Forecast Verification: Past, Present, and Future

By Julie Malmberg, Western Water Assessment

The goal of this article is to provide forecast users with a framework for assessing the quality of any kind of forecast. Also to this end, WWA is co-sponsoring a workshop on Forecast Verification with NOAA's Colorado Basin River Forecast Center and NRCS on February 19th in Denver. The workshop will provide forecast users with the tools to evaluate the overall quality of the forecast. The workshop will emphasize water supply forecasts in the Western United States but the concepts will be applicable to climate forecasts as well. Please contact Christina Alvord for more information: christina.alvord@noaa.gov.

Forecasts are issued by meteorologists, climatologists, and hydrologists to predict future weather, climate, and streamflows for a wide variety of purposes including saving lives, reducing damage to property and crops and even so people can decide what to wear in the morning. Forecast verification is how the quality, skill, and value of a forecast is assessed. The process of forecast verification compares the forecast against a corresponding observation of what actually occurred or an estimate of what occurred. This article discusses some of the many different forecast verification methods, the concept of forecast value to users, and offers some suggestions for forecast users when considering any forecast.

Overview of Forecasts

The three types of forecasts discussed here are weather, climate, and streamflow forecasts. *Weather forecasts* predict the weather that will occur during a short time frame from six hours to two weeks into the future. *Climate forecasts*, also called climate outlooks, predict the average weather conditions for a season or period from several months to years in advance. Climate forecasts will do not predict the weather for a certain day, but predict the average weather over several days or months. Examples of climate forecasts from NOAA are on pages 13–14 of the January 2008 Intermountain West Climate Summary. *Streamflow forecasts* predict water supply conditions, including streamflow at a point or volume for a period, based upon variables like precipitation and snowmelt. Streamflow forecasts can be daily or seasonal time scales. An example of a streamflow forecast map is on page 17.

History of Forecast Verification

In order to create better forecasts, forecasters monitor the forecasts for accuracy and compare different forecasting techniques to see which is better and why (IVMW, 2007). Weather forecasting based upon interpreting weather maps began in the 1850s in the United States, but serious efforts in forecast verification began in the 1880s. In 1884, Sergeant John Finley of the U.S. Army Signal Corps began forecasting tornado occurrences for 18 regions east of the Rocky Mountains. His forecasts were made twice a day and would be either "Tornado" or "No Tornado". This is an example of a dichotomous forecast, where there are only two possible choices. He reported a 95.6-98.6% accuracy for the first three months. However, other scientists pointed out that, ironically, he could have had 98.2% accuracy if he forecasted "No Tornado" for all the regions and all the time periods. A 10-year debate started after Finley's publication, referred to as "The Finley Affair." This debate made forecasters realize the need for valid verification methods in order to improve forecasts, and led to the development of verification methods and practices (Murphy, 1996).

Types of Verification

In order for a forecast to be verified, it must be compared with some "truth." Observational data such as rain gauges, thermometers, stream gauges, satellite data, radar data, eyewitnesses, etc. are used as "truth." In many cases, however, it can be difficult to know the exact "truth" due to instrument error, sampling error, or observation errors. Accurate observations and observation systems, then, are critical to forecast verification.

Forecasters and forecast users have many different ways to verify forecasts and assess quality. Two of the traditional ways are looking at the *accuracy* and the *skill* of the forecast. *Accuracy* is the degree to which the forecast corresponds to what actually happened (i.e. "truth" data) and depends on both the forecast itself and the accuracy of the measurement or observation. As mentioned above, observation data can be a limitation in

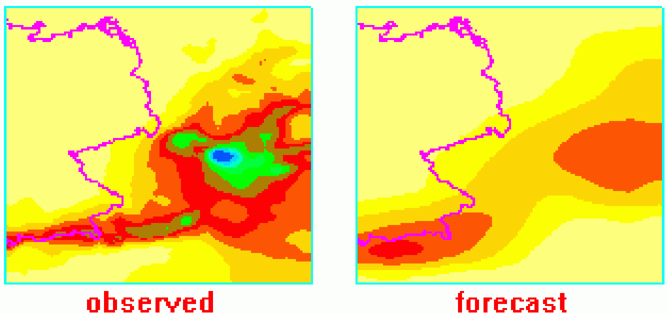

Figure 1a. Observed data versus forecast data (IVMW 2007).

		Observation		
		Yes	No	Total
Forecast	Yes	hits	false alarms	forecast yes
	No	misses	correct negatives	forecast no
	Total	observed yes	observed no	Total

Figure 1b. A contingency table shows what types of errors are being made. A perfect forecasting system would only produce hits and correct negatives.

all verification measures, not just accuracy. In addition, the person verifying the forecast uses expert judgment to decide what makes a forecast accurate. For example, a forecast for a high temperature of 75°F might be considered inaccurate either when the observed high temperature was 76°F or when the high temperature was 85°F.

The second common forecast verification measure is *skill*. Skill is the accuracy of a forecast over a reference forecast. The reference forecast might be random chance, persistence forecasts, climatology, or even another forecast. A random chance forecast would be like flipping a coin to decide whether or not to forecast precipitation. Persistence forecast is forecasting the same conditions that are happening at the time of the forecast. For example, if it is currently snowing, a persistence forecast is for snow to continue. A forecast of climatology is forecasting the average conditions for the forecast period. A "skillful" forecast must show improvement over a reference forecast.

Other measures of forecast quality besides accuracy and skill include bias, resolution, and sharpness. *Bias* measures if forecasts on average are too high or too low relative to the truth. *Resolution* measures the ability of a series of forecasts to discriminate between distinct types of events, even if the forecast itself is wrong. *Sharpness* indicates if the forecasts can predict extreme values. Sharpness is important because forecasters can sometimes achieve high skill scores by predicting average conditions but in some cases the occurrence of extreme events may be more important to users. In general, focusing on just one measure of forecast quality may be misleading. For example, in the case of Findley's forecasts, their apparent high accuracy obscured the fact their skill was less than a constant forecast of no tornado.

Methods of Forecast Verification

Forecast verification methods are chosen depending on the type of verification (accuracy or skill) and the type of forecast (dichotomous, continuous, probabilistic, etc.). Examples of verification methods range from simply "eyeballing" the forecast compared to observations, to statistically and numerically advanced methods.

Eyeballing a forecast is as simple as it sounds and can be use for a variety of forecasts. A forecaster simply looks at the forecast and the observations side by side to see how well they match up (Figure 1a). "Eyeballing" verification is very subjective and can lead to different outcomes depending on the judgment of the individual forecasters looking at the data.

A contingency table is typically used to verify dichotomous forecasts, like the tornado example above, over a period of time. The table shows the "yes" and "no" forecasts and observations (Figure 1b). To find the accuracy of the forecasts, one must sum "hits" and "correct negatives" and divide by the "Total". This will give a number between 0 and 1; the closer to 1, the more accurate the forecast. This type of score can be very misleading in rare events when forecasting "No" will lead to a high "correct negatives" category such as the occurrence of tornados as in the Findley Affair. Numbers in the contingency table can be combined in many other ways than just accuracy. For example, the False Alarm Ratio is the number of events that were forecasted to occur but did not.

One can numerically verify or calculate the error between the forecast and the observed values with the help of graphical representations. Graphical displays, such as scatter or box-and-whisker plots, are used to verify forecasts of continuous variables such as maximum temperature over a period of days. Scatter plots show the observed amount plotted against the forecast amount. An accurate forecast in this case would lie along the diagonal of the scatter plot. Box-and-whisker plots can show the distribution of the observed values relative to the forecasted values, which can provide a measure of the resolution of the forecast. In a well-resolved forecast, the box plot of the forecast would appear to have the same spread as the observed values.

Skill scores can be calculated for almost all types of forecasts, but they are most often used for categorical and probabilistic forecasts, like the seasonal climate outlooks issued by NOAA's Climate Prediction Center (CPC) (see pages 13 and 14). All skill scores measure the fraction of correct forecasts to total forecasts after correcting for the number of correct forecasts a reference forecast – generally persistence, climatology or random chance – would obtain. Three types of skill scores are the Heidke skill score, the Brier skill score, and the Ranked Probability skill score. A score between negative infinity to 1 is calculated, with 1 being a perfect score. If forecasts are consistently better than the reference forecast, the score will be closer to 1, a score of 0 indicates no improvement over the reference forecast, and a negative score indicates the forecast performs worse than the reference forecast. Note that perversely a high negative score may actually provide considerable value if the forecast can be 'inverted'. For this reason, substantial negative skill scores are rarely seen. When comparing skill scores for different forecasts, it is important to use the same method for all forecasts. For example, if you want to compare the CPC seasonal forecast to Klaus Wolter's experimental seasonal guidance, make sure you are looking at either the Heidke or Brier skill score for both.

Forecast Value and Forecast Users

Another important attribute of forecasts is *value*. A forecast might be highly accurate, skillful, unbiased, sharp and well resolved and still not be very useful. A valuable forecast best helps a decision maker. For example, a forecast of clear skies over a desert is probably not very helpful. On the other hand, if a forecast helps a decision maker to gain some benefit, the forecast is considered valuable. Accurately forecasting a drought will help water managers to better prepare for low water supply. Forecasting the April 1st snowpack as early as possible would help improve the annual water management operations. In essence, useful forecasts need a wide variety of attributes including accuracy, skill and value.

NOAA is creating ways to educate decision makers and create better consumers of forecasts. Making forecast verification measures available and explaining the techniques to users will increase the value of forecasts. For example, the Forecast Evaluation Tool and the new verification tools on the NOAA National Weather Service Western Water Supply Application Suite both make verification tools readily available to users (see box). Users will be able to decide which forecasts they want to use for what purpose, and will know the weaknesses, strengths, or biases of particular forecasts. For example, a certain forecast might tend to predict wetter conditions in the spring.

Verifying a forecast should ultimately lead to improvement in the forecasting techniques and an increase in value to the users. Overall, forecasters are starting to understand that they need to think about who is using their forecasts and the value of the forecast to the users, not just the skill score or the accuracy of a forecast. While accuracy is very important, it is not the only element of a good forecast. Whether a forecast is for weather, climate, or streamflows, a user should know what information the forecast provides, how the forecast is verified, and limitations of the forecasts and verification methods. If users are educated about forecasts and forecast verification, they will ultimately be better consumers of those forecasts.

References

Murphy, A.H. 1996. The Finley Affair: A Signal Even in the History of Forecast Verification. *Weather and Forecasting*. 11(1): 3-20.

Third International Verification Methods Workshop (IVMW). 2007. Reading, UK. Available online: http://www.bom.gov.au/bmrc.wefor/staff/eee/verif/verif_web_page.html.

Forecast Verification Websites

Two online tools help make forecast verification techniques accessible and understandable to users: the Forecast Evaluation Tool (FET) for NOAA/CPC seasonal climate outlooks and the NOAA National Weather Service (NWS) Western Water Supply Application Suite for their water supply forecasts.

Forecast Evaluation Tool

FET is an online application to look at the successes of CPC seasonal climate forecasts by climate division, season, and lead time of the forecast. Holly Hartmann, a scientist working for CLIMAS, a NOAA RISA program at the University of Arizona, found that forecast users were hesitant to make decisions based upon forecasts without knowing the track record of forecasts. She then initiated FET. In order to use FET, register for free at http://fet.hwr.arizona.edu/ForecastEvaluationTool/. A tutorial is available at the web page. For more information about FET, see the January 2006 Intermountain West Climate Summary.

NWS Western Water Supply Application Suite

The NOAA/NWS Western Water Supply Application Suite launched in January 2008. This brand new tool allows users to select a state, river, and station and then visualize data and also calculate error statistics and skill statistics. The web page is available at: http://www.nwrfc.noaa.gov/westernwater/. To access the verification section, when you get to the web page, first select "Change Application" and then select the "Verification" tab. At this point, the regional data can be entered. More information is also available by selecting the "About Western Water Supply" tab and then the "Verification" tab.