How to use the climate Forecast Evaluation Tool: Web-based method
yields quick way to test accuracy of seasonal predictions

By Melanie Lenart of the Climate Assessment of the Southwest (CLIMAS)

(A version of this article first appeared in the November 29, 2005 edition of the Southwest Climate Outlook, a CLIMAS publication, available at: http://www.ispe.arizona.edu/climas/index.shtml)

The article describes a tool for users to evaluate climate forecasts, which may be useful for our readers in the Intermountain West Region.

"I could do better by flipping a coin." If this thought has ever crossed your mind while considering a climate forecast, you can test your theory objectively using the web-based Forecast Evaluation Tool (FET). The tool allows for an on-line examination of the successes and failures of past forecasts by climate division, season, and lead time of the forecast.

The Forecast Evaluation Tool grew under the tutelage of Dr. Holly Hartmann based on interviews she conducted with regional decision-makers for The University of Arizona's Climate Assessment for the Southwest (CLIMAS), a program funded by the National Oceanic and Atmospheric Administration (NOAA). Stakeholders revealed that they were hesitant about basing decisions on seasonal climate forecasts without knowing the track records of the forecasts. Western Water Assessment has seen similar attitudes among water managers in Colorado.

With support from a half-dozen other agencies over the years, Hartmann and her team responded by designing the FET to provide customized comparisons of climate forecasts. Although the website continues to evolve and the tool is still under development—it is considered a "beta-test" version—the FET now can compare all forecasts made since 1994 by the National Weather Service's Climate Prediction Center (CPC), the NOAA

branch that issues official government forecasts. Future plans call for similar testing of forecasts issued by other agencies, as well as testing of projections for streamflow (water transport in rivers).

This article serves as a set of easy instructions designed to guide you through the process of using the FET for the first

Figure 1a: FET homepage (http://fet.hwr.arizona.edu/ForecastEvaluationTool).

time to check the performance of the CPC climate forecasts you consider most relevant.

Getting started

Go to the website http://fet.hwr.arizona.edu/ForecastEvaluationTool/ (Figure 1a). Register for the confidential service by providing your name, organization, and email address and choosing a login name and password. After you submit your registration information, you should be able to sign in with no wait. In time, users will have the option to save their evaluation work and other climate information for future reference. Use of the FET is free

of charge and registration information will not be shared with any other organization.

Download Java

Many new computers already have Java installed. If yours doesn't, Java offers a free download of the Sun Java Runtime Environment program (237 kilobytes) needed to show the results of the evaluations. You can access a link to the Java website directly from the FET website. Choose the correct program for your system and follow the installation instructions. Once the program is installed, return to the FET website.

Interpreting climate forecasts tutorial

An optional tutorial introduces users to the concepts and terminology of CPC forecasts. For instance, the tutorial brings home the important point that an Equal Chances or "EC" forecast is tantamount to no forecast at all. To make sure you're interpreting CPC forecasts properly, you can take the five-question self-test at the end. As soon as you submit your answers, you'll see your score as well as the correct answers.

Seasonal climate forecasts use a tercile approach. They consider the probability that climate conditions will fall into one of three categories: above-average, near-average, or below-average. Average is relative to actual conditions observed during a 30-year period—from 1971 through 2000.

Each of the 30 baseline seasons (or years) is divided equally into these three categories, with 33 percent labeled above-average, 33 percent called near-average,
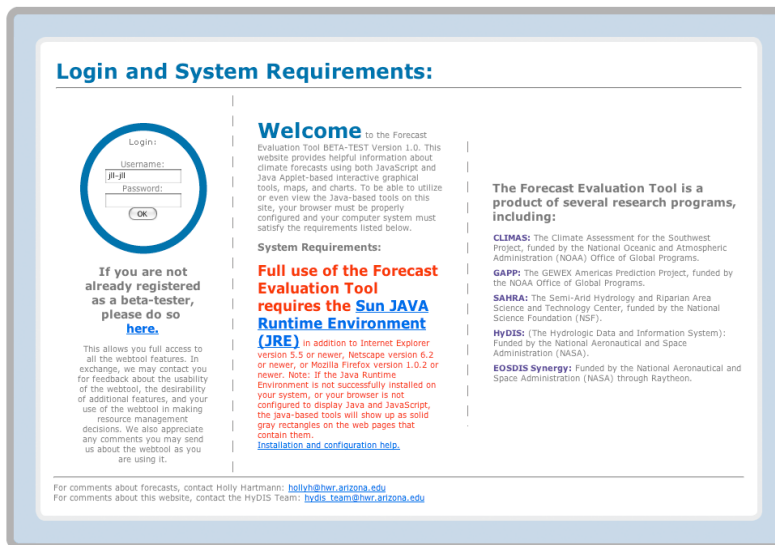
(Continued from p.2)

and 33 percent considered below-average. For example, a forecast that calls for a 40 percent probability of above-average temperature is less certain than a forecast that calls for a 70 percent probability of above-average temperatures. In both cases the projection is for temperatures to fall into the above-average tercile as compared to the actual conditions observed from 1971 through 2000.

White space on the map indicates Equal Chances (EC) of falling into any of the three terciles (i.e., no forecast). Only rarely does the CPC issue a forecast predicting near-average temperatures, indicated by gray shading.

Climate forecast performance

On the FET home page, you'll also see options to "Explore the Forecasts," to consider "How do the forecasts relate to my specific situation?" and to evaluate "Forecast Performance." Select "Forecast Performance" to follow the example here.

This is where you can test and compare how CPC forecasts have performed in the past, based on the forecasts issued since 1994. Here we take a step-by-step approach to testing a seasonal forecast's success:

1. The "National Weather Service Climate Prediction Center" option is automatically selected, so there's no need to do anything. (In the future, other options will become available.)

2. Select NWS CPC seasonal climate outlooks (contiguous states).

3. Select precipitation.

4. Select a forecast season, in groups of three months, by sliding the shaded box with your cursor and then clicking on it. The months are listed by their first initial only. Choose DJF to get the three-month seasonal outlook for December, January, and February. The selected grouping will show up below the shaded area as DJF. (If you want to do more than one three-month period, click your mouse upon each selection and you'll see the selected months listed below.)

5. Select the month or months during

which the forecast was issued. Click in the boxes for each year you want. We'll select N (November) for each available year (1994–2004). The three-month seasonal forecasts are issued up to a year in advance and updated every month.

6. You now have the opportunity to select the type of statistical test you'd like to apply to the forecasts. Select the "False Alarm Rate" option. Brief descriptions of the other options (e.g., Probability of Detection, Brier Score) are included at the end of this article.

7. Once you have made your choices, hit "Submit" to launch the program. When the results appear, read the box at the top under "You Chose" to make sure the computer accurately recognized all your choices. (For example, if you did not click on your season selection, the default "All Seasons" will appear.)

8. The results will include national maps color-coded by division and a color bar below that explains the legend (Figure 1b). For these comparisons, the 344 NOAA climate divisions have been grouped into 102 larger divisions. Colorado Wyoming, and Utah have eleven total divisions under this system, with some divisions that overlap other states. You can see the actual value for a climate division by holding your cursor over it.

Frequency of Forecast Results

Regardless of which category you select, you will first see a map indicating the Frequency of Forecast Results. This shows how often a forecast was actually made about the season of interest by climate division. A value of 0.322 means a forecast covered some or all of the division about 32.3 percent of the time since 1994, when forecasts were available more than one month ahead. Scroll down to see the results you were seeking.

False Alarm Rate

This comparison considers how often the projected forecast turns out to be wrong, using the category that was predicted to be most likely. To convert

the resulting climate division score into a percentage, just multiply the value by 100. So if forecasters called for wet conditions three times, but they only occurred twice, the false alarm rate would be 0.333 or 33 percent. Note that, in this case, low scores are good. To consider how often an issued forecast was accurate, just subtract the False Alarm Rate score from 1 (or the percentage from 100). In this theoretical example, the forecast was accurate 66 percent of the time. In the actual example tested here, scores ranged from 0.5 to 0.857 for "wet" conditions and from 0 to 0.75 for "dry" conditions (Figure 1b). Water managers have indicated they find the False Alarm Rate particularly relevant.

Show Data Behind the Map

If you want to see the forecasts that were considered for the evaluation, click on a climate division of interest and then click on the "Show the Data Behind the Map" option. First you'll see a description of how to interpret bubble plots, including a sample bubble plot. Then you'll see the data used for the climate division of interest for the season(s) and years indicated.

Besides the False Alarm Rate, there are a number of other options available for evaluating forecasts. To try other techniques, return to the Climate Forecast Performance page. (If you can't find it, return to the FET homepage and select "Forecast Performance.")

Modified Heidke Score

This selection is intended for use by the National Weather Service (NWS) forecasters who have historically used this approach to evaluate forecasts. It is included on the FET site because NWS forecasters receive instruction in use of this tool as part of their ongoing climate training courses, as explained by NWS Climate Services Chief Robert Livezey. He feels that for those not familiar with the Heidke system, the other methods provided (e.g. Frequency of Forecasts, Probability of Detection, False Alarm Rate, Brier Score,
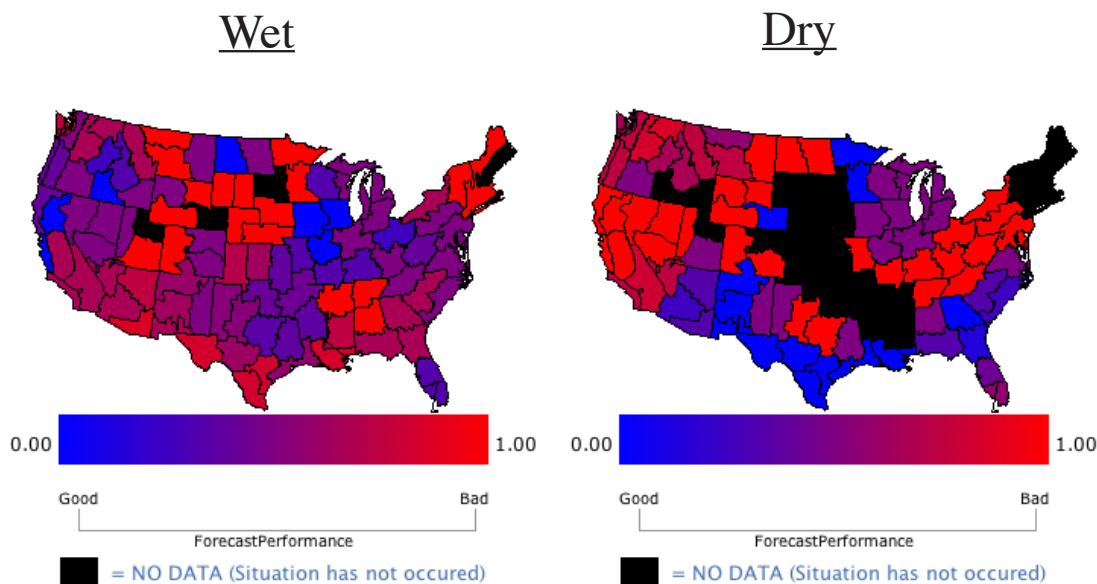
(Continued from p.3)

Wet

Dry

Figure 1b: An example result of the Forecast Evaluation Tool. The False Alarm Rate results for climate forecasts issued in November for the December-February season. Southeastern Utah's and south-central Colorado's winter forecasts tended to be the most successful of all the climate divisions in Colorado Wyoming and Utah, especially for predicting drier than average conditions (map at right). For example, the 0 scores in those divisions indicate that every forecast for dry conditions on the last decade panned out. Forecasts for wet winters in the Intermountain West region only came to pass about half the time or less (map at left).

and Ranked Probability Score) are more useful to understand the forecast performance.

Probability of Detection

This analysis indicates how often a forecast was made for non-average conditions compared to the total number of times it actually occurred. Your results will include separate maps for forecasts of above-average events (wet or warm) versus below-average events (dry or cool). To convert the resulting climate division score into a percentage, just multiply the resulting value by 100. A score of 0.346 for detecting wet conditions for the selected season means the CPC issued a forecast calling for above-average precipitation in about 34.6 percent of the cases in which precipitation tallies registered as above-average. Emergency managers have indicated they find these scores useful.

Ranked Probability and Brier Scores

While the Brier score differentiates categories into wet and dry (or warm

and cool), the Ranked Probability score provides one lumped result for both conditions. Other than that, they have similar features. Both scores take into consideration the strength of the issued forecast. So, if above-average conditions prevail as the CPC had predicted, a forecast issued with a 70 percent probability gets a higher score than one issued with a 40 percent probability. Similarly, the 70 percent probability forecast takes a bigger penalty than the 40 percent probability if conditions turn out to be average—and an even bigger hit if conditions turn out to be below-average.

The Brier and Ranked Probability skill scores represent the proportion of time above and beyond what would be expected by chance (33 percent). That's partly why a climate division with a Probability of Detection score of 0.517 can translate into a Brier skill score of 0.086. This also explains why some of the skill scores turn up negative, indicating the viewer theoretically could have done better just by flipping a three-sided coin.

Customize your options

Now you have the know-how to consider how forecasts fare during a variety of seasons with a number of different lead times, using evaluation approaches that suit your needs. The website has many other features to explore on your own.

Want to know more?

If you have any questions about how the website works, you can send an email to: hydis_team@hwr.arizona.edu.