

COMPARING NWS POP FORECASTS TO THIRD-PARTY PROVIDERS

P.3

J. Eric Bickel (The University of Texas at Austin)*, Eric Floehr (Intellovations, LLC), and
Seong Dae Kim (University of Alaska Anchorage)

1. INTRODUCTION

Bickel and Kim (2008) (hereafter BK) analyzed approximately 169,000 probability of precipitation (POP) forecasts provided by The Weather Channel (TWC) over a 14-month period, spanning 2004-2006 at 42 US locations. BK found that TWC's near-term forecasts (less than a three-day lead time) were relatively well calibrated. Longer-term forecasts were less reliable. This performance was driven by TWC's forecasting policies and tools. For example, TWC artificially avoids POPs of 0.5.

In this paper, we use a much larger database than BK to analyze and compare the reliability POP forecasts provided by National Weather Service (NWS), CustomWeather (CW), and TWC. Specifically, we analyze 7 million POPs covering a 12-month period (01 November 2008 through 31 October 2009) at 753 stations across the US. This larger dataset confirms the results of BK and extends their analysis in important respects. First, we provide verification results for two additional providers of POP forecasts, including the NWS. Second, we analyze whether third party forecasts are more skilled than those of the NWS.

The Weather Channel® is the leading provider of weather information to the general public via its cable television network and interactive website 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 websites, according to Nielsen/NetRatings.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 US 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. Finally, in Section 5 we conclude.

2. VERIFICATION OF PROBABILITY FORECASTS

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

2.1 Distributional Measures

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

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 denoted $p(f,x|l)$ and completely describes the performance of the forecasting system. A perfect forecasting system would ensure that $p(f,x|l) = 0$ when $f \neq x$. Lead times for the TWC may obtain integer values ranging from 1 (one day ahead) to 9 (the last day in a 10-day forecast). BK 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 4 days. CW provides POPs from 1 to 14 days ahead.

Since

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

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 . The forecasts and observations are independent if and only if $p(x|f,l) = p(x|l)$. A set of forecasts is well *calibrated* if $p(1|f) = f$ 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 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. For

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

¹ <http://press.weather.com/company.asp>.

example, $p(f|1, I)$ is the relative frequency of forecasts when precipitation occurred, and $p(f|0, I)$ 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 $p(f|x, I) = p(f|I)$.

2.2 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

$$\bar{f}_I = \sum_F f p(f | I) = E_{F|I}[f],$$

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

$$\bar{x}_I = \sum_x x p(x | I) = E_{x|I}[x].$$

The *mean error (ME)* is

$$ME(f, x | I) = \bar{f}_I - \bar{x}_I \quad (2)$$

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

$$MSE(f, x | I) = E_{F, x|I}[(f - x)^2]. \quad (3)$$

The *climatological skill score (SS)* is

$$SS(f, x | I) = 1 - MSE(f, x | I) / MSE(\bar{x}_I, x | I). \quad (4)$$

Since

$$MSE(\bar{x}_I, x | I) = E_{F, x|I}[(\bar{x}_I - x)^2] = \sigma_x^2,$$

where σ_x^2 is the variance of the observations, the SS can be written as

$$SS(f, x | I) = \frac{\sigma_x^2 - MSE(f, x | I)}{\sigma_x^2} \quad (5)$$

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

It is important to note that ranking providers by MSE and SS may not yield the same results if their observation windows differ. For example, as will be explained below, CW provides a 24-hour POP, while both the NWS and TWC provide 12-hour POPs. Thus, CW's forecasting task is more difficult and they will likely have higher a MSE even if they are just as skilled as the other providers. To correct for this, SS divides by the variance of the observations, which is a measure of forecast difficulty. This adjustment will be greater for 24-hour POP forecasts than for 12-hour forecasts.

3. DATA GATHERING PROCEDURE

We wish to analyze both the absolute performance of each provider and their performance relative to each other. To analyze absolute performance, we would like to include as many observations as possible. However,

to make comparisons as fair as possible, we must use the same observation window for each provider. As mentioned in §1, the observation window we selected is 01 November 2008 through 31 October 2009. However, even within this window there are times when one or two providers posted an invalid forecast (e.g., a POP greater than 1). To correct for this, at the end of the paper, we provide a "head-to-head" comparison where we only include forecasts in situations where all three providers provided a valid forecast.

3.1 POP Forecasts

POP forecasts were collected from the public websites of each provider at 6pm ET each day. CW's 15-day forecasts were collected from www.myforecast.com. TWC's 10-day forecasts were collected from www.weather.com (they do not forecast out 15-days). Finally, NWS's POP forecasts were collected from the forecast-at-a-glance section at www.weather.gov. The forecast-at-a-glance provides forecasts for four full day-parts 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-hour forecasts and are valid for the entire 24-hour local day. For TWC, the POP forecasts are valid from 7am to 7pm local time. For the NWS, the day-part POP forecasts on the forecast-at-a-glance section are valid 7am-7pm UTC. For example, this corresponds to 7am-7pm EST, 8am-8pm EDT, and 4am-4pm PST. Additionally, through said correspondence, it was identified that NWS does not display POPs of 0%, nor will it display POPs of 10% except for the Western Region. 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%.

Forecasts were collected from identical zip code/ICAO stations for all providers. However, because the NWS website does not provide forecast-at-a-glance forecasts for Alaskan cities, the NWS POP forecasts do not include POP forecasts for 16 locations within Alaska.

3.2 Precipitation Observations

Forecasts were collected to match observations obtained from the National Climatic Data Center Quality-Controlled Local Climatic Data product. Observation stations from the ASOS/AWOS observation network were selected that could be matched with a zip code centroid lying within 10km 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-hour summary observations of the station. For TWC and NWS, the appropriate summation of hourly precipitation observations was used. As the hourly observations were reported in local time, conversion to UTC was

performed to ensure the proper 12-hour 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 weren't 21 or more hourly observations the observation was invalidated. If the daily high or low temperature reported was not within 5 degrees of the high and low calculated from the hourly observations, or the daily reported precipitation total was not within 0.1 inch of the summed hourly precipitation observations, the observation was invalidated. The cross-checking between the daily reported and the hourly observations ensured there were a complete set of hourly observations to construct 12-hour 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 16 TWC forecasts, 11 CW forecasts, and 67 NWS forecasts were invalidated due to audit.

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

Including both missing observational and forecast data and forecasts and observations invalidated in audit, 3.42% of possible TWC forecasts are not present, 3.52% of possible CW forecasts are not present, and 13.08% of possible NWS forecasts are not present (including the missing Alaskan cities) with the majority of missing data due to missing observations due to site maintenance, or the observation being invalidated due to hourly quality issues (not having enough or not matching closely enough with the daily observation).

3.3 Data Summary

In what follows, we exclude forecast-observation pairs for which there are fewer than 40 observations. A cutoff of 40 is a 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 $p = \frac{1}{2}$ then N should be greater than 40.

Before beginning our analysis, we summarize our forecast and observation data in Table 1. We obtained about 250,000 POPs for each lead-time-provider combination. In total, we obtained 985,584 POP forecasts for the NWS, 2,388,921 for TWC and 3,721,084 for CW—yielding a total of 7,095,591 POP observations. The difference in the number of observations by lead time is a result of a data validation process; we have removed invalid forecasts and there were rare times when provider website issues or other technical problems prevented collection.

In the case of the NWS and TWC, precipitation was observed about 23% of the time. Precipitation is more frequent in our CW observations because CW is providing a 24-hour POP.

The NWS tends to under forecast the POP, as is evidenced by their negative MEs. For example, its four-day-ahead (4da) POP averages 0.097 even though precipitation was observed at the rate of 0.228, yielding a ME of -0.131; which is significantly biased. TWC's MEs are lower than that of the NWS. However, they under forecast POPs from 3da to 7da. CW under forecasts the POP for all lead times and their 1da forecast is quite biased. However, CW outperforms TWC and the NWS in terms of mean error in five of the first nine lead times.

Overall, the MEs are -0.066, -0.014, and -0.031 for the NWS, TWC, and CW, respectively. TWC's 1da, 2da, 8da, and 9da forecasts are particularly free from bias.

4. FORECAST VERIFICATION

BK separately considered cool (October-March) and warm (April-September) seasons. In the interest of space, we do not make this division in this paper. These results are available by request from the corresponding author.

4.1 Calibration-Refinement Factorization

Fig. 1 displays a calibration or attributes diagram (Hsu; Murphy 1986) for the NWS, TWC and CW 1da POP forecasts. The line at 45° represents POPs that are perfectly calibrated, i.e., $p(1|f, f) = f$. Based on the normal approximation to the binomial distribution, we establish a 99% credible interval around this line of perfect calibration and label is "Calibrated". There is a 1% chance a forecast-observation pair would lay 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

$$f \pm \Phi^{-1}(.995) \left(\frac{f(1-f)}{N} \right)^{1/2}, \quad (6)$$

where Φ^{-1} is the inverse of the standard normal cumulative ($\Phi^{-1}(.995) = 2.576$) and N is the number of forecasts.² If a forecast-observation pair lies outside this range then the forecast is not well calibrated.

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 skill score (SS) is equal to zero and according to Equation (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 climatological frequency of precipitation \bar{x}_1 .

² This is identical to a two-tailed t-test with a 1% level of significance.

Table 1. Summary of forecast and observation data

Lead Time (Days)	National Weather Service					The Weather Channel					Custom Weather					Lowest Absolute ME
	Number of Forecasts	Precip. Obs. (x = 1)	Avg. POP Forecast	Freq. of Precip.	Mean Error (ME)	Number of Forecasts	Precip. Obs. (x = 1)	Avg. POP Forecast	Freq. of Precip.	Mean Error (ME)	Number of Forecasts	Precip. Obs. (x = 1)	Avg. POP Forecast	Freq. of Precip.	Mean Error (ME)	
1	249,486	57,310	0.200	0.230	-0.030	265,431	60,512	0.242	0.228	0.014	265,848	87,041	0.278	0.327	-0.050	TWC
2	250,573	57,548	0.186	0.230	-0.043	265,446	60,514	0.229	0.228	0.001	265,847	86,875	0.314	0.327	-0.013	TWC
3	245,585	56,310	0.169	0.229	-0.061	265,434	60,502	0.199	0.228	-0.029	265,856	87,043	0.312	0.327	-0.016	CW
4	239,942	54,778	0.097	0.228	-0.131	265,420	60,586	0.198	0.228	-0.031	265,850	87,180	0.307	0.328	-0.021	CW
5						265,463	60,672	0.198	0.229	-0.031	265,699	87,033	0.305	0.328	-0.023	CW
6	-	-	-	-	-	265,435	60,889	0.195	0.229	-0.035	265,708	87,164	0.303	0.328	-0.025	CW
7	-	-	-	-	-	265,423	60,917	0.182	0.230	-0.047	265,202	87,149	0.294	0.329	-0.034	CW
8	-	-	-	-	-	265,438	60,932	0.246	0.230	0.016	265,872	87,286	0.296	0.328	-0.032	TWC
9	-	-	-	-	-	265,431	60,827	0.240	0.229	0.011	265,867	87,344	0.299	0.329	-0.029	TWC
10	-	-	-	-	-						265,863	87,457	0.292	0.329	-0.037	CW
11	-	-	-	-	-	-	-	-	-	-	265,868	87,432	0.290	0.329	-0.039	CW
12	-	-	-	-	-	-	-	-	-	-	265,871	87,347	0.292	0.329	-0.037	CW
13	-	-	-	-	-	-	-	-	-	-	265,867	87,277	0.287	0.328	-0.041	CW
14	-	-	-	-	-	-	-	-	-	-	265,866	87,110	0.289	0.328	-0.039	CW
Total	985,586	225,946	0.164	0.229	-0.066	2,388,921	546,351	0.214	0.229	-0.014	3,721,084	1,220,738	0.297	0.328	-0.031	TWC

The dotted lines are $p(x|f, l=1)$ or the relative frequency with which precipitation was observed for each forecast. We see that most of the POPs are not well calibrated. For example, in the case of the NWS, only POPs are 0.1, 0.2, and 0.9 are well calibrated, while 0.3 is very close. POPs between 0.4 and 0.8 are under forecast. POPs of 1.0 are not surprisingly over forecast.

In the case of TWC, POPs of 0.2 and below are significantly miscalibrated and exhibit negative skill, echoing the results of BK (see BK for a discussion). TWC's midrange POPs follow a similar pattern as the NWS.

CW's 1da POPs are significantly 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 $f + 0.01$ than for a POP of f ($f < 1$). This behavior is weakest for POPs between 0.48 and 0.56.

The grey area in Fig. 1 presents the frequency $p(f|l=1)$ with which different POPs are forecast. The three providers differ in this respect. For example, the TWC is mostly likely to provide a POP of 0 or 0.1 one day ahead. In addition, TWC is just as likely to provide POPs in particular regions (e.g., 0.2 to 0.3 and 0.4 to 0.6). CW concentrates its 1da POPs around 0.05. The large fraction POPs at 0 provided by the NWS stems from our assumption that the lack of POP is equivalent to a POP of 0. As discussed above, we believe this assumption is reasonable and the one likely to be made by users.

The calibration diagrams for the NWS's 1da to 4da forecasts are displayed in Fig. 2. The performance of the 1da to 3da forecasts are similar. The forecasts are miscalibrated and biased. Performance noticeably declines for the 4da forecast. Again, the lack of forecasts at 0.1 is a result of our assumption that the absence of a POP forecast is equivalent to a POP of 0.

TWC's 1da to 9da calibration diagrams appear in Fig. 3. Performance for the 2da forecasts is similar to the 1da. Midrange POPs are significantly miscalibrated beginning with the 3da forecast. TWC's performance decreases markedly after six days. As was discussed in BK, 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 six days. Thus, the verification results of the 7da – 9da 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 1da – 6da performance is much better. However, when humans do intervene, they introduce considerable bias into the low-end POP forecasts.

We also notice that TWC's POP forecasts are more "lumpy" than the other providers and exhibit odd preferences for particular POPs. This is most evident in the longer lead times and we see that TWC avoids POPs of 0.5. As discussed in BK, this behavior is intentional because TWC believes that viewers will interpret a POP of 0.5 as a lack of knowledge. This policy significantly degrades the quality of TWC's forecasts.

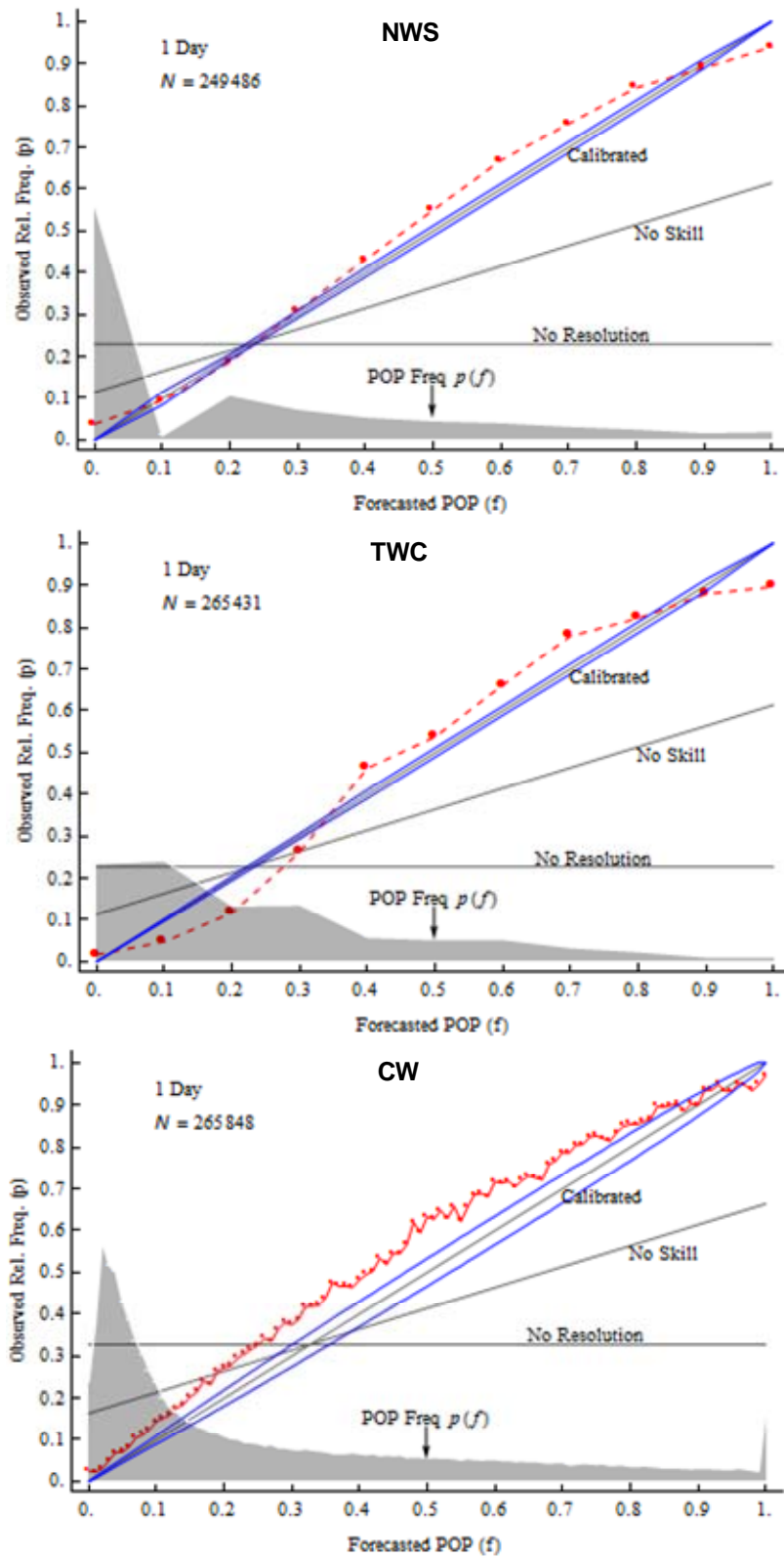


Fig. 1. Calibration diagram for NWS, TWC and CW's 1da POP forecasts

CW's calibration diagrams for their 1da to 9da forecasts appear in Fig. 4. The performance of their 10da to 14da forecasts are similar to the 9da and are omitted. CW's 2da forecast is much better than their 1da forecast, with the 1da significant bias having been removed. In addition, many of the 2da POPs are well calibrated. CW's performance changes dramatically after six days. Their 7+da forecasts are quite poor. First, we notice these forecasts exhibit almost no resolution—their forecast has almost nothing to do with the frequency of precipitation. Second, we notice that the

pattern of POP forecasts is markedly different. Rather than the smooth and continuous pattern observed in the 1da to 6da forecasts, the 7+da forecasts are concentrated at particular POPs and spike at 0.3. We notified CW's Geoff Flint (Founder and CEO) of this phenomena and he noted that CW is "having to work with low resolution data beyond day 7 [our 6da 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."

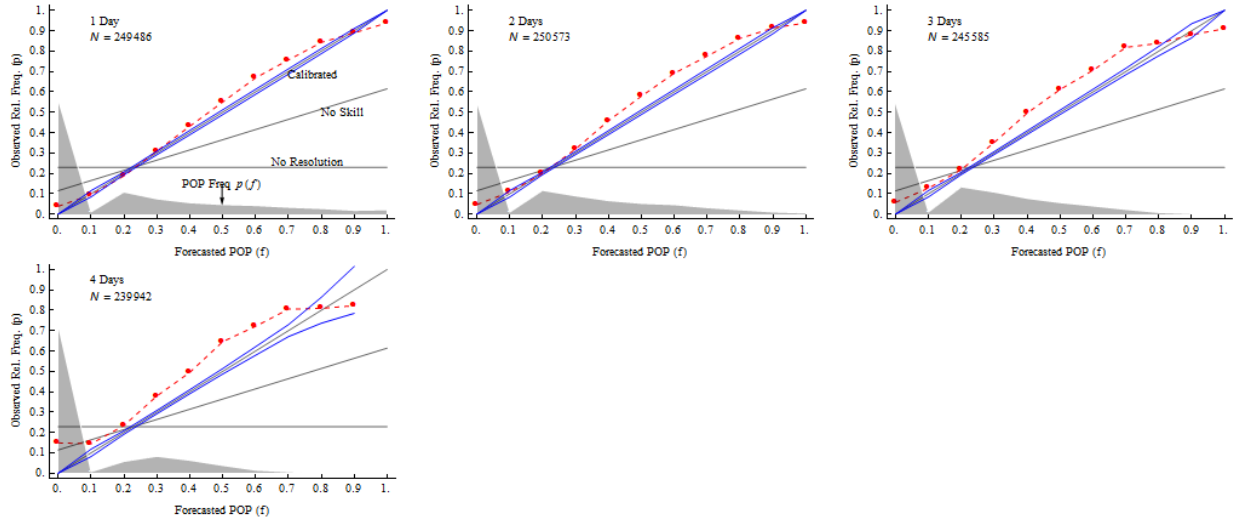


Fig. 2. NWS calibration diagrams for 1da to 4da POP forecasts

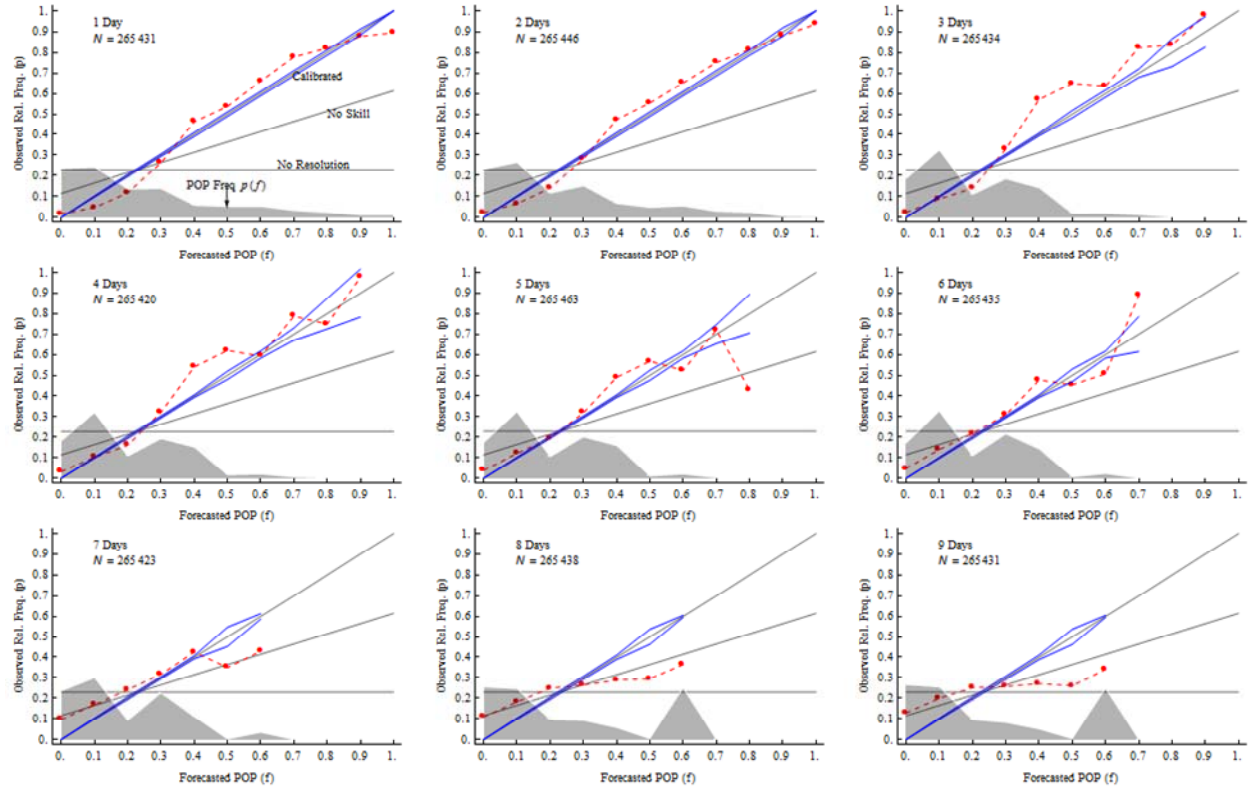


Fig. 3. TWC calibration diagrams for 1da to 9da POP forecasts

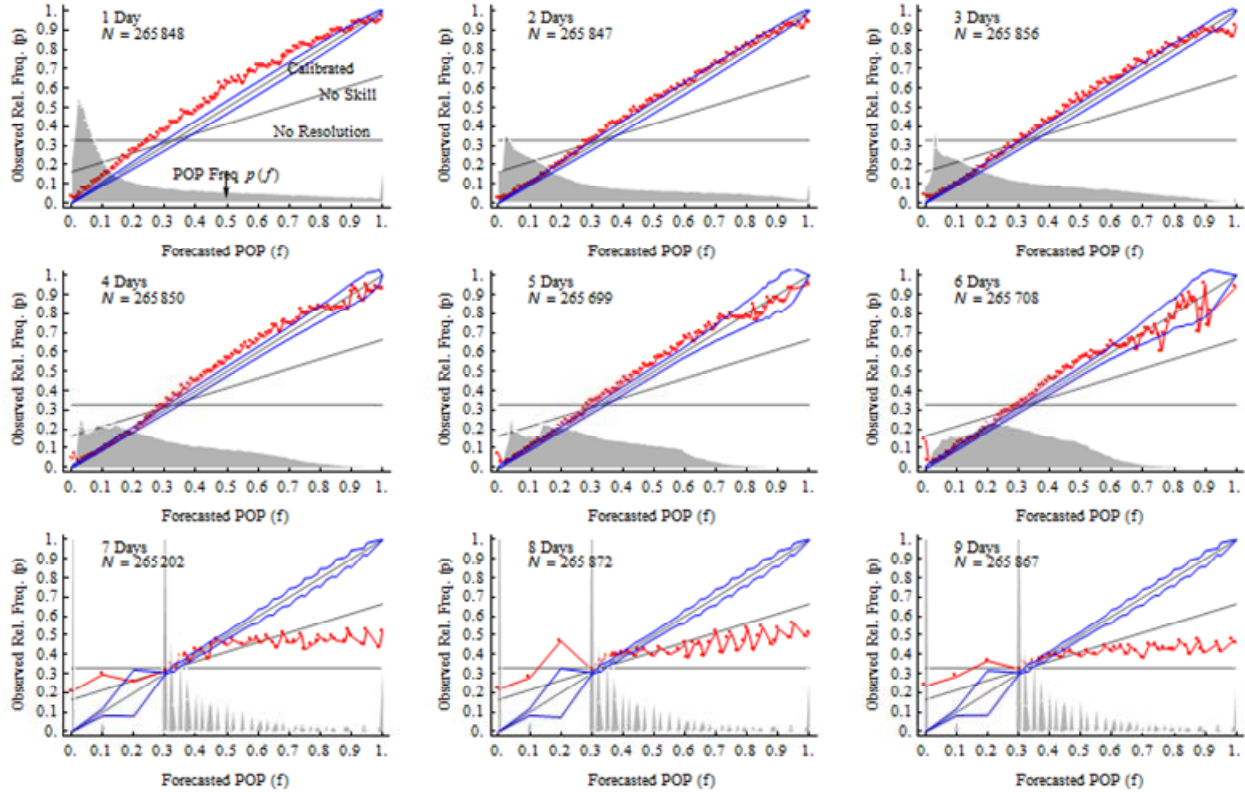


Fig. 4. CW calibration diagrams for 1da to 9da POP forecasts

Fig. 5 presents the MSE and SS for each provider as a function of lead time. TWC and NWS have the lowest MSE, but one must remember that CW is forecasting a 24-hour rather than a 12-hour POP. The NWS's skill score is between 42% and 27% for the 1da to 3da forecasts. For comparison, Murphy and Winkler (1977) found an overall SS approximately 31% for the NWS.

We see that the NWS has a lower MSE and SS than TWC for the 1da to 3da forecasts. By looking at the SS we see that the NWS outperforms CW as well for the 1da and 2da forecasts. CW has a slight advantage for the 3da forecast.

CW dominates TWC for the 2da to 6da forecasts. At this point, CW's forecast skill collapses; CW's SS for

their 7+da forecasts are negative, meaning they are worse than simply forecasting the climatological average. TWC also exhibits negative skill 8 and 9 days ahead.

4.2 Likelihood-Base Rate Factorization

Fig. 6 displays the likelihood functions, $p(f|x=1, I)$ and $p(f|x=0, I)$ for the NWS, TWC, and CW 1da POP forecasts. We see that the providers are skilled at forecasting a lack of precipitation, but struggle to identify precipitation. For example, given that it precipitated ($x = 1$), the NWS was almost equally likely to give a forecast between 0.2 and 1.0. TWC was more likely to forecast a

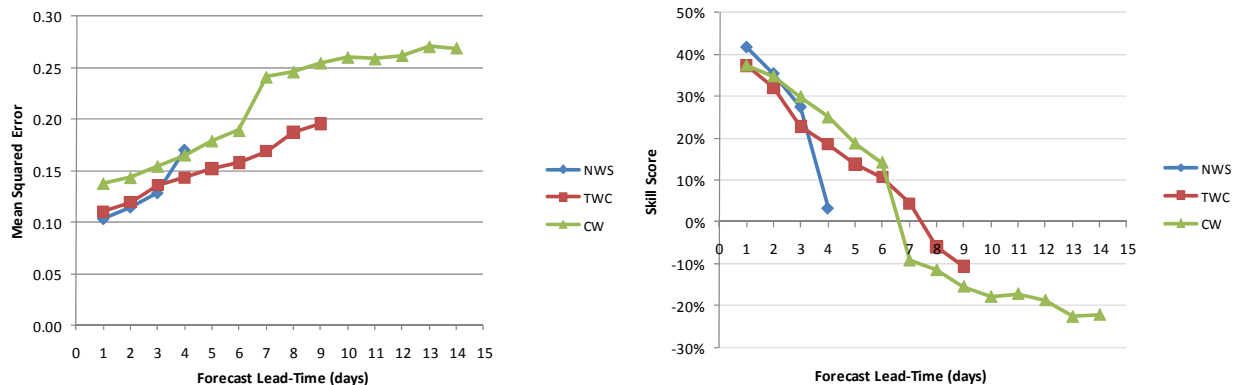


Fig. 5. MSE (left) and SS (right) comparison

mid-range POP in this situation. CW's forecasts, on the other hand, are sharper, as evidenced by the spike at POPs near 1.0. In situations where precipitation was not observed, all three providers were more likely to provide a low POP. CW's performance is especially impressive in this regard.

Figures 7 through 9 display the likelihood functions for each provider and lead time. We notice that as lead

time increases the likelihood functions begin to overlap. For example, CW's 7+da likelihoods are completely overlapping, highlighting the independence of precipitation observations and their forecasts. The likelihood plots also highlight TWC's preference for particular POPs.

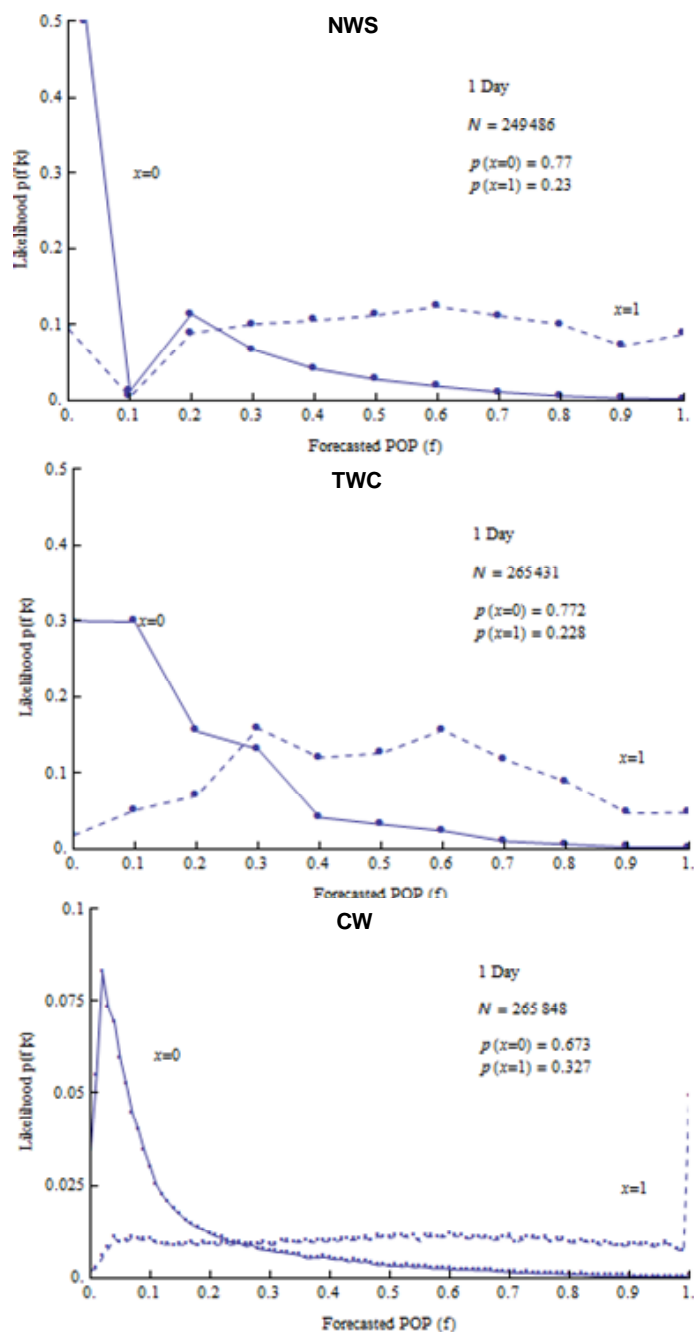


Fig. 6. Likelihood diagrams for NWS, TWC and CW's 1da POP forecasts

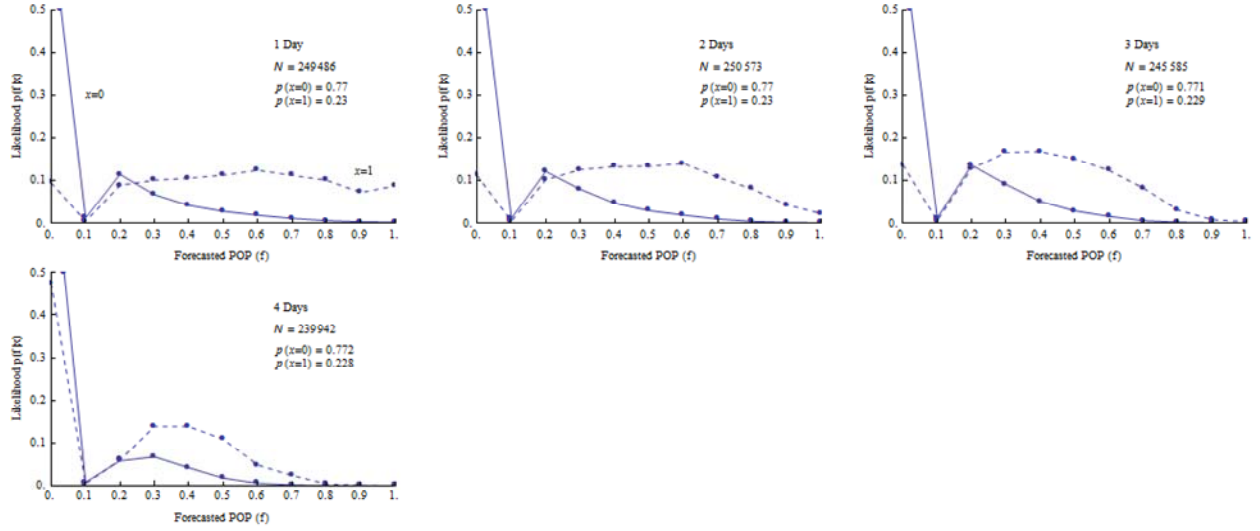


Fig. 7. NWS likelihood diagrams for 1da to 4da POP forecasts

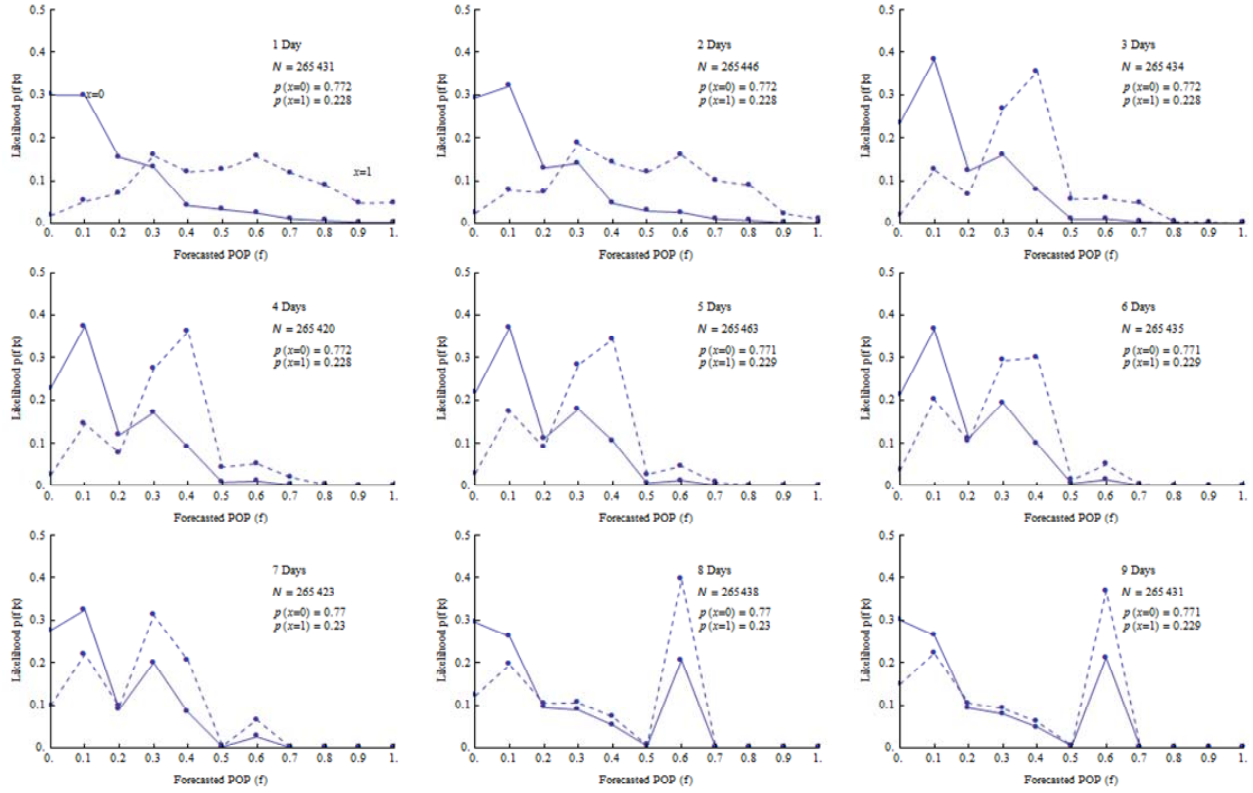


Fig. 8. TWC likelihood diagrams for 1da to 9da POP forecasts

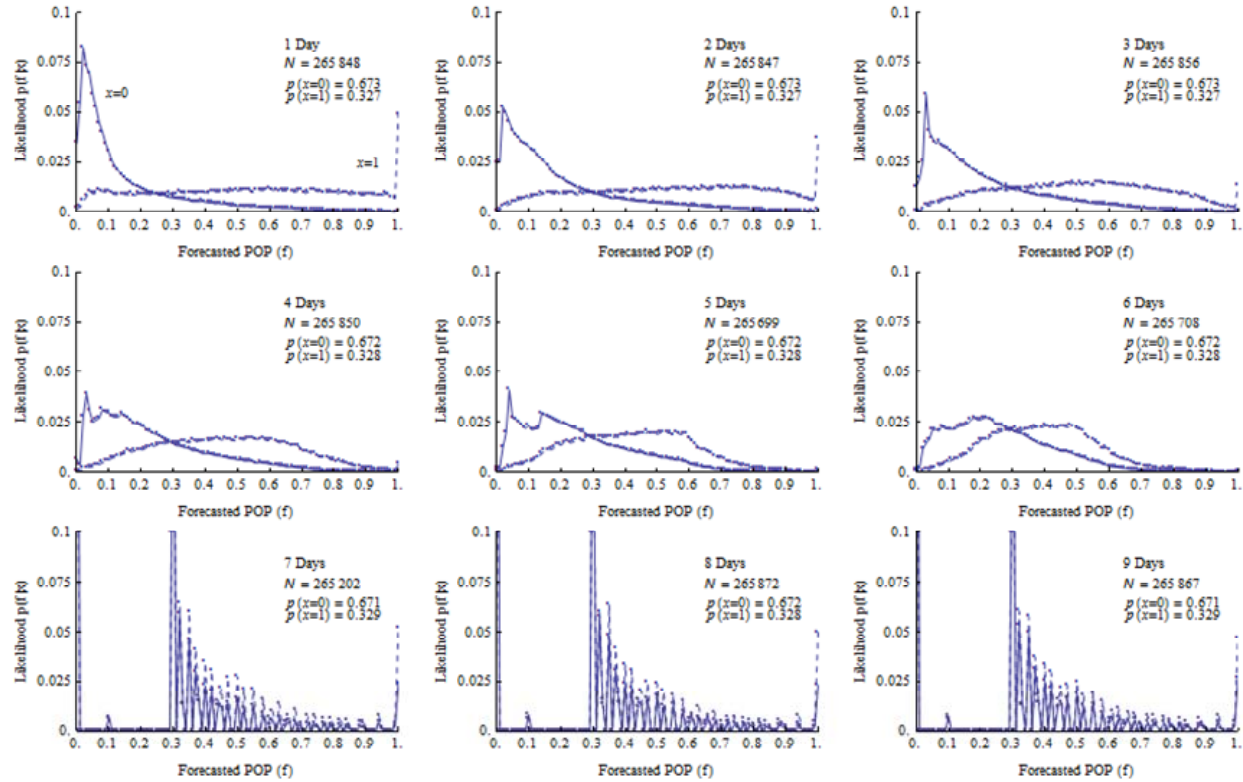


Fig. 9. CW likelihood diagrams for 1da to 9da POP forecasts

4.3 Head to Head Comparison

In this section we only consider forecasts for which all three providers published a valid forecast. This reduces our dataset, but ensures that we compare the exact same forecasting events.

Table 2 summarizes the forecast and observation data. Our lead time is limited to four days, since this is the maximum lead time provided by the NWS. In total, we consider over 2.8 million POPs. For the four days, we have excluded 29,042 POPs (2.9%), 105,187 (9.9%), and 106,857 (10.0%) for the NWS, TWC, and CW respectively. More data is excluded for the TWC and CW because we are now excluding the Alaskan cities from their dataset, in order to match the NWS.

The MEs given in Table 2 are nearly identical to those of Table 1. We see that the NWS's 4da forecast and CW's 1da forecast are quite biased. Again, TWC has the lowest average ME.

The MSE and SS comparison is shown in Fig. 10, results of which are nearly identical to Fig. 5. The NWS's 1da and 2da POP forecasts exhibit more skill than either CW or TWC. However, the NWS's performance begins to decline at 3da and their 4da forecast is quite poor. TWC's 1da to 3da POP forecasts are dominated by both the NWS and CW. CW's 4da forecast also exhibits more skill than TWC.

Table 2. Summary of forecast and observation data for head-to-head competition

Lead Time (Days)	National Weather Service					The Weather Channel					Custom Weather					Lowest Absolute ME
	Number of Forecasts	Precip. Obs. (x = 1)	Avg. POP Forecast	Freq. of Precip.	Mean Error (ME)	Number of Forecasts	Precip. Obs. (x = 1)	Avg. POP Forecast	Freq. of Precip.	Mean Error (ME)	Number of Forecasts	Precip. Obs. (x = 1)	Avg. POP Forecast	Freq. of Precip.	Mean Error (ME)	
1	245 446	56 254	0.200	0.229	-0.029	245 446	55 872	0.240	0.228	0.013	245 446	80 389	0.278	0.328	-0.050	TWC
2	245 902	56 181	0.186	0.228	-0.043	245 902	55 815	0.227	0.227	0.000	245 902	80 373	0.313	0.327	-0.014	TWC
3	240 946	55 137	0.168	0.229	-0.060	240 946	54 774	0.195	0.227	-0.032	240 946	78 807	0.311	0.327	-0.016	CW
4	224 250	50 888	0.100	0.227	-0.127	224 250	50 655	0.193	0.226	-0.033	224 250	73 003	0.303	0.326	-0.022	CW
Total	956 544	218 460	0.165	0.228	-0.064	956 544	217 116	0.214	0.227	-0.013	956 544	312 572	0.301	0.327	-0.026	TWC

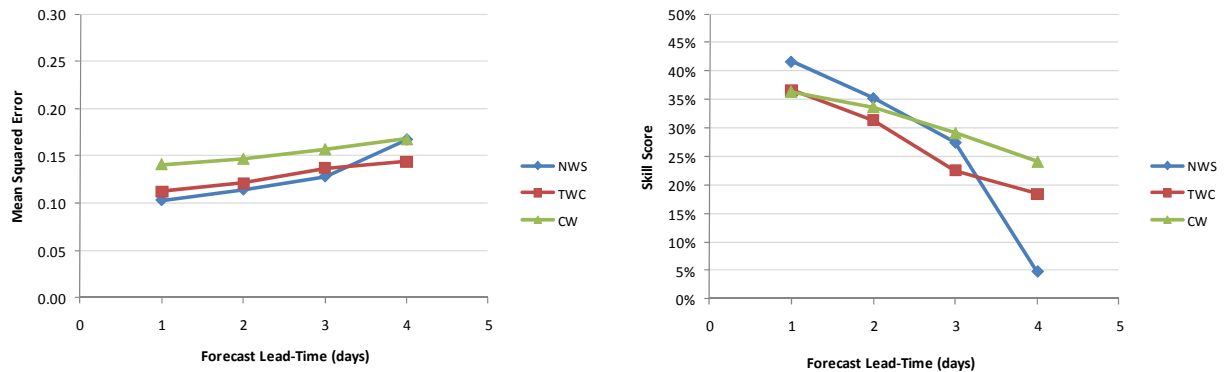


Fig. 10. MSE (left) and SS (right) for head-to-head comparison

5. DISCUSSION AND CONCLUSIONS

The NWS POP forecasts exhibit positive skill for lead times of four days or less. Their one- and two-day ahead forecasts are more skilled than either TWC or CW and, thus, the third party providers fail to add value in this respect. CW is the most skilled from three to six days, after which their performance decreases substantially. In fact, CW's POP forecasts with lead times of 7 days or greater are less skilled than a forecaster that simply reported the climatological average. Thus, users should not rely on these forecasts. TWC is only the most skilled in the case of their 7 day-ahead forecast and their 8 and 9 day forecasts are worse than climatology.

Which forecast should users rely on? This is a rather difficult question. While the NWS has the most skill 1 to 2 days ahead, CW provides 24-hour forecasts at the 0.01 level. If one needs this degree of precision or requires 24-hour forecasts then the results in this paper could be used to de-bias the CW forecasts. If one only needs 12-hour forecasts at 0.1 intervals then the NWS provides the best service for 1- and 2-day forecasts. None of the providers supplies a skilled forecast beyond 6 days and users should probably ignore these long-term forecasts.

References

Bickel, J. E., and S.-D. Kim, 2008: Verification of The Weather Channel Probability of Precipitation Forecasts. *Monthly Weather Review*, **136**(12), 4867-4881.

Brier, G. W., 1950: Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**(1), 1-3.

Hsu, W.-r., and A. H. Murphy, 1986: The Attributes Diagram: A Geometrical Framework for Assessing the Quality of Probability Forecasts. *International Journal of Forecasting*, **2**285-293.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons Ltd, 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. *Applied Statistics*, **26**(1), 41-47.

—, 1987: A General Framework for Forecast Verification. *Monthly Weather Review*, **115**1330-1338.

—, 1992: Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**435-455.

Acknowledgements

The authors would like to thank Geoff Flint of CustomWeather, Dr. Bruce Rose of The Weather Channel, and Ron Jones of the National Weather Service for help in properly interpreting their respective forecasts.