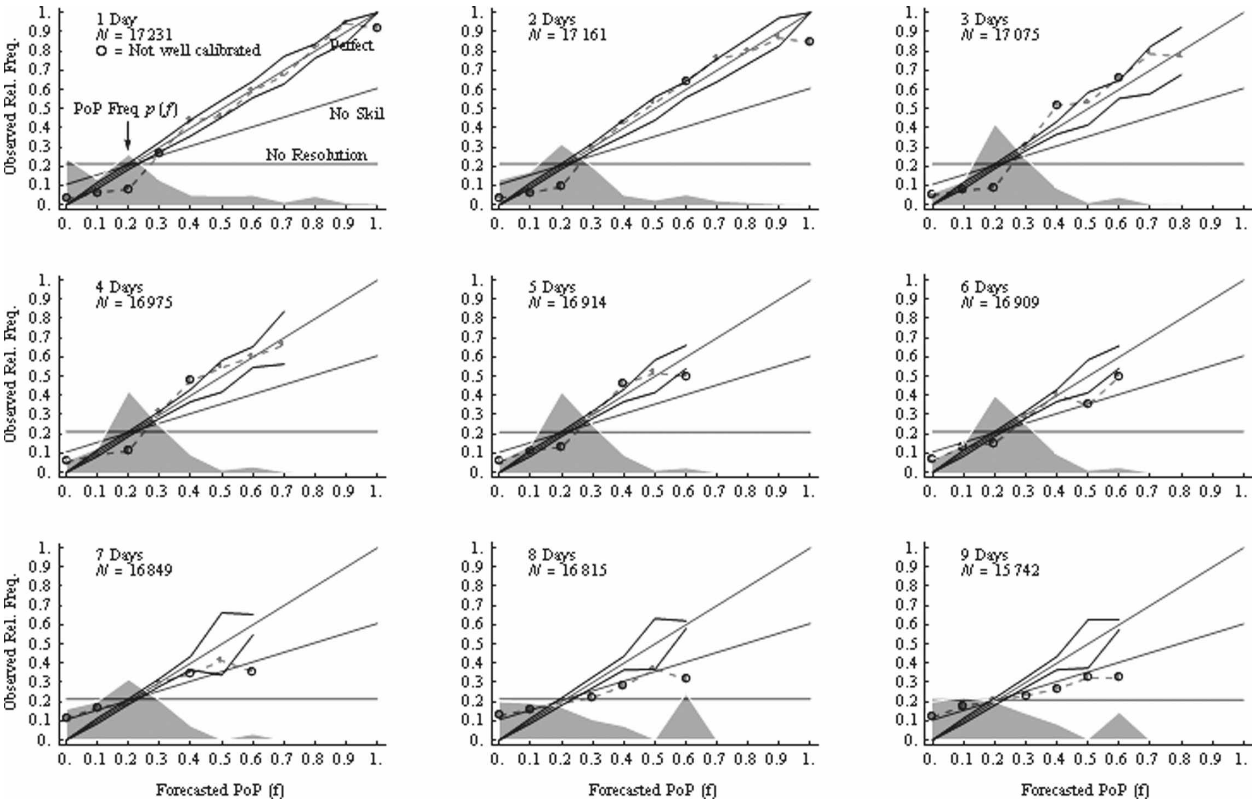


FIG. 3. Calibration diagrams for 1–9-day lead times.

tion for the same day, precipitation occurred only 5.5% of the time.

PoPs of 0.4 and above, excluding 1.0, can be taken at face value and used directly in decision making. However, PoPs of 0.3 and below or 1.0 require adjustment—sometimes significant.

Figure 3 presents the calibration diagrams for lead times of 1–9 days. The 1-day forecasts exhibit the same behavior as the 0-day forecasts: PoPs from 0.0 to 0.2 and 1.0 are miscalibrated. The calibration of midrange PoPs begins to degrade with lead time. Performance decreases markedly beginning with the 7-day forecasts. For example, most of the PoP forecasts lay along the no skill line for lead times of 7 days or longer. While predictability does decrease with lead time, calibration performance should not; a forecast of f should occur $f \times 100\%$ of the time whether it was a forecast for the next hour or the next year.

These phenomena can be explained in part by TWC's forecasting procedure (B. Rose 2007, personal communication). The meteorologists at TWC receive guidance from a mixture of numerical, statistical, and climatological inputs provided by computer systems. The human forecasters rarely intervene in forecasts beyond 6 days. Thus, the verification results of the 7–9-day forecasts represent the "objective" machine guidance being provided to TWC's human forecasters. In this respect, the human forecasters appear to add considerable skill, since the 0–6-day calibration performance is so much better.

However, when humans do intervene, they introduce considerable bias into the low-end PoP forecasts. This bias could be a by-product of the intervention tools used by the human forecasters. The forecasters do not directly adjust the PoPs, but instead change what is known as the sensible weather forecast. For example, they might change partly cloudy to "isolated thunder." When this change is made, a computer algorithm determines the "smallest" change that must be made in a vector of weather parameters to make them consistent with the sensible weather forecast. A PoP of 29% is the cutoff for a dry forecast and therefore, it appears as though this intervention tool treats all "dry" PoPs as being nearly equivalent. This also might explain the curious dip in forecast frequency at 0.1 in both the 0- and 1-day forecasts.

The frequency of forecasts highlights additional challenges with the machine guidance. The most likely 8- and 9-day forecasts are 0.0 and 0.6, with a forecast of 0.5 being very unlikely. TWC appears to avoid forecasts

TABLE 4. Summary measures of forecasting performance at different lead times.

Lead		Variance			Skill
time	MSE	Forecasts	Observations	Correlation	score
0	0.095	0.053	0.148	0.615	35.9%
1	0.113	0.055	0.167	0.575	32.4%
2	0.127	0.036	0.167	0.499	24.2%
3	0.140	0.019	0.167	0.416	16.1%
4	0.147	0.016	0.167	0.352	11.8%
5	0.152	0.014	0.166	0.289	8.1%
6	0.158	0.015	0.167	0.243	5.4%
7	0.167	0.019	0.167	0.177	0.4%
8	0.188	0.049	0.167	0.176	−12.0%
9	0.179	0.038	0.165	0.158	−8.2%

of 0.5. We can even see the "ghost" of the 0.6 peak in the shorter-term human-adjusted forecasts. Forecasts as extreme as 0.0 or 0.6 are difficult to justify far into the future. For example, the frequency of precipitation conditional on the forecast ranges from 0.12 to 0.32 for the 9-day forecast. It appears that TWC's forecasts would need to be constrained to this range if they were intended to be well calibrated.

Table 4 presents several summary measures of forecasting performance. The mean-square error [Eq. (5)] ranges from 0.095 to 0.188. The variance of the forecasts is less than the variance of the observations, but much less stable. The correlation between the forecasts and the observations begins at 0.615 and declines quickly with lead time. The same-day skill score is approximately 36% and declines with lead time. The 8- and 9-day computer forecasts exhibit negative skill—using the computer forecasts directly induces more error than using climatology. For comparison, Murphy and Winkler (1977) found an overall SS for a sample of National Weather Service forecasts, averaged over all lead times, of approximately 31%.

b. *Likelihood-base-rate factorization*

Figure 4 displays the likelihood functions (or discrimination plots), $p(f|1, l)$ and $p(f|0, l)$ for TWC's 0-day PoP forecasts. Given that precipitation did not occur, it is likely TWC forecast a PoP of either 0.0 or 0.2. Likewise, it is unlikely that PoPs greater than 0.6 were forecast in this situation. However, if precipitation did occur, a range of PoPs from 0.3 to 0.8 were almost equally likely to have been forecast. Ideally, one would hope to see $p(f|1, l)$ peak at high PoPs and decline to the left.

Figure 5 displays likelihoods for the remainder of lead times. The degree of overlap between the likelihood functions increases rapidly with lead time, as the forecasts lose their ability to discriminate and skill scores fall. The peaks at a PoP of 0.6 are even more pronounced in the likelihood graphs.

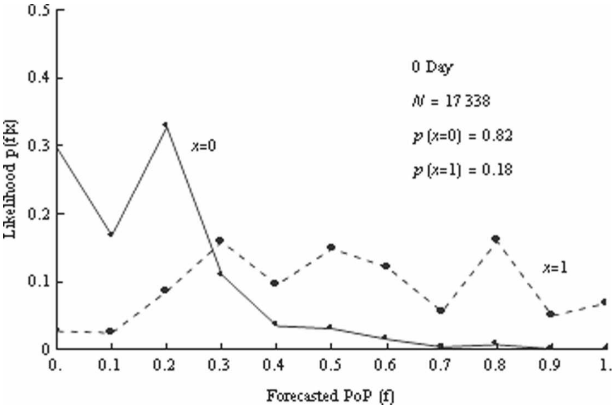


FIG. 4. Likelihood functions for TWC same-day forecasts.

c. *Warm and cool seasons*

Following Murphy and Winkler (1992), we gain additional insight into TWC's forecasts by analyzing their performance during warm (April–September) and cool (October–March) months. Table 5 summarizes the forecast and observation data by season. Approximately 60% of our dataset covers the cool season because we gathered data from 2 November 2004 to 16 January 2006. The sum of the number of forecasts for the cool and warm seasons is lower than the totals presented in Table 2 because we have excluded PoPs that were forecast fewer than 40 times. For example, a same-day PoP of 0.9 was forecast only 26 times during the warm-season and has therefore been excluded from the warm-season analysis (17 388 − 10 374 − 6938 = 26).

The frequency of precipitation was lower during the warm season than during the cool season. Yet, TWC forecast higher PoPs during the warm season, resulting in a larger mean error. For example, the 0-day warm season PoP was 0.086 too high on average.

Figure 6 compares the 0-day PoP calibration in the cool and warm seasons. The most likely forecast in the cool season was 0.0, even though precipitation occurred more frequently than during the warm season. The cool season is not well calibrated for low (0.0–0.2) or high (0.8–1.0) PoPs, whereas the lower half of the PoP range performs poorly during the warm season—TWC overforecasts PoPs below 0.5 during the warm season. Overall, the warm season is not as well calibrated as the cool.

Figure 7 contrasts the cool and warm calibration for 1–9-day forecasts. The calibration performance between the two seasons is similar. However, the cool-season PoPs tend to be sharper because they forecast

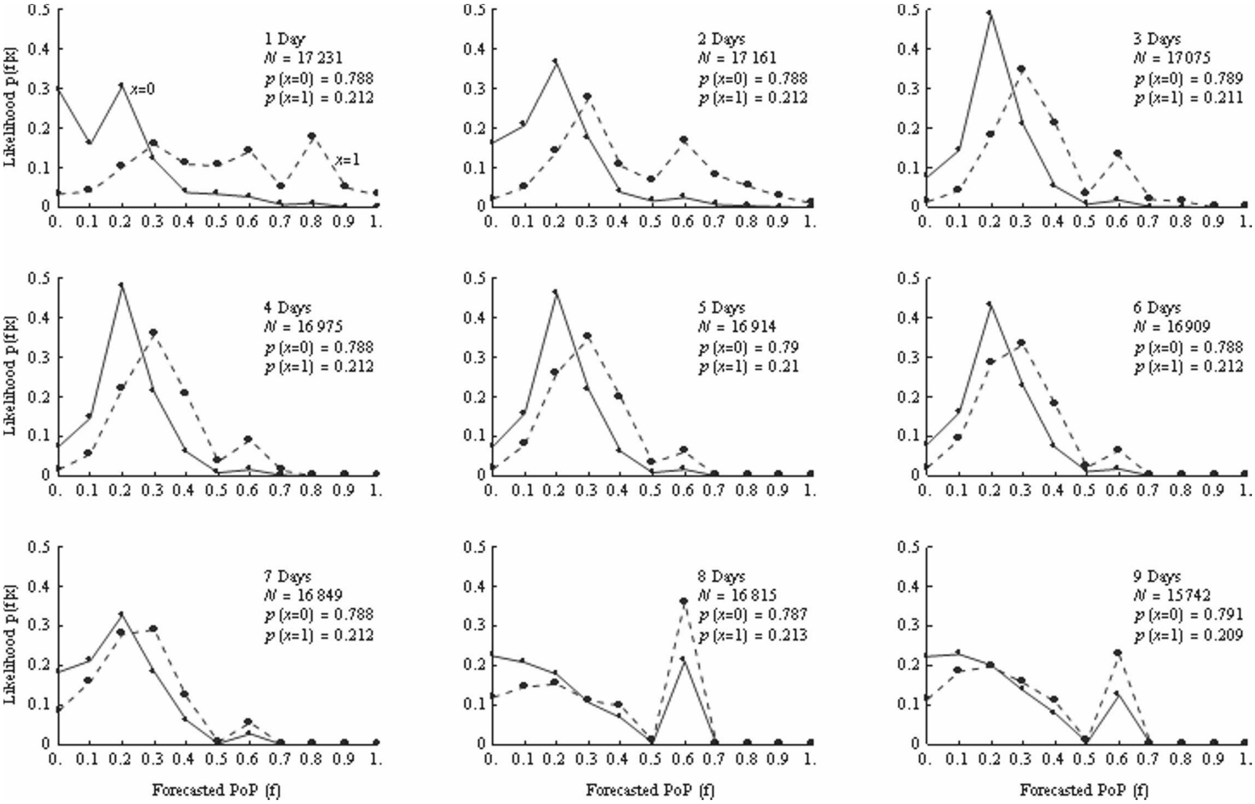


FIG. 5. Likelihood diagrams for 1–9-day lead times.

TABLE 5. Summary of forecast and observation data for cool and warm seasons.

Cool season

Lead time (days)	No. of forecasts	Precipitation observations ($x = 1$)	Avg PoP forecast	Frequency of precipitation	ME
0	10 374	1877	0.210	0.181	0.029
1	10 374	2232	0.229	0.215	0.014
2	10 296	2204	0.231	0.214	0.017
3	10 256	2212	0.237	0.216	0.022
4	10 216	2196	0.232	0.215	0.017
5	10 170	2164	0.224	0.213	0.011
6	10 149	2201	0.225	0.217	0.008
7	10 117	2199	0.190	0.217	−0.027
8	10 080	2188	0.243	0.217	0.026
9	8998	1904	0.239	0.212	0.027

Warm season

Lead time (days)	No. of forecasts	Precipitation observations ($x = 1$)	Avg PoP forecast	Frequency of precipitation	ME
0	6938	1222	0.262	0.176	0.086
1	6765	1341	0.262	0.198	0.064
2	6799	1380	0.252	0.203	0.049
3	6789	1381	0.248	0.203	0.044
4	6745	1404	0.244	0.208	0.036
5	6744	1386	0.240	0.206	0.035
6	6760	1387	0.240	0.205	0.035
7	6722	1377	0.209	0.205	0.004
8	6695	1373	0.296	0.205	0.091
9	6709	1371	0.216	0.204	0.012

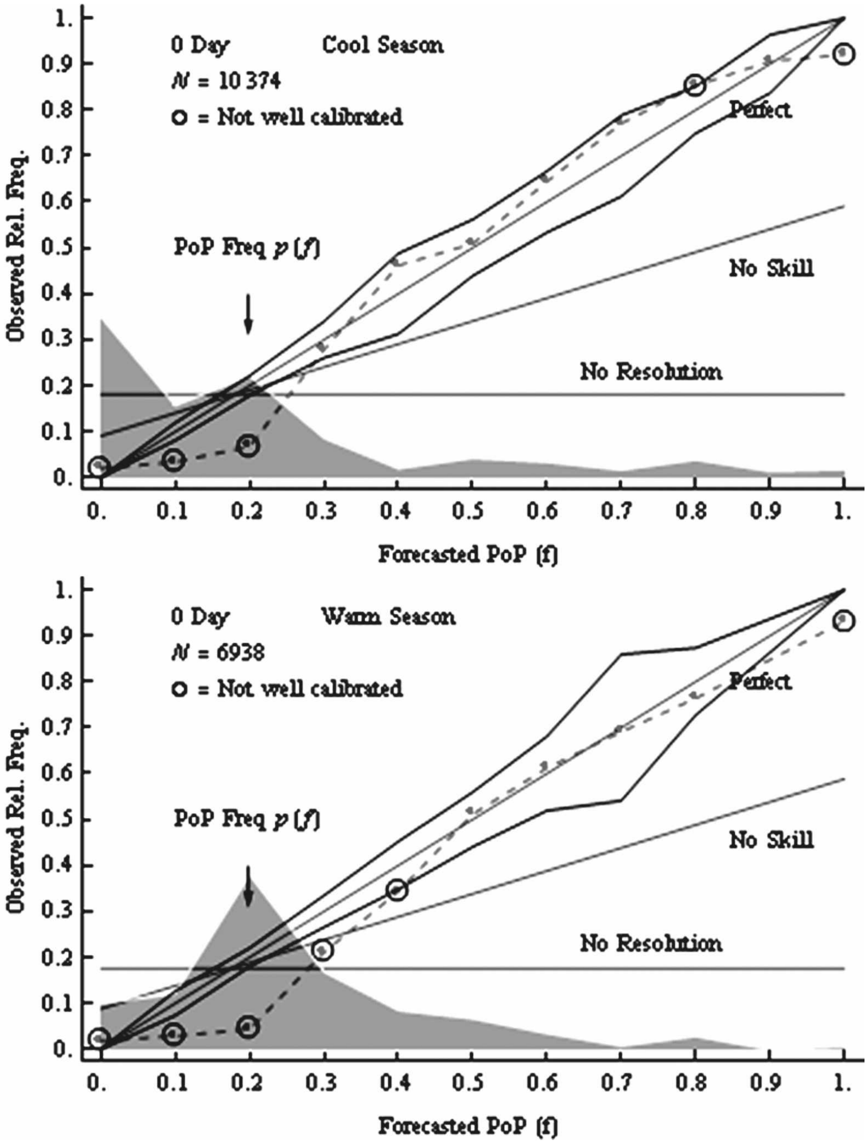


FIG. 6. Same-day PoP calibration in warm and cool seasons.

0.0 more frequently. One noticeable difference in forecast behavior is the increased frequency of 0.3 PoPs during the warm season.

Table 6 compares the skill scores and correlations between the two seasons. Warm-season forecasts are about half as skillful as the cool season. Cool-season skill scores begin at about 44% and decline to 0% by day 7. Warm-season skill scores are about 50% lower. For comparison, Murphy and Winkler (1992) found skill scores of 57%, 38%, and 30% for the 0-, 1-, and 2-day forecasts during the cool season and 37%, 24%, and 21% during the warm season, respectively. TWC's performance is on par with these earlier studies in the cool season, if somewhat worse for same-day forecasts. Warm-season performance appears to lag previous studies.

We can better understand the drivers of the difference between warm and cool seasons by decomposing the MSE given in Eq. (5) as follows (Murphy and Winkler 1992):

$$\mathrm{MSE}(f, x|l) = \sigma_x^2 + E_{F|l}[f - p(x|f, l)]^2 - E_{F|l}[\bar{x}_l - p(x|f, l)]^2. \qquad (10)$$

The second term on the rhs of (10) is a measure of calibration or refinement. The last term is the resolution (Murphy and Daan 1985). Figure 8 plots the MSE for the cool and warm seasons according to this factor-

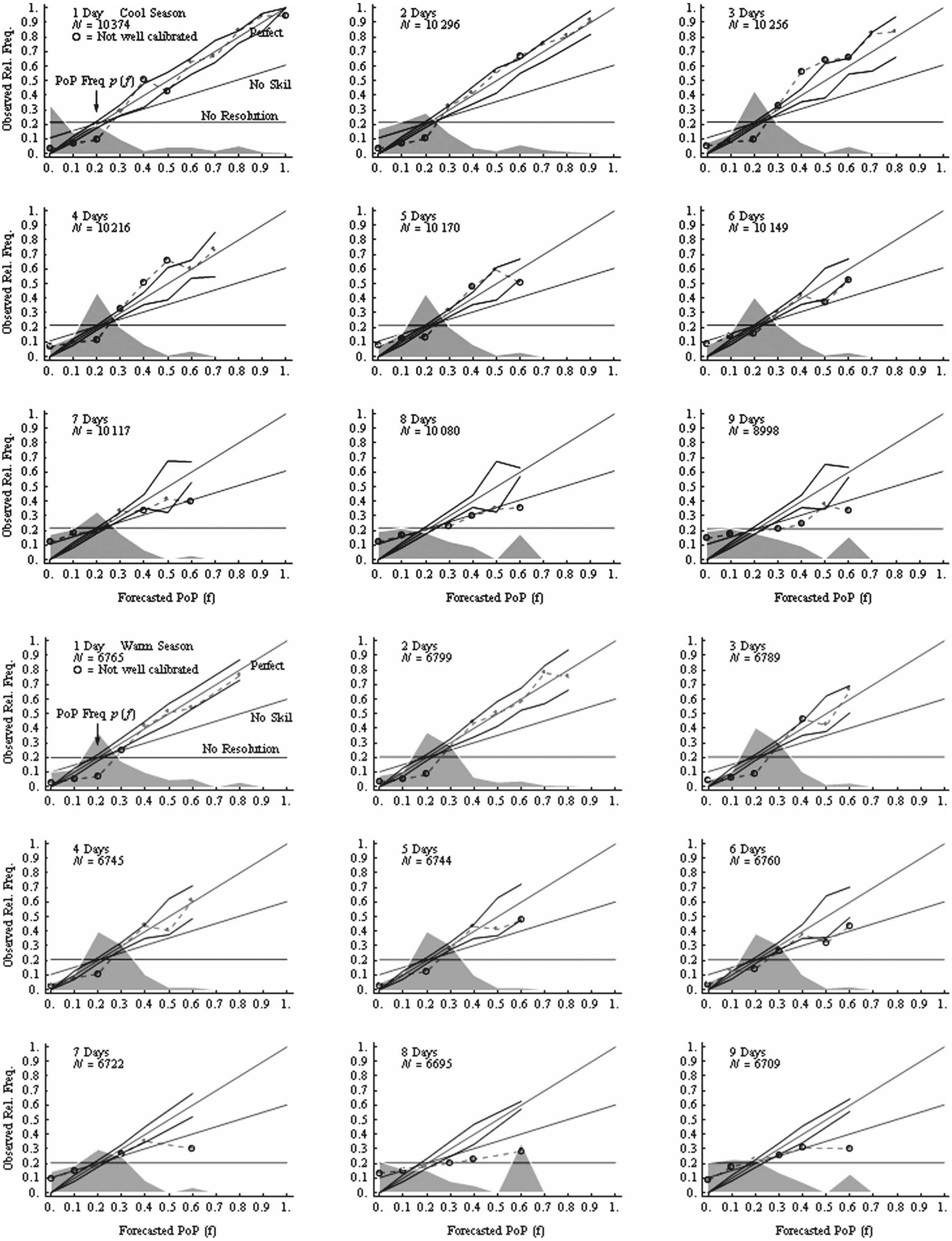


FIG. 7. Comparison of PoP calibration in (top three rows) cool and (bottom three rows) warm seasons for 1–9-day lead times.

TABLE 6. Comparison of cool- and warm-season summary measures.

Cool season					
		Variance			
Lead time	MSE	Forecasts	Observations	Correlation	Skill score
0	0.083	0.064	0.148	0.670	44.3%
1	0.103	0.067	0.169	0.627	39.2%
2	0.119	0.042	0.168	0.543	29.1%
3	0.137	0.021	0.169	0.450	19.0%
4	0.146	0.018	0.169	0.376	13.7%
5	0.153	0.016	0.168	0.300	8.9%
6	0.159	0.016	0.170	0.254	6.1%
7	0.169	0.019	0.170	0.184	0.7%
8	0.179	0.042	0.170	0.199	−5.3%
9	0.183	0.040	0.167	0.150	−9.8%

Warm season					
		Variance			
Lead time	MSE	Forecasts	Observations	Correlation	Skill score
0	0.112	0.034	0.145	0.527	22.5%
1	0.128	0.032	0.159	0.468	19.4%
2	0.137	0.022	0.162	0.413	15.4%
3	0.144	0.013	0.162	0.360	11.2%
4	0.150	0.013	0.165	0.317	9.1%
5	0.152	0.012	0.163	0.276	6.9%
6	0.156	0.013	0.163	0.229	4.1%
7	0.163	0.019	0.163	0.171	−0.2%
8	0.200	0.058	0.163	0.152	−22.7%
9	0.172	0.035	0.163	0.170	−5.9%

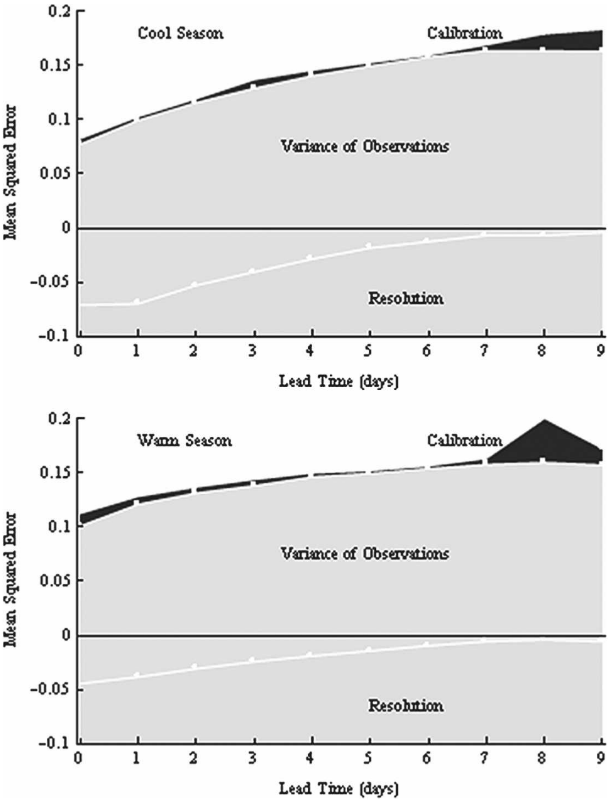


FIG. 8. MSE decomposition for cool and warm seasons.

ization. Note that we have displayed the negative of the resolution (the lowest area) so that higher resolution lowers the MSE, as in Eq. (10). We see that cool-season forecasts have better resolution (more negative) than the warm season. In addition the cool season exhibits better calibration for near-term (2 days or less) and long-term (7 days or more) PoP forecasts. The variance of the observations is slightly lower in the warm season.

The best measure of a probability distribution's sharpness is its entropy H (Cover and Thomas 1991), which is given by

$$H(p) = -\sum_i p_i \log(p_i). \qquad (11)$$

The logarithm can be to any base, but we will use base 2. Entropy is at a minimum in the case of certainty and at a maximum when the probabilities are uniform. In the case of binary forecasts, the maximum entropy is $\log_2(2) = 1$. Entropy can also be thought of as a measure of the amount of information contained in a probability assessment, with lower entropies conveying greater information content.

Suppose a forecaster provides a PoP of f. The entropy of this forecast is $-[f \log_2(f) + (1 - f) \log_2(1 - f)]$. We can therefore associate an entropy to each of TWC's forecasts. Figure 9 plots the average entropy of TWC forecasts for the cool and warm seasons as a function of lead time. In addition, the entropy of a climatological forecast, based on Table 5, is also displayed. In the case of the cool season, we see that TWC forecasts have less entropy (more information) than climatology. The 0- and 1-day forecasts are much narrower than forecasts based solely on climatology because a PoP of 0.0 is forecast often. Entropy increases with lead time as one would expect, but suddenly drops for lead times of 7–9 days. Because these forecasts are not calibrated, we see this drop in entropy as not a result of superior information. Rather, the long-term forecasts are too sharp. The warm season entropies are closer to climatology, but also drop significantly after 6 days.

The 0-day likelihood functions for the cool and warm seasons are compared in Fig. 10. Given that precipitation was not observed, the most likely forecast during the cool season was 0.0, whereas it was 0.2 during the warm season. If precipitation was observed, it was much more likely that a lower PoP was forecast during the warm season than during the cool season. We also notice peaks at 0.8 in the event of precipitation. Figure 11 compares the likelihoods for the remaining

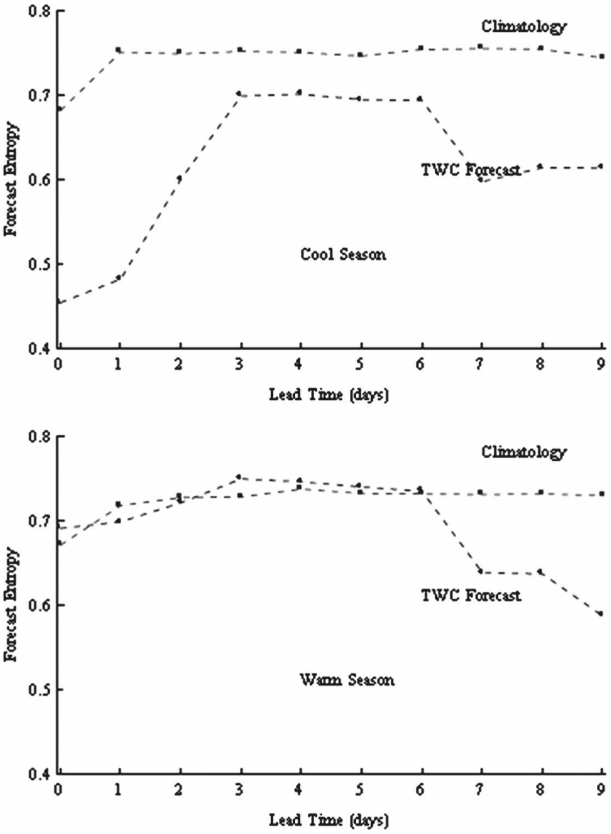


FIG. 9. Forecast entropy for cool and warm seasons.

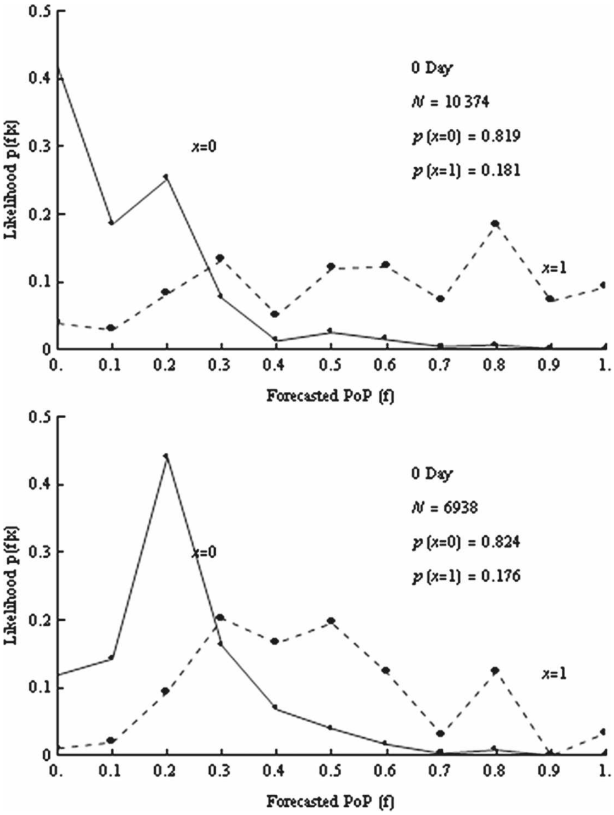


FIG. 10. Cool- and warm-season same-day likelihood functions.

lead times. The overlap between the likelihood functions is greater during the warm season. We also observe peaks at particular probabilities. For example, if precipitation occurred during the warm season, it is almost certain that TWC did not forecast a PoP of 0.7 1-day ahead. Likewise, the 0.6 peaks are prominent in both seasons. Again, one would hope to see the likelihood function given precipitation peak at high PoPs and monotonically decline to the left. TWC's forecasts are good at identifying a lack of precipitation, but are not particularly strong at identifying precipitation—especially during the warm season.

5. Conclusions

TWC's forecasts exhibit positive skill for lead times less than 7 days. Midrange PoPs tend to be well calibrated, but performance decreases with lead time and worsens during the warm season. PoPs below 0.3 and above 0.9 are miscalibrated and biased. Overall, almost all lead times exhibit positive bias and the same-day bias is significant, especially during the warm season.

As discussed previously, there is no reason, per se, that calibration performance should decrease with lead time. Rather, the difficulty of the forecasting task should be reflected in the sharpness of the forecasts. TWC's long-term forecasts are too sharp. Apparently, one cannot reasonably forecast a 0% or 60% chance of precipitation 8 or 9 days from now, much less provide these forecasts nearly 40% of the time.

There seem to be two primary areas in which TWC could improve its forecasts: the machine guidance provided to human forecasters and the intervention tool used by these forecasters to arrive at sensible forecasts. The long-term forecasts, which are unedited by humans, exhibit a tendency to provide extreme forecasts and to artificially avoid 0.5. Perhaps revisions/additions to these models could improve performance. If not, TWC might want to consider intervening in these forecasts as well. The intervention of human forecasters increases skill, but also introduces bias. The intervention tool uses a least squares procedure to adjust underlying weather variables. Perhaps other approaches, such as the application of maximum entropy techniques (Jaynes 1957), would improve performance. Maximum entropy techniques would avoid producing narrow and biased forecasts.

Performance during the warm season is noticeably worse; even though the variance of the observations is

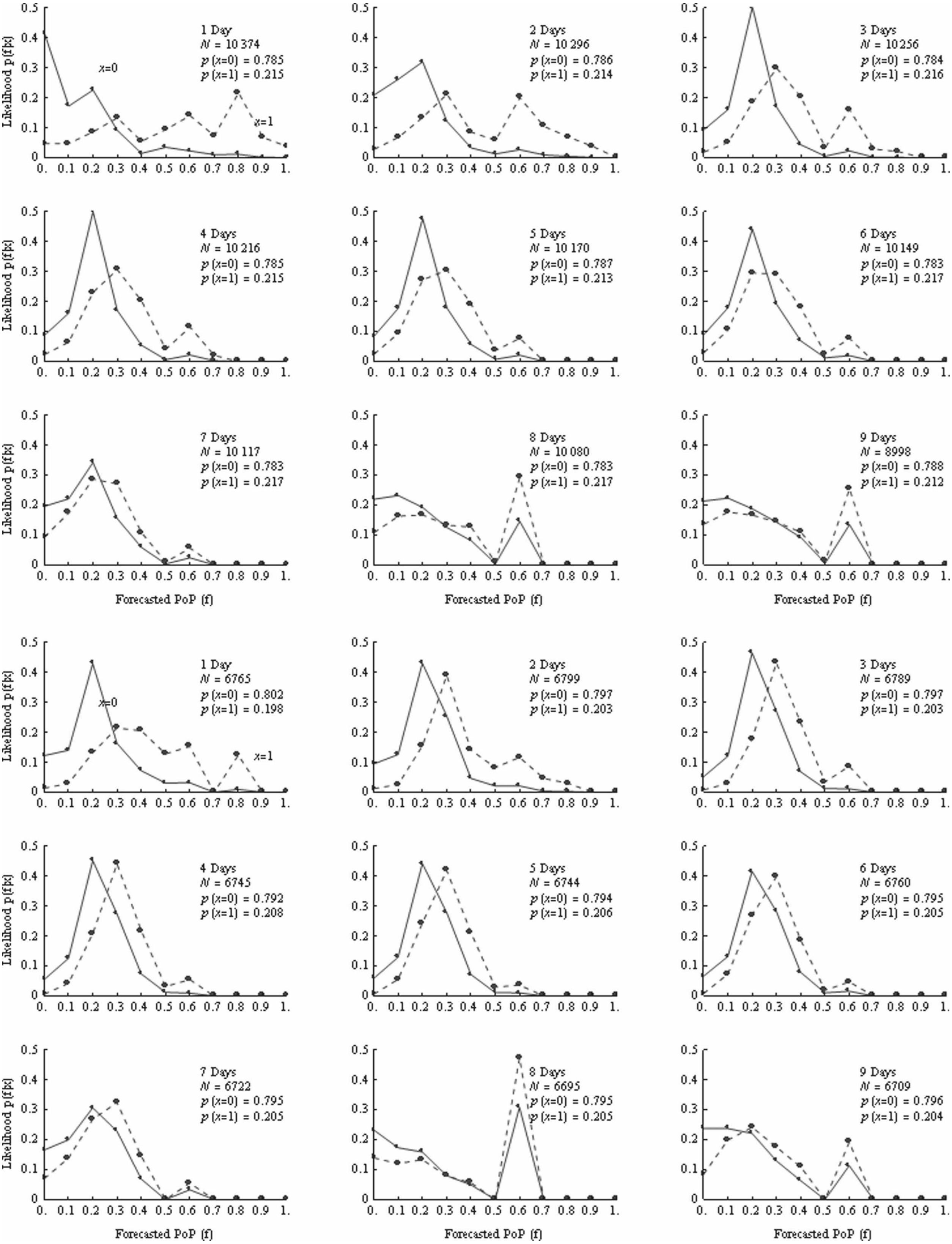


FIG. 11. Likelihood functions for cool and warm season for 1–9-day lead times.

lower (see Table 6). This suggests that TWC should concentrate its attention on improving PoPs during this time.

In addition, providing PoPs at 0.05 intervals (i.e., 0.05, 0.10, . . . , 0.95) might be helpful and enable TWC to avoid forecasts of 0.0 and 1.0, which will not be well calibrated.

Acknowledgments. The authors gratefully acknowledge the assistance and guidance of Dr. Bruce Rose at The Weather Channel and the suggestions of two anonymous reviewers.

REFERENCES

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

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory.* John Wiley and Sons, 542 pp.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

Jaynes, E. T., 1957: Information theory and statistical mechanics. *Phys. Rev.,* **106,** 620–630.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 240 pp.

Katz, R. W., and A. H. Murphy, Eds., 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 222 pp.

Murphy, A. H., and R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Stat.,* **26,** 41–47.

——, and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy, and R. W. Katz, Eds., Westview Press, Inc., 379–437.

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

——, and ——, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.