Comparing NWS PoP Forecasts to Third-Party Providers

J. ERIC BICKEL

The University of Texas at Austin, Austin, Texas

ERIC FLOEHR

Intellovations, LLC, Columbus, Ohio

SEONG DAE KIM

University of Alaska Anchorage, Anchorage, Alaska

(Manuscript received 27 May 2010, in final form 14 January 2011)

ABSTRACT

In this paper, the authors verify probability of precipitation (PoP) forecasts provided by the National Weather Service (NWS), The Weather Channel (TWC), and CustomWeather (CW). The *n*-day-ahead forecasts, where *n* ranges from 1 to 3 for the NWS, from 1 to 9 for TWC, and from 1 to 14 for CW, are analyzed. The dataset includes almost 13 million PoP forecasts, or about 500 000 PoPs per provider per day of lead time. Data were collected over a 2-yr period (1 November 2008–31 October 2010) at 734 observation stations across the contiguous United States. In addition to verifying these PoP forecasts in an absolute sense, relative comparisons are made between the providers. It is found that, in most cases, TWC does not add skill to NWS forecasts. Perhaps most interestingly, it is also found that CW does have the ability to forecast PoPs at a resolution of 0.01.

1. Introduction

Bickel and Kim (2008, hereafter BK08) analyzed approximately 169 000 probability of precipitation (PoP) forecasts provided by The Weather Channel (TWC) over a 14-month period, spanning 2004–06 at 42 U.S. locations. They found that TWC's near-term forecasts (less than a 3-day lead time) were relatively well calibrated, while longer-term forecasts were less reliable. This performance was driven by TWC's forecasting policies and tools. For example, it was found that TWC artificially avoids PoPs of 0.5.

In this paper, we use a much larger dataset than BK08 to analyze and compare the reliability of PoP forecasts provided by the National Weather Service (NWS), CustomWeather (CW), and TWC. Specifically, we analyze almost 13 million PoPs covering a 24-month period (1 November 2008–31 October 2010) at 734 stations across the contiguous United States. Our analysis confirms the results of BK08 and extends their analysis in four important respects. First, as mentioned, we use a much larger dataset, both in terms of geographic locations and observations per location. Second, we provide verification results for two additional providers of PoP forecasts, including the NWS. Third, we analyze whether third-party forecasts are more skilled than those of the NWS. Finally, we segment our results by the National Oceanic and Atmospheric Administration (NOAA) climate regions.

TWC is the leading provider of weather information to the general public via its cable television network and interactive Web site (see online at http://www.weather.com/). TWC's cable network is available in 97% of cable-TV homes in the United States and reaches more than 99 million households. The Internet site, providing weather forecasts for 100 000 locations worldwide, averages over 41 million unique users per month and is the most popular source of online weather, news and information, according to Nielsen/NetRatings (more information available online at http://press.weather.com/company.asp).

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

DOI: 10.1175/2011MWR3525.1

CustomWeather, Inc., is a San Francisco–based provider of syndicated weather content. They generate local weather forecasts for over 200 countries worldwide, establishing it as the industry leader for global location-based coverage at both the U.S. and international levels. CustomWeather provides sophisticated weather products to leading companies in a variety of industries including media, energy, travel, wireless, and the Web.

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 of this study. Finally, in section 5 we present our conclusions.

2. Verification of probability forecasts

The forecast verification literature is extensive. See Katz and Murphy (1997) and Jolliffe and Stephenson (2003) for excellent overviews. In this paper, we adopt the distribution-oriented framework proposed by Murphy and Winkler (1987, 1992). This framework was described in BK08, but is repeated here for convenience.

a. Distributional measures

Let F be the finite set of possible PoP forecasts and let $f \in [0, 1]$ be a particular forecasted value. In practice, forecasts are given in discrete intervals, 0.1 being common and used by the NWS and TWC. On the other hand, CW provides PoP forecasts at 0.01 intervals.

Here X is the set of possible precipitation observations, while x denotes a particular observation. We assume that x may obtain only the value 1 in the event of precipitation and 0, otherwise.

Both F and X are random variables and their empirical frequency distribution, given a particular lead time l, is denoted $p(f, x|l)$. To simplify the exposition we will state lead time in days, rather than hours. This distribution completely describes the performance of the forecasting system. For example, a perfect forecasting system would ensure that $p(f, x|l) = 0$, when $f \neq x$.

Lead times for TWC may obtain integer values ranging from 1 (1-day ahead) to 9 (the last day in a 10-day forecast). BK08 also analyzed TWC's same-day forecast, but we do not consider these forecasts in this paper. In the case of the NWS, we analyze lead times from 1 to 3 days. CustomWeather provides PoPs from 1 to 14 days ahead.

Since

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

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. The second term $p(x|f, l)$ is the conditional distribution of the observation given the forecast. For example, $p(x = 1|f, l)$ is the relative frequency of precipitation when the forecast was f and is equal to the conditional mean $\mu_{X|F,l} = E_X[X|F = f, l]$, where E_X denotes the expectation taken with respect to X. 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 $p(x = 1|f) = f$ 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, that is, the forecasts are categorical. Forecasting the climatological average or base rate will be well calibrated, but not sharp. Likewise, perfectly sharp forecasts will generally 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. For example, $p(f|x = 1, l)$ is the relative frequency of forecasts when precipitation occurred, and $p(f|x = 0, l)$ 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

$$\mu_{F|l} = E_{F|l}[F] = \sum_F fp(f|l).$$

Likewise, the sample climatological frequency of precipitation, indexed by lead time, is

$$\mu_{X|l} = E_{X|l}[X] = \sum_X xp(x|l).$$

The mean error (ME) is

$$\mathrm{ME}(F, X|l) = \mu_{F|l} - \mu_{X|l} \qquad (2)$$

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

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

The MSE can be factored as follows (Murphy and Winkler 1992):

$$\mathrm{MSE}(F, X|l) = \sigma^2_{X|l} + E_{F|l}[(\mu_{X|F,l} - F)^2] - E_{F|l}[(\mu_{X|F,l} - \mu_{X|l})^2]. \quad (3)$$

The first term is the variance of the observations and is a measure of forecast difficulty. Since X is binary, $\sigma^2_{X|l} = p(x = 1|l)p(x = 0|l)$. The second term is an overall measure of forecast reliability or conditional basis (conditional on the forecast). The last term is the resolution (Murphy and Daan 1985) and measures the degree to which the conditional forecasts deviate from the unconditional frequency of precipitation. Reliability and resolution are under the control of the forecaster, while the variance of the observations is not.

The climatological skill score (SS) is

$$\mathrm{SS}(F, X|l) = 1 - \mathrm{MSE}(F, X|l)/\mathrm{MSE}(\mu_{X|l}, X|l). \quad (4)$$

Since

$$\mathrm{MSE}(\mu_{X|l}, X|l) = E_{F,X|l}[(\mu_{X|l} - X)^2] = \sigma^2_{X|l},$$

the SS can be written as

$$\mathrm{SS}(F, X|l) = \frac{\sigma^2_{X|l} - \mathrm{MSE}(F, X|l)}{\sigma^2_{X|l}}, \quad (5)$$

and we see that SS measures the proportional amount by which the forecast reduces our uncertainty regarding precipitation, as measured by variance. The SS may also be written as the sum of resolution and reliability (positively oriented), normalized for forecast difficulty (Toth et al. 2003). Specifically, substituting Eq. (3) into Eq. (5) we have

$$\mathrm{SS}(F, X|l) = \mathrm{SS}_{\mathrm{Res}} + \mathrm{SS}_{\mathrm{Rel}} = \frac{E_{F|l}[(\mu_{X|F,l} - \mu_{X|l})^2]}{\sigma^2_{X|l}} + \left\{-\frac{E_{F|l}[(\mu_{X|F,l} - F)^2]}{\sigma^2_{X|l}}\right\}. \quad (6)$$

Thus, SS will be positive when the reward for resolution exceeds the penalty for miscalibration. If the forecasts are perfectly reliable, perhaps after transformation, then $\mu_{X|F} = f$ and $\mu_X = \mu_F$ and

$$\mathrm{SS}(F, X|l) = \frac{E_{F|l}[(F - \mu_{F|l})^2]}{\sigma^2_{X|l}} = \frac{\sigma^2_{F|l}}{\sigma^2_{X|l}}. \quad (7)$$

Thus, in the case of perfectly calibrated forecasts, the SS is the ratio of the variance of the forecasts to the variance of the observations. The variance of the forecasts is a measure of forecast sharpness.

It is important to note that ranking providers using MSE instead of SS may not yield the same results if their forecast windows differ. For example, as will be explained below, CW provides a 24-h (24 h) PoP, while both the NWS and TWC provide 12-h PoPs. In the dataset we consider, the variance of observations for a 24-h window is larger than the variance of observations for a 12-h window. Thus, CW's forecasting task is more difficult and they will have higher a MSE even if they are just as skilled as the other providers. To correct for this, SS normalizes by the variance of the observations, which, again, is a measure of forecast difficulty.

3. Data-gathering procedure

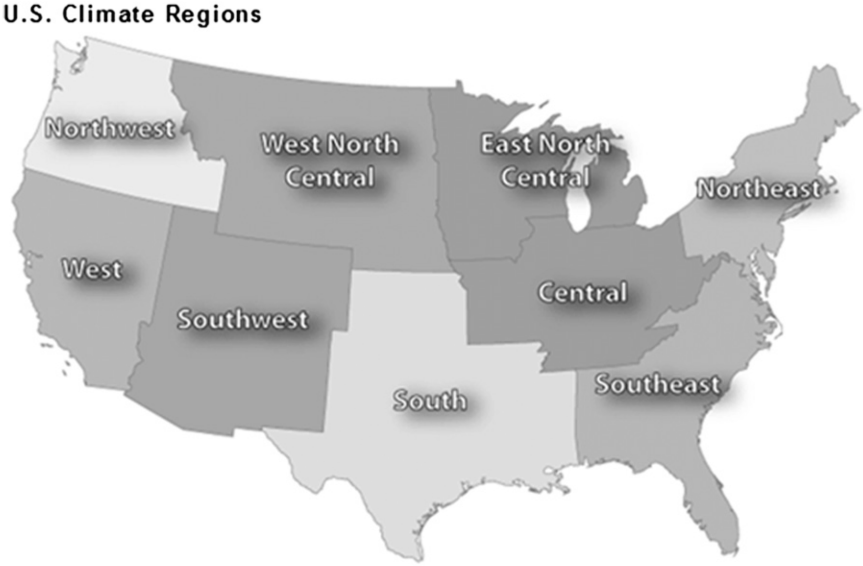
We gathered forecasts and observations for nine NOAA climate regions, shown in Fig. 1, dividing the contiguous United States. The regions are the following: Northwest (NW), West (W), west north central (WNC), Southwest (SW), east north central (ENC), South (S), central (C), Northeast (NE), and Southeast (SE). As discussed below, we further segregate our data into cool (October–March) and warm (April–September) seasons. We segregate our data in this way in an attempt to pool data with similar climatological frequencies. While it is well known that conclusions reached in aggregate may fail to hold for specific pools (Simpson 1951; Hamill and Juras 2006), as we show below, this is not a significant issue in our case.

Forecasts were collected from identical zip code/International Civil Aviation Organization (ICAO) stations for all providers and match observations obtained from the National Climatic Data Center quality-controlled local climatic data product.

We used the same observation time frame for each provider: 1 November 2008–31 October 2010. However, as described in section 3b, there were times when one or two providers posted an invalid forecast (e.g., a PoP greater than 1). While these specific forecasts are not included, we do not exclude the forecasts of all providers when one or two providers are not included. As we showed earlier (Bickel et al. 2010), this issue only affects about 3% of the data and does not alter our conclusions.

a. PoP forecasts

PoP forecasts were collected from the public Web sites of each provider at 1800 EST each day. [CustomWeather's 15-day forecasts were collected online at http://www.myforecast.com. TWC's 10-day forecasts were collected online at http://www.weather.com/. (TWC does not forecast out 15 days). Finally, NWS's PoP



Central
Kentucky (KY)
Illinois (IL)
Indiana (IN)
Missouri (MO)
Ohio (OH)
Tennessee (TN)
West Virginia (WV)

East North Central
Iowa (IA)
Michigan (MI)
Minnesota (MN)
Wisconsin (WI)

Northeast
Connecticut (CT)
Delaware (DE)
Maine (ME)
Maryland (MD)
Massachusettes (MA)
New Hampshire (NH)
New Jersey (NJ)
New York (NY)
Pennsylvania (PA)
Rhode Island (RI)
Vermont (VT)

Northwest
Idaho (ID)
Oregon (OR)
Washington (WA)

South
Arkansas (AK)
Louisianna (LA)
Kansas (KS)
Mississippi (MS)
Oklahoma (OK)
Texas (TX)

Southeast
Alabama (AL)
Florida (FL)
Georgia (GA)
North Carolina (NC)
South Carolina (SC)
Virginia (VA)

Southwest
Arizona (AZ)
Colorado (CO)
New Mexico (NM)
Utah (UT)

West
California (CA)
Nevada (NV)

West North Central
Montana (MT)
Nebraska (NE)
North Dakota (ND)
South Dakota (SD)
Wyoming (WY)

All Regions
Central
East North Central
Northeast
Northwest
South
Southeast
Southwest
West
West North Central

FIG. 1. NOAA climate regions used in the analysis. (Source: http://www.ncdc.noaa.gov/temp-and-precip/us-climate-regions.php.)

forecasts were collected from the forecast-at-a-glance section at http://www.weather.gov/.[1]] The forecast-at-a-glance provides forecasts 4 days beyond the current-day forecast. Because of the time of collection (late afternoon) the first day collected was the ''next day'' forecast.

From correspondence with each provider, we determined the valid timing of each PoP forecast. For CW, the PoP forecasts are 24-h forecasts and are valid for the entire 24-h local day. For TWC, the PoP forecasts are valid during 0700–1900 local time. For the NWS, the day-part PoP forecasts on the forecast-at-a-glance section, which we use here, are valid during 1200–2400 UTC. This corresponds, for example, to 0700–1900 EST, 0800–2000 EDT, and 0400–1600 PST. Therefore, lead times and forecast–observation windows are not perfectly matched among providers or regions and differ in two ways. First, the NWS and TWC 12-h observation windows differ for regions that contain observations that are not in the eastern time zone. As one moves west one time zone, the NWS observation (and forecast) window moves 1 h earlier, while the TWC observation window remains fixed at 0700–1900 LT. This difference is greatest in the west. The primary implication of this difference is that the difficulty of TWC and NWS forecasting tasks may not be identical. The difficulty of forecasting

[1] The forecast-at-a-glance appears on local forecast pages returned from entering a city, state, or zip code in the search box on the top-left side of the front page.

TABLE 1. Forecast lead time by provider for the NE, S, SW, and W climate regions.

NWS	1 day	2 day	3 day
NE	13–25 h	37–49 h	61–73 h
S	13–25 h	37–49 h	61–73 h
SW	13–25 h	37–49 h	61–73 h
W	13–25 h	37–49 h	61–73 h

				Diff relative to NWS		
TWC	1 day	2 day	3 day	1 day	2 day	3 day
NE	13–25 h	37–49 h	61–73 h	0%	0%	0%
S	14–26 h	38–50 h	62–74 h	8%	3%	2%
SW	15–27 h	39–51 h	63–75 h	15%	5%	3%
W	16–28 h	40–52 h	64–76 h	23%	8%	5%

				Diff relative to NWS		
CW	1 day	2 day	3 day	1 day	2 day	3 day
NE	6–30 h	30–54 h	54–78 h	−54%	−19%	−11%
S	7–31 h	31–55 h	55–79 h	−46%	−16%	−10%
SW	8–32 h	32–56 h	56–80 h	−38%	−14%	−8%
W	9–33 h	33–57 h	57–81 h	−31%	−11%	−7%

precipitation during 0400–1600 LT may not be the same as forecasting a 0700–1900 window (e.g., precipitation may be less likely in the predawn hours or more likely in the late afternoon). In reality, as we show below, this effect appears to be minor. The second, and more significant, difference among the providers is that forecast lead time varies. For example, when we gathered forecasts for the West at 1500 PST (1800 EST), TWC 1-day lead time was 16 h while the NWS 1-day lead time was 13 h, a difference 3 h or 23% (3/13). Table 1 presents the 1–3-day lead times in hours,[2] for the 4 climate regions that are completely within a single time zone: Northeast (ET), South (CT), Southwest (MT), and West (PT). As one moves west, TWC is at a disadvantage relative to NWS, but this effect becomes muted for longer lead times. For example, TWC's 3-day lead time is 3 h longer than the NWS in the western region, but this is only 5% longer (64 vs 61 h). All else being equal, TWC should exhibit the same degree of skill as NWS in the Northeast, but lower skill as one moves west; this difference should lessen with increasing lead time. As we will show, TWC's performance does not conform to this expectation.

As mentioned above, we will discuss lead times in terms of days, with the caution that the definition for TWC and CW changes slightly as one moves west (see Table 1). Regions that are not completely within a single time zone will have a lead time that is some combination of the values shown in Table 1. We could have instead segregated our data by time zone or altered our collection timing such that each lead time was 12 h. The latter would have increased the difficulty of the collection task significantly. The former is simply the reality of segregating data by natural climate regions, which do not fit perfectly into man-made time zones.

Additionally, through said correspondence and NWS directives, it was identified that NWS does not display PoPs of 0%, nor will it generally display PoPs of 10% except in certain convective weather situations to better describe isolated precipitation. Therefore, a lack of a PoP forecast on NWS's forecast-at-a-glance was interpreted in this paper to be a PoP of 0%. We believe this assumption is reasonable and the conclusion most users would reach.

It was further discovered that the NWS fails to post 4-day PoPs in certain climate regions with high frequency. For example, in the western region, the NWS failed to post a 4-day PoP almost 97% of the time. For this reason, we have excluded all NWS 4-day PoPs from our analysis. Thus, we will consider only NWS 1–3-day PoPs.

b. *Precipitation observations and verification process*

Observation stations from the Automated Surface Observing System (ASOS) and the Automated Weather Observing System (AWOS) networks were selected that could be matched with a zip code centroid lying within 10 km of the observation station. Forecasts from TWC and NWS were queried via this matching zip code, while CW forecasts were queried via the ICAO code of the observation station.

PoP forecasts were verified against the precipitation reported from the observation station. For CW, a precipitation event was considered when measureable precipitation was reported in the 24-h summary observations of the station. For TWC and NWS, the appropriate summation of hourly precipitation observations was used. A precipitation event was considered when measureable precipitation was reported during the 12-h valid window. As the hourly observations were reported in local time, conversion to UTC was performed to ensure the proper 12-h valid window was used for each NWS forecast, taking into account the time zone and daylight savings time observance of the station.

There were a number of audits performed on both collected forecasts and observational data to ensure that both were valid. For observations, if there were not 21 or more hourly observations the observation was invalidated. If the daily high or low temperature reported was not within 5° of the high and low calculated from the hourly observations, or the daily reported precipitation total was not within 0.1 in. of the summed hourly precipitation observations, the observation was invalidated. The cross checking between the daily reported and the

[2] The reader can extrapolate longer lead times by simply adding 24 h for each day.

TABLE 2. Summary of forecast and observation data by lead time for each provider.

Cool season (October–March)												
	National Weather Service				The Weather Channel				Custom Weather			
Lead time (days)	No. of forecasts	Avg POP forecast	Frequency of precipitation	Mean error (ME)	No. of forecasts	Avg POP forecasts	Frequency of precipitation	Mean error (ME)	No. of forecasts	Avg POP Forecast	Frequency of precipitation	Mean Error (ME)
1	222 717	0.201	0.218	−0.017	237 508	0.248	0.216	0.032	236 497	0.276	0.307	−0.030
2	222 086	0.188	0.219	−0.031	237 396	0.231	0.216	0.015	236 496	0.300	0.307	−0.007
3	217 743	0.168	0.218	−0.050	237 390	0.196	0.216	−0.020	236 499	0.299	0.307	−0.008
4	—	—	—	—	237 445	0.193	0.216	−0.024	236 503	0.291	0.308	−0.017
5	—	—	—	—	237 387	0.189	0.216	−0.028	236 502	0.286	0.307	−0.022
6	—	—	—	—	237 519	0.182	0.216	−0.034	236 505	0.287	0.307	−0.020
7	—	—	—	—	237 407	0.166	0.217	−0.050	236 505	0.286	0.307	−0.021
8	—	—	—	—	237 399	0.239	0.216	0.023	236 515	0.292	0.306	−0.014
9	—	—	—	—	237 399	0.232	0.216	0.016	236 514	0.297	0.307	−0.010
10	—	—	—	—	—	—	—	—	236 528	0.301	0.307	−0.007
11	—	—	—	—	—	—	—	—	236 522	0.300	0.308	−0.008
12	—	—	—	—	—	—	—	—	236 536	0.300	0.307	−0.007
13	—	—	—	—	—	—	—	—	236 538	0.301	0.307	−0.007
14	—	—	—	—	—	—	—	—	236 531	0.305	0.307	−0.002
Tot	662 546	0.186	0.218	−0.032	2 136 850	0.208	0.216	−0.008	3 311 191	0.294	0.307	−0.013
Warm season (April–September)												
	National Weather Service				The Weather Channel				Custom Weather			
Lead time (days)	No. of forecasts	Avg POP forecast	Frequency of precipitation	Mean error (ME)	No. of forecasts	Avg POP forecasts	Frequency of precipitation	Mean error (ME)	No. of forecasts	Avg POP forecast	Frequency of precipitation.	Mean Error (ME)
1	248 348	0.189	0.220	−0.030	257 965	0.225	0.218	0.006	259 250	0.260	0.324	−0.064
2	248 266	0.175	0.219	−0.045	257 971	0.216	0.218	−0.002	259 248	0.308	0.323	−0.015
3	247 564	0.158	0.219	−0.060	257 975	0.192	0.218	−0.026	259 241	0.307	0.324	−0.017
4	—	—	—	—	257 973	0.190	0.219	−0.029	259 238	0.304	0.324	−0.020
5	—	—	—	—	257 983	0.190	0.219	−0.029	259 215	0.302	0.324	−0.022
6	—	—	—	—	257 963	0.189	0.220	−0.031	259 205	0.302	0.326	−0.024
7	—	—	—	—	257 951	0.177	0.220	−0.043	259 201	0.285	0.326	−0.042
8	—	—	—	—	257 944	0.249	0.221	0.028	259 201	0.288	0.326	−0.039
9	—	—	—	—	257 946	0.241	0.219	0.022	259 210	0.284	0.325	−0.041
10	—	—	—	—	—	—	—	—	259 198	0.278	0.325	−0.047
11	—	—	—	—	—	—	—	—	259 204	0.276	0.325	−0.049
12	—	—	—	—	—	—	—	—	259 185	0.276	0.325	−0.048
13	—	—	—	—	—	—	—	—	259 174	0.275	0.324	−0.048
14	—	—	—	—	—	—	—	—	259 167	0.276	0.324	−0.048
Total	744 178	0.174	0.219	−0.045	2 321 671	0.208	0.219	−0.012	3 628 937	0.287	0.325	−0.037

hourly observations ensured there were a complete set of hourly observations to construct 12-h precipitation totals. Forecasts were also invalidated if the PoP was not between 0% and 100%. They were invalidated if there was an error with collection, or were of suspicious quality. A total of 10 TWC forecasts, 6845 CW forecasts, and 19 010 NWS forecasts were invalidated due to an audit.

Additionally, ASOS/AWOS stations are down for maintenance at least 1 day every few months, in which case data was not collected. Also, because of network issues and provider Web site issues, there were times when a forecast could not be collected.

The theoretical maximum number of forecast–observation pairs per provider per lead time is the number of stations (734) times the number of days of the study (730), or 535 820. Including both missing observational and forecast data and forecasts and observations invalidated in an audit, 5.21% of possible TWC forecasts, 5.29% of possible CW forecasts, and 6.44% of possible NWS forecasts are not present with the majority of missing data due to missing observations due to site maintenance, or the observation being invalidated as a result of hourly quality issues (not having enough or not matching closely enough with the daily observation).

c. *Data summary*

Before beginning our analysis, we summarize our forecast and observation data in Table 2. We obtained

TABLE 3. Variance of forecasts and observations by lead time for each provider.

Cool season (October–March)

	National Weather Service			The Weather Channel			Custom Weather		
Lead time (days)	No. of forecasts	Variance of forecasts	Variance of obs	No. of forecasts	Variance of forecasts	Variance of obs	No. of forecasts	Variance of forecasts	Variance of obs
1	222 717	0.093	0.170	237 508	0.071	0.169	236 497	0.093	0.213
2	222 086	0.074	0.171	237 396	0.057	0.170	236 496	0.084	0.213
3	217 743	0.055	0.171	237 390	0.029	0.169	236 499	0.069	0.213
4	—	—	—	237 445	0.025	0.170	236 503	0.049	0.213
5	—	—	—	237 387	0.022	0.170	236 502	0.037	0.213
6	—	—	—	237 519	0.019	0.169	236 505	0.029	0.213
7	—	—	—	237 407	0.021	0.170	236 505	0.067	0.213
8	—	—	—	237 399	0.047	0.169	236 515	0.067	0.213
9	—	—	—	237 399	0.046	0.169	236 514	0.069	0.213
10	—	—	—	—	—	—	236 528	0.071	0.213
11	—	—	—	—	—	—	236 522	0.071	0.213
12	—	—	—	—	—	—	236 536	0.071	0.213
13	—	—	—	—	—	—	236 538	0.073	0.213
14	—	—	—	—	—	—	236 531	0.075	0.213

Warm season (April–September)

	National Weather Service			The Weather Channel			Custom Weather		
Lead time (days)	No. of forecasts	Variance of forecasts	Variance of obs	No. of forecasts	Variance of forecasts	Variance of obs	No. of forecasts	Variance of forecasts	Variance of obs
1	248 348	0.062	0.171	257 965	0.045	0.171	259 250	0.062	0.219
2	248 266	0.049	0.171	257 971	0.040	0.170	259 248	0.062	0.219
3	247 564	0.037	0.171	257 975	0.025	0.171	259 241	0.049	0.219
4	—	—	—	257 973	0.023	0.171	259 238	0.039	0.219
5	—	—	—	257 983	0.022	0.171	259 215	0.030	0.219
6	—	—	—	257 963	0.021	0.172	259 205	0.024	0.220
7	—	—	—	257 951	0.023	0.172	259 201	0.071	0.220
8	—	—	—	257 944	0.056	0.172	259 201	0.070	0.220
9	—	—	—	257 946	0.057	0.171	259 210	0.070	0.219
10	—	—	—	—	—	—	259 198	0.069	0.219
11	—	—	—	—	—	—	259 204	0.067	0.219
12	—	—	—	—	—	—	259 185	0.067	0.219
13	—	—	—	—	—	—	259 174	0.068	0.219
14	—	—	—	—	—	—	259 167	0.067	0.219

about 250 000 PoPs for each season lead time–provider combination. In total, we obtained 1 406 724 NWS, 4 458 521 TWC, and 6 940 128 CW PoP forecasts–observation pairs —yielding a total of 12 805 373. The difference in the number of observations by lead time is a result of the data validation process described above.

In the case of the NWS and TWC, precipitation was observed about 22% of the time. Precipitation is more frequent in our CW observations, occurring about 31% of the time, since CW is providing a 24-h PoP.

4. Forecast verification

As shown in Table 2, when averaged over all regions, the NWS tends to underforecast the PoP, as is evidenced by their negative MEs in both the cool and warm seasons. For example, their 3-day warm-season PoP averages 0.158, while precipitation was observed at the rate of 0.219, yielding a ME of −0.060. TWC under forecasts 3–7-day PoPs in the cool season and 2–7-day PoPs during the warm season. CustomWeather underforecasts the PoP for all lead times in both seasons; their warm-season 1-day forecast is the most biased PoP in our dataset.

Table 3 presents the variance in the forecasts and the variance in the observations by season and lead time. As expected, since they are both forecasting a 12-h PoP, the variance in the NWS and TWC observations are nearly identical; CW's 24-h observations exhibit greater variance, confirming that their forecasting task is more difficult. We also notice that the variance in the observations is slightly higher in the warm season, but is essentially independent of lead time. This last point follows since $\sigma^2_{X|l} = p(x = 1|l)p(x = 0|l)$ and the unconditional probability of precipitation within a 12- or 24-h window does not depend upon lead time. These results are consistent with Murphy and Winkler (1992), whose 12-h observation variance ranged from 0.163 to 0.173.

In the interest of space, we do not present observation and forecast variance by region, but summarize a few important results here. During the cool season, the 1–3-day variance of observations in the West for the NWS and TWC averaged 0.148 and 0.143, respectively—a difference of about 3%. Thus, TWC's forecasting task was easier, but the difference is quite small. The results for the Southwest, South, and Northeast are nearly identical during the cool season. During the warm season, the 1–3-day average variance of observations in the West was 0.055 for the NWS and 0.049 for TWC, a difference of about 11%. Again, the TWC's task was slightly easier. TWC's task in the Southwest was slightly more difficult than the NWS during the warm season, with an average observation variance of 0.133 compared to 0.128 for NWS (about 4%). The results in the South and Northeast are nearly identical.

The variance of the NWS's forecasts (Table 3) decreases with lead time, as one would expect. TWC's forecasts exhibit substantially less variance than the NWS, meaning that TWC's near-term forecasts are sharper than NWS. In addition, TWC forecast variances initially decrease, but then *increase* beginning with the 7-day forecast. In fact, during the warm season, the variance in TWC's forecasts are greater for the 8–9-day PoPs than all other lead times, including the 1-day forecast. Thus, TWC's long-term forecasts are sharper than their near-term forecasts, which is clearly problematic. This behavior is related to the TWC's forecasting procedures. As BK08 explained, TWC's 7–9-day forecasts represent the ''objective'' guidance provided by their computer forecasting systems. Human forecasters do not intervene in these cases, as they do for the 1–6-day forecasts. Similar behavior is observed in CW's forecasts: forecast variance decreases, nearly linearly, with lead time until day 7, at which point it increases dramatically.

Skill scores and MSE for the cool and warm seasons, averaged over each geographic region, are presented in Table 4 (we present regional results below). The NWS's cool-season skill scores are 0.469, 0.409, and 0.329, for the 1-, 2-, and 3-day forecasts, respectively. In the warm season, the corresponding NWS SS are 0.358, 0.295, and 0.223. For comparison, based on a much smaller dataset, Murphy and Winkler (1992) found NWS 1–3-day SS of 0.567, 0.376, and 0.295 in the cool season and 0.365, 0.240, and 0.210 for the warm season.

The TWC's cool-season skill scores are 0.420, 0.369, and 0.269, for the 1-, 2-, and 3-day forecasts, respectively. In the warm season, the corresponding SS are 0.330, 0.278, and 0.199. For comparison, BK08 found TWC 1–3-day SS of 0.443, 0.392, and 0.291 in the cool season and 0.225, 0.194, and 0.154 for the warm season. Thus, BK08's SS were slightly higher in the cool season and lower in the warm season. The exact reason for this unknown, but BK08's dataset was much smaller, containing only about 10 000 forecasts per day during the cool season and 6800 during the warm season—compared to the nearly 250 000 we use here. In addition, BK08's observation window was 2 November 2004–16 January 2006.

Considering only the NWS and TWC, we see that NWS's forecasts exhibit more skill. For example, TWC's skill scores are about 5 percentage points lower than the NWS during the cool season and about 2 percentage points lower during the warm season. As detailed in section 3a, the definitional differences in lead time tend to favor the NWS, which certainly affects our results to some extent. However, this definitional difference is smallest for 3-day PoPs and we see that the cool-season SS difference between TWC and NWS is the *largest* at this point—0.269 versus 0.329 a difference of 0.060 or 18%. TWC's 8- and 9-day forecasts exhibit negative skill and are, thereby, worse than forecasting the sample climatological frequency of precipitation.

CustomWeather's 1- and 2-day SS are lower than the NWS, while their 3-day PoP is more skillful, especially during the warm season. While SS normalizes for forecast difficulty, the reader must keep in mind that CW is providing a 24-h PoP forecast; a more skillful 24-h PoP does not imply that a 12-h CW PoP would also be more skillful. CustomWeather's +7-day forecasts are quite poor.

Figure 2 displays the normalized reliability and resolution [Eq. (7)] for each provider. Performance is clearly separated into two domains in the case of TWC and CW. For example, TWC's performance decreases markedly beyond 7 days. In the case of CW, forecasts beyond 6 days are remarkably poor. Within the range of reasonable forecast performance, we see that the resolution term is much larger than the reliability penalty for miscalibration. Furthermore, as one would expect, resolution decreases with lead time. On the other hand, reliability is nearly independent of lead time. This too should be expected since there is no reason that calibration per se should decrease with lead time.

The results discussed above were averaged across all climate regions. Table 5 presents SS by region and season for each provider. The cells with bold (italic) font in are the highest (lowest) skill scores for a particular lead time and provider. The SS by region is widely dispersed. For example, the NWS's 1-day SS during the cool season ranges from a low of 0.282 in the west north central to a high of 0.516 in the Southeast. Lower SS tend to be associated with lower observational variance, since it is harder to improve on climatology (the sample average)

TABLE 4. Cool- and warm-season skill scores by lead time for each provider.

Cool season (October–March)									
	National Weather Service			The Weather Channel			Custom Weather		
Lead time (days)	No. of forecasts	Mean square error (MSE)	Skill score (SS)	No. of forecasts	Mean square error (MSE)	Skill score (SS)	No. of forecasts	Mean square error (MSE)	Skill score (SS)
1	222 717	0.090	0.469	237 508	0.098	0.420	236 497	0.119	0.438
2	222 086	0.101	0.409	237 396	0.107	0.369	236 496	0.127	0.405
3	217 743	0.114	0.329	237 390	0.124	0.269	236 499	0.139	0.345
4	—	—	—	237 445	0.133	0.214	236 503	0.153	0.282
5	—	—	—	237 387	0.143	0.154	236 502	0.167	0.213
6	—	—	—	237 519	0.150	0.117	236 505	0.180	0.154
7	—	—	—	237 407	0.160	0.056	236 505	0.234	−0.099
8	—	—	—	237 399	0.176	−0.041	236 515	0.239	−0.124
9	—	—	—	237 399	0.183	−0.080	236 514	0.249	−0.171
10	—	—	—	—	—	—	236 528	0.256	−0.203
11	—	—	—	—	—	—	236 522	0.257	−0.205
12	—	—	—	—	—	—	236 536	0.259	−0.218
13	—	—	—	—	—	—	236 538	0.268	−0.258
14	—	—	—	—	—	—	236 531	0.270	−0.269

Warm season (April–September)									
	National Weather Service			The Weather Channel			Custom Weather		
Lead time (days)	No. of forecasts	Mean square error (MSE)	Skill score (SS)	No. of forecasts	Mean square error (MSE)	Skill score (SS)	No. of forecasts	Mean square error (MSE)	Skill score (SS)
1	248 348	0.110	0.358	257 965	0.114	0.330	259 250	0.151	0.312
2	248 266	0.121	0.295	257 971	0.123	0.278	259 248	0.154	0.295
3	247 564	0.133	0.223	257 975	0.137	0.199	259 241	0.163	0.255
4	—	—	—	257 973	0.143	0.164	259 238	0.173	0.211
5	—	—	—	257 983	0.148	0.134	259 215	0.184	0.159
6	—	—	—	257 963	0.154	0.102	259 205	0.192	0.128
7	—	—	—	257 951	0.165	0.038	259 201	0.239	−0.086
8	—	—	—	257 944	0.188	−0.093	259 201	0.242	−0.102
9	—	—	—	257 946	0.197	−0.147	259 210	0.248	−0.132
10	—	—	—	—	—	—	259 198	0.253	−0.153
11	—	—	—	—	—	—	259 204	0.253	−0.155
12	—	—	—	—	—	—	259 185	0.257	−0.171
13	—	—	—	—	—	—	259 174	0.262	−0.196
14	—	—	—	—	—	—	259 167	0.262	−0.197

if weather does not vary, but this relationship is not perfect. For example, the lowest observation variance in the warm season occurred in the West, yet, this region had two of the largest skill scores. As was the case when averaged over all regions, the NWS outperforms TWC. For example, in the Northeast, where the lead times and observation windows are the same for both providers, the NWS SS are 11%–25% higher than TWC in the cool season and 1%–23% higher than TWC during the warm season. Likewise, in the Southeast, where forecasting tasks are also closely matched, the NWS performance exceeds TWC. In fact, TWC's SS only exceeds the NWS in 3-season lead time–region combinations: 1- and 2-day forecasts in the West, where it is at the greatest lead-time disadvantage, during the cool season (0.508 vs 0.476 and 0.438 vs 0.426) and 3-day forecasts in the east north central during the warm season (0.190 vs 0.185). Thus, the aggregation of data across regions and lead-time differences does not appear to alter our main conclusion regarding the superior performance of the NWS, relative to TWC. Of course, this does not imply that the NWS outperforms TWC for all possible locations.

During the cool season, the west north central generates the lowest SS for all three providers for most lead times. Performance is mixed during the warm season. For example, both the NWS and TWC perform well in the West, while this is the worst performing region for CW. In fact, CW performance in the West is much worse than other regions.

As was the case when we averaged over all regions, we can gain additional insight into forecasting performance by looking at the components of skill score. Figure 3 presents the normalized reliability and resolution for the NWS's 1–3-day forecasts. First, we see that resolution is

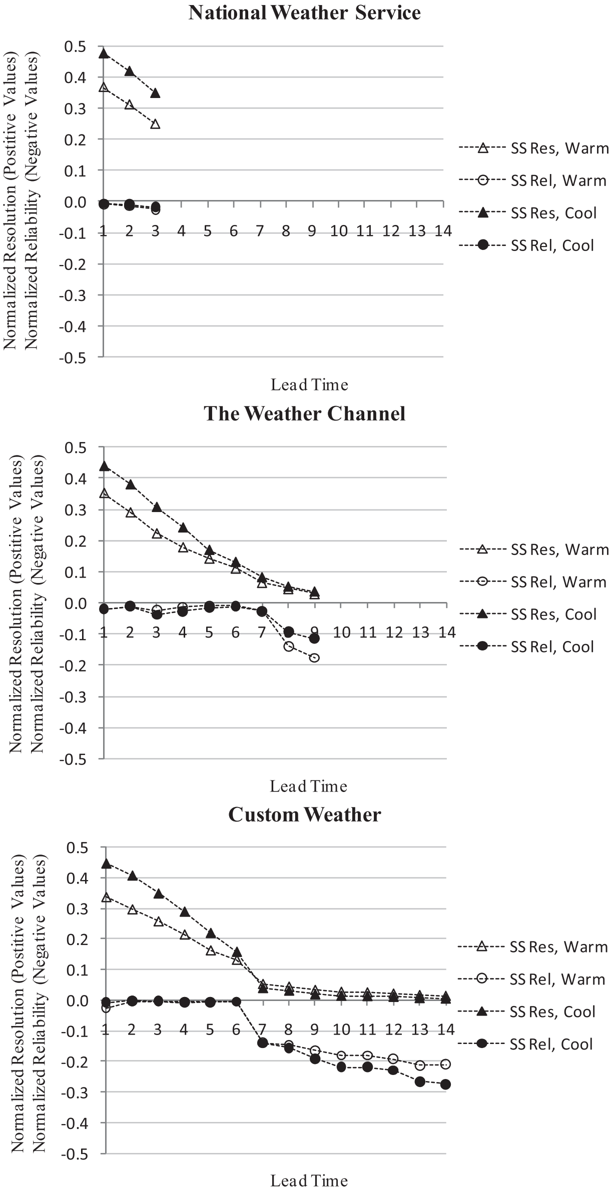


FIG. 2. Skill score components by season for each provider.

the primary determinant of the SS, being an order of magnitude larger than the normalized reliability. Second, even after normalizing for the variance of the observations, large differences between regions remain. For example, for the 1-day forecasts, normalized resolution varies from 0.26 in the Southwest to 0.41 in the Northwest. Finally, as lead time is increased we see a steady decline in normalized resolution and reliability (down and to the left).

a. Calibration-refinement factorization

Figure 4 displays calibration or attributes diagrams (Hsu and Murphy 1986) for the NWS 1–3-day PoP forecasts, for the warm season, averaged over all regions. In the interest of space, we do not present cool-season or regional results, which are available from the corresponding author by request. A line at 45° (omitted here) represents PoPs that are perfectly reliable or calibrated [i.e., $p(x = 1|f, l) = f$]. Based on the normal approximation to the binomial distribution, and assuming that forecasts are independent, we establish a 99% credible interval around this line of perfect calibration and label this region ‘‘calibrated.’’ There is a 1% chance a forecast–observation pair would lay outside this region (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

$$f \pm \Phi^{-1}(0.995)\left[\frac{f(1-f)}{N}\right]^{1/2}, \qquad (8)$$

where Φ^{-1} is the inverse of the standard normal cumulative $[\Phi^{-1}(0.995) = 2.576]$ and N is the number of forecasts.[3] We omit PoPs that were forecasted fewer than 40 times. A cutoff of 40 is a common in hypothesis testing. The variance of the 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. If a forecast–observation pair lies outside the range established by Eq. (8) then we say the forecast is not well calibrated. Calibration is important if users take PoPs at face value, which is likely the case in practice; few users would know how to calibrate the forecasts, which they could now do using these research results. As mentioned, the calibrated region assumes PoP forecasts are independent. This is certainly not true in reality, but the degree of dependence is unknown and we do not attempt to account for it here. Accounting for dependence would widen the calibration interval.

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 calibrated is labeled ‘‘no skill.’’ Along this line the SS is equal to zero and according to Eq. (5), the forecast does not reduce uncertainty in the observation; points above (below) this line exhibit positive (negative) skill. The three lines cross at the sample climatological frequency of precipitation $\mu_{X|l}$.

The dots are the relative frequency with which precipitation was observed for each forecast, $p(x|f, l)$. We see that most of the NWS PoPs are not well calibrated.

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

TABLE 5. Cool- and warm-season skill scores by region, lead time, and provider. The cells with bold (italic) font are the highest (lowest) skill scores for a particular lead time and provider.

Lead time (days)	NWS cool-season SS									NWS warm-season SS								
	Climate region									Climate region								
	C	ENC	NE	NW	S	SE	SW	W	WNC	C	ENC	NE	NW	S	SE	SW	W	WNC
1	0.489	0.377	0.497	0.460	0.464	**0.516**	0.392	0.476	*0.282*	0.368	0.344	0.395	**0.395**	0.309	0.299	*0.256*	0.388	0.320
2	0.410	0.326	0.434	0.402	0.406	**0.447**	0.337	0.426	*0.247*	0.300	0.272	0.320	0.326	0.251	0.244	*0.218*	**0.339**	0.256
3	0.304	0.270	0.354	0.337	0.327	0.327	0.269	**0.318**	*0.192*	0.207	0.185	0.248	0.243	0.184	0.178	*0.164*	**0.289**	0.180
Variance of obs	0.188	0.174	0.195	**0.222**	0.141	0.170	*0.109*	0.146	0.132	0.189	0.179	0.198	0.158	0.156	**0.203**	0.125	*0.054*	0.164

Lead time (days)	TWC cool-season SS									TWC warm-season SS								
	Climate region									Climate region								
	C	ENC	NE	NW	S	SE	SW	W	WNC	C	ENC	NE	NW	S	SE	SW	W	WNC
1	0.432	0.298	0.441	0.405	0.414	0.478	0.309	**0.508**	*0.199*	0.345	0.329	**0.380**	0.363	0.265	0.268	*0.222*	0.357	0.276
2	0.369	0.267	0.380	0.348	0.370	0.421	0.281	**0.438**	*0.178*	0.294	0.267	0.316	**0.318**	0.214	0.221	*0.193*	0.291	0.223
3	0.255	0.215	0.266	0.215	0.274	0.286	0.209	**0.351**	*0.156*	0.203	0.190	0.192	0.221	0.161	0.154	*0.141*	**0.235**	0.160
4	0.192	0.166	0.204	0.159	0.222	0.232	0.172	**0.280**	*0.113*	0.161	0.139	0.153	0.192	0.126	0.126	*0.110*	**0.209**	0.128
5	0.117	0.108	0.126	0.138	0.163	0.165	0.124	**0.223**	*0.075*	0.125	0.094	0.123	0.158	0.103	0.103	0.091	**0.198**	*0.077*
6	0.097	0.072	0.087	0.083	0.125	0.121	0.088	**0.186**	*0.040*	0.079	*0.035*	0.096	0.141	0.071	0.081	0.059	**0.177**	0.041
7	0.027	0.012	0.025	0.011	0.074	0.063	0.040	**0.100**	*−0.018*	0.000	*−0.016*	−0.008	0.073	0.014	0.040	0.015	**0.159**	−0.016
8	−0.070	−0.093	−0.078	−0.005	−0.070	−0.039	−0.115	**0.038**	*−0.127*	−0.135	−0.178	−0.099	**0.041**	*−0.199*	−0.087	−0.162	−0.010	−0.133
9	−0.120	−0.132	−0.122	−0.035	−0.125	−0.045	−0.157	**−0.024**	−0.152	−0.193	−0.245	−0.156	**−0.006**	*−0.254*	−0.125	−0.236	−0.075	−0.207
Variance of obs	0.187	0.175	0.194	**0.224**	0.138	0.172	0.111	0.143	0.132	0.192	0.182	0.197	0.160	0.157	**0.201**	0.133	*0.050*	0.169

Lead time (days)	CW cool-season SS									CW warm-season SS								
	Climate region									Climate region								
	C	ENC	NE	NW	S	SE	SW	W	WNC	C	ENC	NE	NW	S	SE	SW	W	WNC
1	0.464	0.376	**0.488**	0.400	0.420	0.467	0.257	0.412	*0.208*	0.273	0.268	0.337	**0.399**	0.263	0.220	0.176	*0.090*	0.202
2	0.427	0.334	**0.436**	0.390	0.389	0.430	0.266	0.392	*0.196*	0.262	0.239	0.275	**0.372**	0.252	0.227	0.213	*0.090*	0.233
3	0.334	0.260	**0.379**	0.337	0.328	0.347	0.240	0.363	*0.170*	0.213	0.202	0.234	**0.325**	0.199	0.179	0.187	*0.144*	0.193
4	0.262	0.207	0.275	0.290	0.270	0.273	0.206	**0.310**	*0.135*	0.179	0.144	0.180	**0.275**	0.157	0.144	0.152	0.146	*0.143*
5	0.185	0.138	0.190	0.227	0.213	0.209	0.157	**0.264**	*0.078*	0.124	0.108	0.111	**0.184**	0.113	0.104	0.108	0.163	*0.088*
6	0.116	0.081	0.119	0.166	0.160	0.144	0.112	**0.196**	*0.044*	0.087	0.057	0.091	**0.155**	0.072	0.088	0.081	0.125	*0.055*
7	−0.167	−0.138	−0.184	−0.008	−0.110	*−0.216*	0.032	**0.033**	−0.080	−0.167	*−0.256*	−0.154	−0.029	−0.143	−0.082	−0.022	**0.018**	−0.151
8	−0.165	−0.137	−0.179	−0.058	−0.189	*−0.219*	−0.049	**−0.003**	−0.123	−0.179	*−0.308*	−0.208	−0.035	−0.136	−0.069	−0.055	**0.001**	−0.150
9	−0.226	−0.147	−0.196	−0.135	*−0.286*	−0.237	−0.085	**−0.058**	−0.163	−0.225	*−0.264*	−0.238	**−0.040**	−0.181	−0.130	−0.072	−0.068	−0.184
10	−0.232	−0.173	−0.247	−0.157	*−0.306*	−0.271	**−0.095**	−0.119	−0.208	−0.231	*−0.269*	−0.242	**−0.056**	−0.230	−0.181	−0.089	−0.095	−0.184
11	−0.226	−0.161	−0.195	−0.142	−0.306	*−0.373*	−0.159	**−0.110**	−0.174	−0.226	*−0.285*	−0.236	**−0.044**	−0.232	−0.176	−0.086	−0.084	−0.216
12	−0.277	**−0.119**	−0.220	−0.144	−0.308	*−0.339*	−0.140	−0.204	−0.218	−0.243	*−0.291*	−0.223	−0.099	−0.268	−0.193	−0.109	**−0.090**	−0.228
13	−0.332	−0.195	−0.229	−0.203	−0.321	*−0.407*	**−0.184**	−0.285	−0.185	−0.301	*−0.306*	−0.288	−0.093	−0.296	−0.187	−0.119	**−0.080**	−0.237
14	−0.328	**−0.179**	−0.288	−0.220	*−0.399*	−0.368	−0.185	−0.237	−0.192	−0.303	*−0.330*	−0.270	−0.120	−0.248	−0.211	−0.116	**−0.034**	−0.280
Variance of obs	0.227	0.215	0.233	**0.247**	0.188	0.212	*0.144*	0.193	0.173	0.239	0.239	**0.242**	0.199	0.202	0.235	0.166	*0.065*	0.222

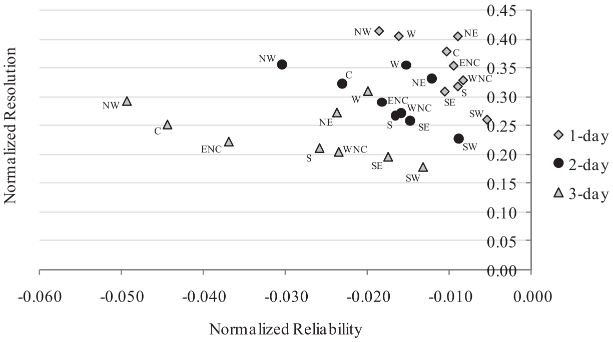


FIG. 3. Comparison of NWS SS components as a function of lead time (warm season).

For example, only 1-day PoPs of 0.1 and 0.9 are well calibrated, while 0.2 and 0.3 are very close. Some miscalibration is undoubtedly caused by the decision to forecast at a resolution of one-tenth (e.g., 0.1, 0.2). For example, shifting 1-day PoPs between 0.4 and 0.8 to the right by 0.05 would improve calibration. In other words, a PoP of 0.4 summarizes all PoPs between 0.4 and 0.5 and might be better thought of as a PoP of 0.45.

TWC's 1–8-day calibration diagrams appear in Fig. 5; 9-day performance is similar to 8-day performance and is omitted. Some PoPs are considerably miscalibrated. For example, when TWC forecasted a PoP of 0.7, precipitation occurred 84% of the time, which is more frequent than when TWC gave a 1-day PoP of 0.8. Whether or not a difference of 0.14 between the forecast and the observation is important depends, of course, on the decision situation. In some cases, such as planning a picnic, users may not distinguish between and 0.70 and a 0.84 chance of rain. On the other hand, a utility might plan differently in these two situations. As was true with the NWS, increasing midrange PoPs by 0.05 would improve calibration.

TWC's performance decreases markedly after 6 days. As was discussed in BK08, 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 1–6-day performance is much better.

CustomWeather's 1–8-day calibration diagrams appear in Fig. 6; 9–14-day results are similar to 8-day results and are omitted. Their 1-day PoPs are considerably biased (as we also saw in Table 1), but still exhibit positive skill. It is quite interesting that CW does seem to be able to forecast at the 0.01 level. That is, in most cases, it was more likely to precipitate for a PoP of f than for a PoP of $f - 0.01$ ($f > 0$). CustomWeather's 2-day forecast is much better than their 1-day forecast, with the 1-day bias having been removed. In addition, many of the 2-day PoPs

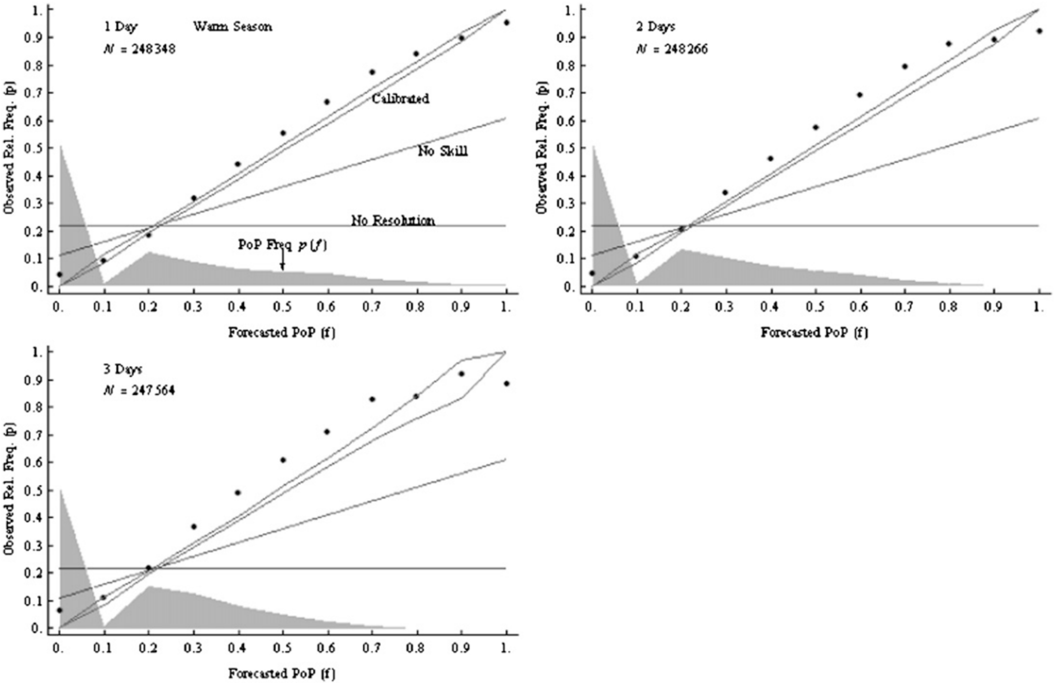


FIG. 4. NWS calibration diagrams for 1–3-day PoP forecasts.

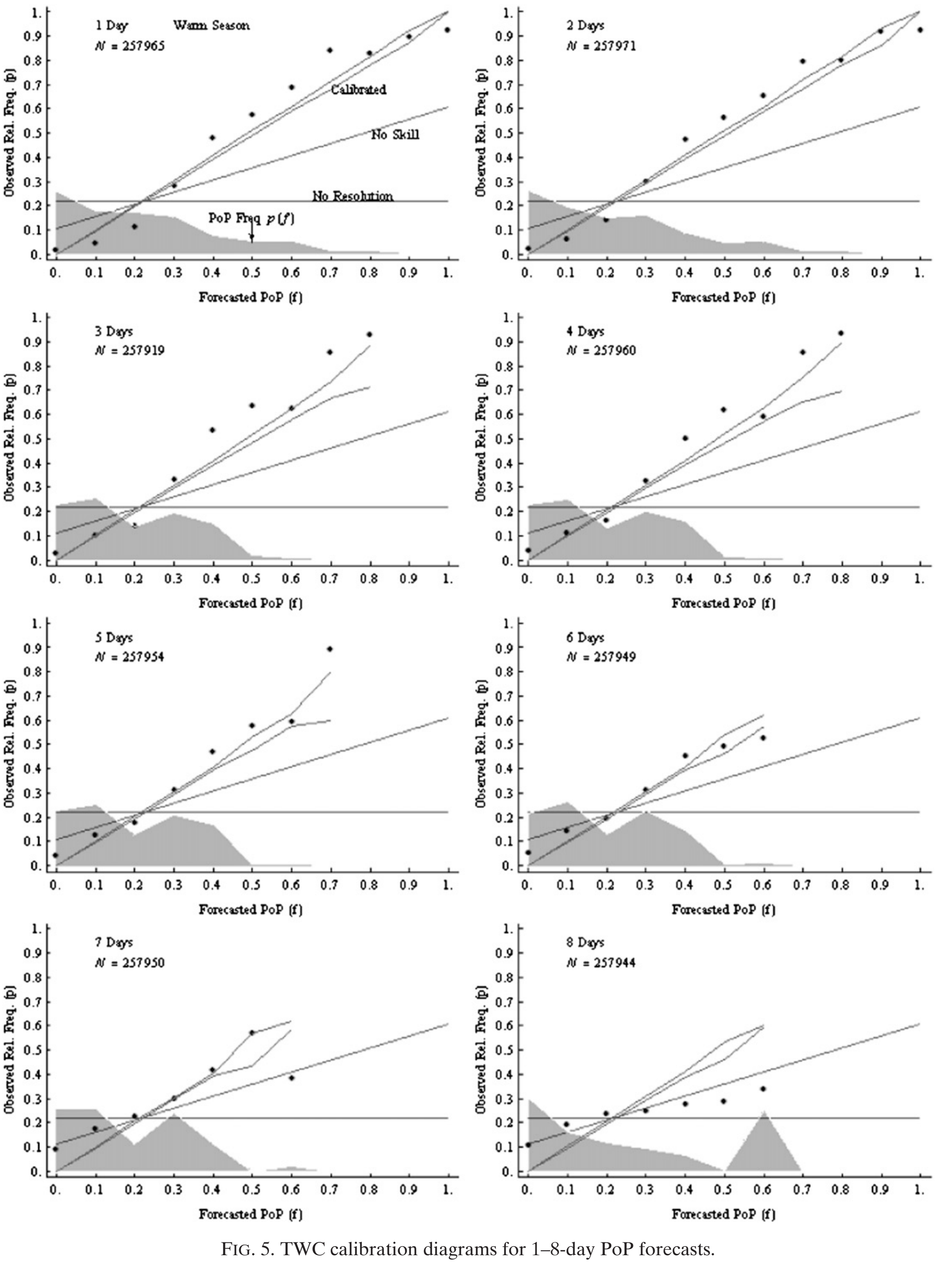


FIG. 5. TWC calibration diagrams for 1–8-day PoP forecasts.

are well calibrated. CustomWeather's performance changes dramatically after 6 days. Their +7-day forecasts are quite poor; these forecasts exhibit almost no resolution.

The gray areas in Figs. 4–6 present the frequency $p(f)$ with which different PoPs are forecast. The three providers differ substantially in this regard. Beginning with the NWS (Fig. 4), we see very few 0.1 PoP forecasts. As discussed in section 3a, the NWS often, but not always, fails to report PoPs below 0.2 and we treat the failure to report a PoP as a forecast of 0%. Beyond PoPs of 0.2, the NWS PoP frequency decreases monotonically and smoothly, as one might expect. TWC's forecasts (Fig. 5), on the other hand, appear to be concentrated at particular PoPs, while avoiding others. This

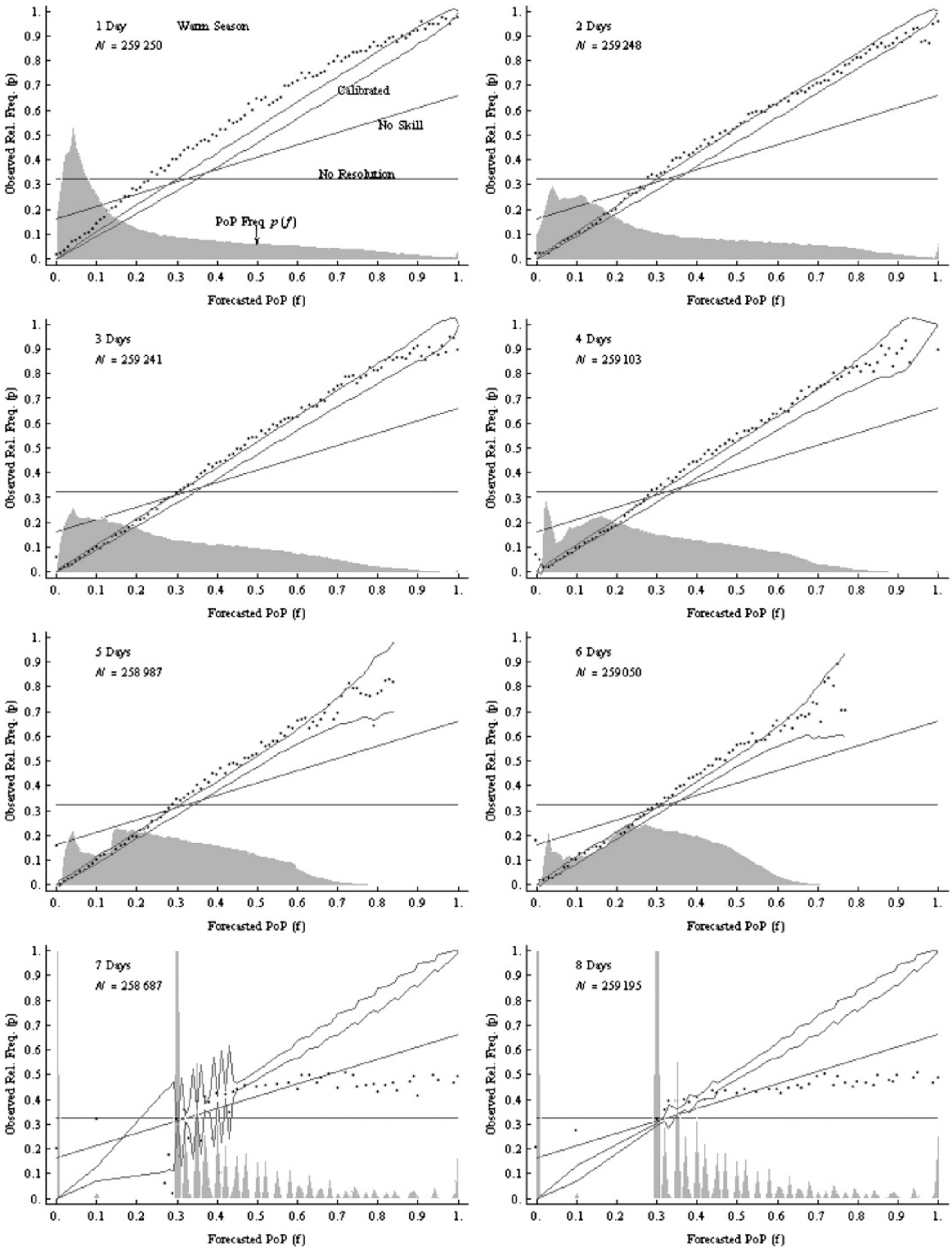


FIG. 6. CW calibration diagrams for 1–8-day PoP forecasts.

is most evident in the longer lead times. For example, we see that TWC provides relatively few 0.2 PoPs for 3–7-day lead times, despite the fact that the average occurrence of precipitation in our sample was approximately 0.2 (see Table 2). One would think that PoPs should become more concentrated near the climatological average as lead time increases. This phenomenon is clearly evident in CW's 6-day forecasts (Fig. 6). Instead of becoming concentrated around climatology, TWC's long-term PoPs center on 0.0 and 0.6, with a very noticeable gap at 0.5. As discussed in BK08, this behavior is *intentional* because TWC believes that users will interpret a PoP of 0.5 as a lack of knowledge (after all, there are only two possible outcomes), when, in fact, a forecast of

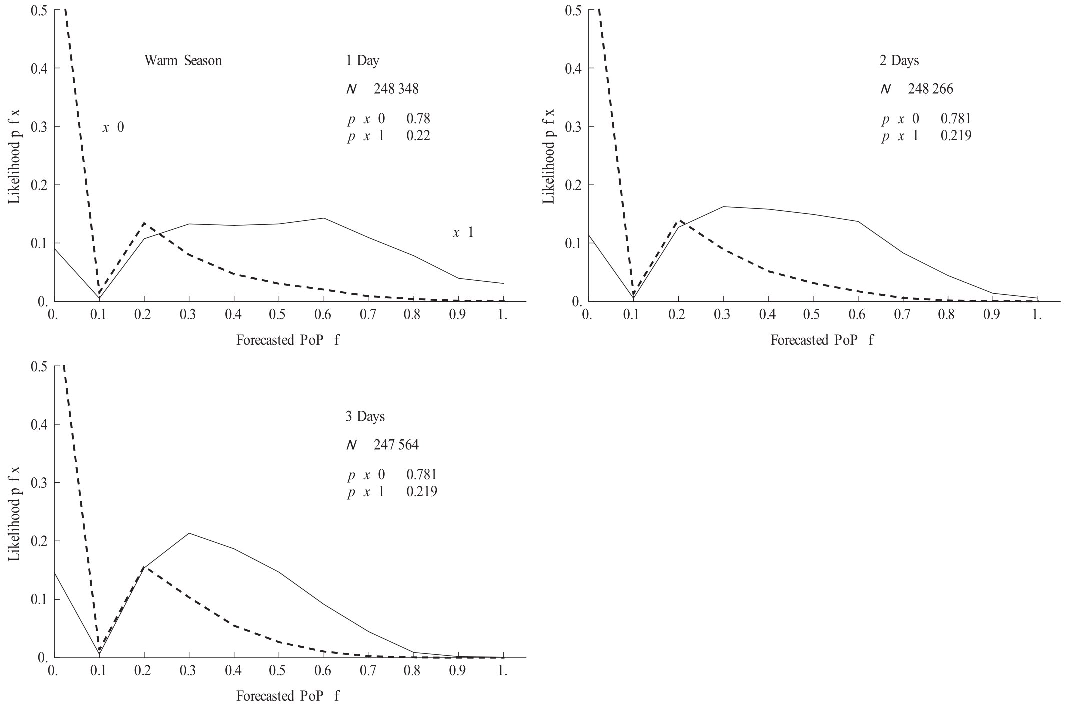


FIG. 7. NWS likelihood diagrams for 1–3-day PoP forecasts.

0.5 is more than twice the climatological average and thus a strong statement regarding the chance of precipitation. This policy degrades the quality of TWC's forecasts.

While CW's pattern of 1–6-day PoP forecasts (Fig. 6) exhibit a few irregularities (e.g., the lower likelihood of PoPs around 0.1 in the 5-day forecast), it is generally concentrated at the long-term average precipitation rate and decreases monotonically with increasing PoP. This is, again, quite interesting since CW forecasts with a resolution of 0.01. The +7-day forecasts are markedly different, however. Rather than the smooth and continuous pattern observed in the 1–6-day forecasts, the +7-day forecasts are concentrated at particular PoPs and avoid others altogether. We notified G. Flint (founder and CEO of CW) of this phenomena and he noted that CW is "having to work with low resolution data beyond day 7 [our 6-day forecast] that doesn't actually provide . . . substantive PoP values so we had to derive them from precipitation totals. This methodology obviously needs improvement so this is certainly something that we need to work on."

b. Likelihood base-rate factorization

Figure 7 displays the likelihood functions, $p(f|x = 1, l)$ and $p(f|x = 0, l)$ for the NWS 1–3-day warm-season PoP forecasts averaged over all regions. Again, regional and cool-season results are available from the corresponding author by request. We see, by the small spike at 0 and the lack of a spike at 1, that the NWS is more skilled at forecasting a lack of precipitation than precipitation. For example, given that it precipitated ($x = 1$; the solid line), the NWS was almost equally likely to give a forecast between 0.2 and 0.6. Furthermore, they were very unlikely to have given a PoP of 0.9 or 1.0 when it precipitated 1-day later. The likelihood functions cross between PoPs of 0.2 and 0.3, which contains the sample climatological frequency of precipitation.

TWC's 1–8-day likelihood graphs appear in Fig. 8; longer lead times are similar to the 8-day results and are omitted. Again, we see TWC's predilection for forecasting particular PoPs and avoiding 0.5. The very large spike at a PoP of 0.6 in their 8-day forecast is especially telling; TWC was more likely to provide a PoP of 0.6 eight days out than any other forecast. Even when it did not precipitate, a PoP of 0.6 was provided over 20% of the time.

CustomWeather's 1–8-day likelihoods are shown in Fig. 9; longer lead times are similar to the 8-day results and are omitted. Their near-term forecasts are sharper than other providers, as evidenced by the small spike at PoPs near 1.0.[4] This could be due to the fact that CW forecasts at a resolution of one-hundredth. In situations

[4] Since CW forecasts at a resolution of 0.01, rather than 0.1, their PoP frequency tends to be about one-tenth that of TWC or the NWS, which explains the scale of the vertical axis in these graphs.

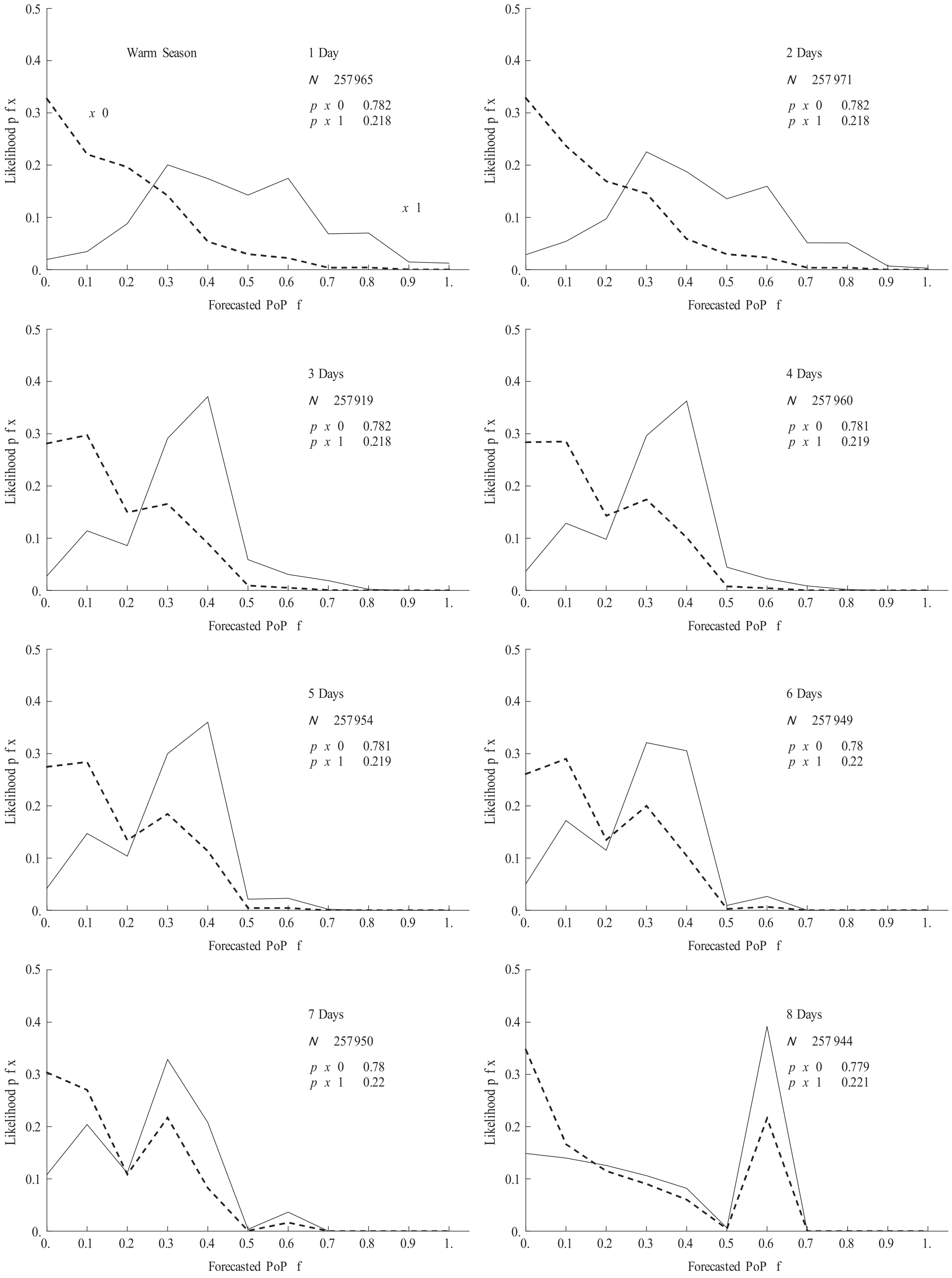


FIG. 8. TWC likelihood diagrams for 1–8-day PoP forecasts.

where precipitation was not observed, all three providers were more likely to provide a low PoP. CustomWeather's performance is especially impressive in this regard. Their +7-day likelihoods, on the other hand, are completely overlapping, highlighting the independence of precipitation observations and their forecasts.

5. Discussion and conclusions

In this study, we have analyzed the absolute and relative performance of the NWS, TWC, and CW's PoP forecasts. In an absolute sense, all three providers exhibit positive skill: the NWS from 1 to 3 days, TWC from

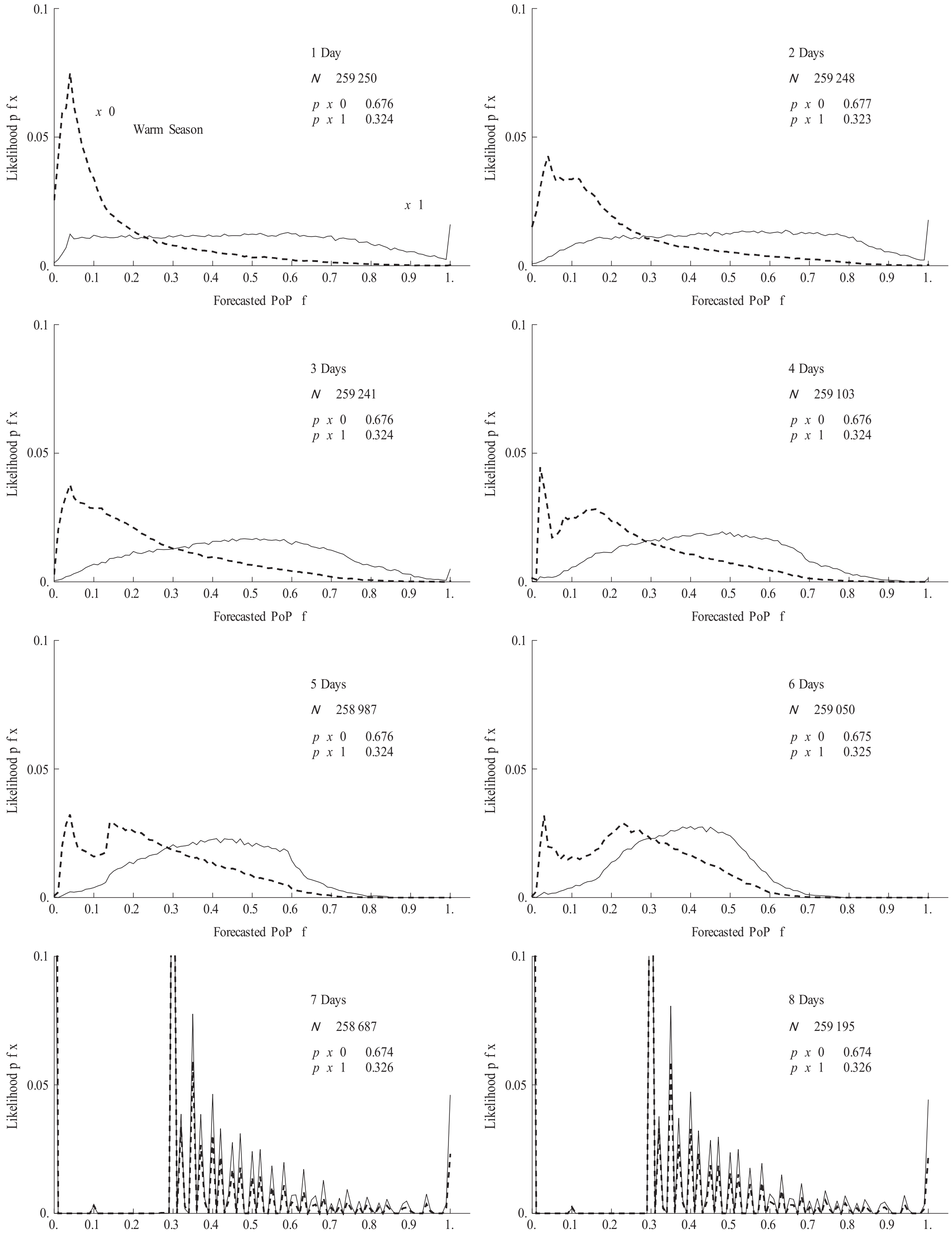


FIG. 9. CW likelihood diagrams for 1–8-day PoP forecasts.

1 to 7 days, and CW from 1 to 6 days. Yet, individual PoP forecasts are miscalibrated, considerably so in some cases. Our analysis could be used to calibrate future forecasts from these providers. Doing so, as shown by the resolution component in Fig. 2, would result in positive skill for all lead times, but would reduce the sharpness of TWC and CW's +7-day forecasts. The sharpness of the existing forecasts implies a level of certainty about future precipitation that is not borne out in the observations.

We have also reconfirmed BK08's findings that TWC's PoP forecasts encode some odd forecasting behaviors; their avoiding PoPs of 0.2 and 0.5, being the most striking examples. These behaviors seem rooted in easily changeable policies, rather than in the difficulty of the forecasting task.

Perhaps the most interesting feature we found was the ability of CW to forecast at a resolution of 0.01. In most cases, but not always, when CW provided a PoP of f, it was more likely to precipitate than when they gave a forecast of $f - 0.01$.

In a relative sense, we only compare the PoP forecasts of TWC and NWS. Even here, our ability to draw definitive conclusions is hampered by differences in lead-time and observation windows. However, in situations where these parameters were identical (e.g., the Northeast region), or nearly so (e.g., the Southeast region), the NWS SS exceed those of the TWC. This performance, combined with TWC's forecasting behaviors outlined above, led us to believe that TWC is not adding skill above and beyond what is in the NWS forecasts. Yet, TWC is adding value by providing forecasts that cover the 0700–1900 LT window, which is probably of more relevance to many users in the western region, for example, than a 0400–1600 LT window, which is provided by the NWS.

To summarize, we hope that this analysis will help the NWS, TWC, and CW provide better forecasts and help users better interpret and use these forecasts in their decision making.

Acknowledgments. The authors thank Geoff Flint of CustomWeather; Dr. Bruce Rose of The Weather Channel; and Bob Glahn, David Ruth, and Ron Jones of the National Weather Service for help in properly interpreting their respective forecasts. In addition, the authors thank three anonymous reviewers and the associate editor for their helpful comments and suggestions. This research was partially supported by NSF Grant CMMI-0907794.

REFERENCES

Bickel, J. E., and S. D. Kim, 2008: Verification of the weather channel probability of precipitation forecasts. *Mon. Wea. Rev.,* **136,** 4867–4881.

——, W. Floehr, and S. D. Kim, 2010: Comparing NWS PoP forecasts to third-party providers. Preprints, *20th Conf. on Probability and Statistics,* Atlanta, GA, Amer. Meteor. Soc., P3. [Available online at http://ams.confex.com/ams/90annual/techprogram/paper_161669.htm.]

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

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real or is it the varying climatology. *Quart. J. Roy. Meteor. Soc.,* **132,** 2905–2923.

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.

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 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.

Simpson, E. H., 1951: The interpretation of interaction in contingency tables. *J. Roy. Stat. Soc.,* **13A** (2), 238–241.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability of ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 137–163.