FORECASTING WEEK.2 FORECAST SKILL

Wilbur Y. Chen
Climate Prediction Center
NCEP/NWS/NOAA

The skill level of week.2 forecasts and physical ways to raise the level of skill are investigated using the NCEP's daily operational ensemble forecasts. The anomaly correlation (AC) skill scores over the North Pacific/North American sector are evaluated. Occasionally, a week.2 forecast (Day 8-14 mean) can turn out to be an outlier, as shown in Figure 2, where the red curve in the upper panel displays the AC skill scores during October 21 2003 to March 5 2004. On Day 24, the Week.2 forecast skill was down to less than -0.8. Catching an outlier at forecast time, such as this one, is a challenging task. Looking into various objective ways for identify an inferior forecast is the subject of this investigation.

Figure 1 gives an illustration: the conventional ensemble-mean forecast of T850 versus an enhanced version by objectively catching and eliminating the inferior forecasts among the ensemble members. The upper-left panel is the spaghetti chart of all 12 perturbation runs. The panel below it is their ensemble mean. The northwest Pacific region is seen here to be near neutral, being evened out by conflicting forecasts of positive and negative anomalies. However, after elimination of inferior forecasts, as shown in the right-hand panels, this region becomes anomalously positive. The verification (the bottom panel) proves that, indeed, the objectively inferior-forecast-eliminated 3-member ensemble mean is a much better one.

Contact: Wilbur.Chen@noaa.gov

As shown in Fig.1, the 12-member ensemble-mean forecast yields only 0.1 in AC skill score, while the objectively improved 3-member ensemble-mean forecast turns out to be better than 0.6.

Three objective ways to identify the inferior forecasts are introduced here. They are:

1) The internal persistency of an individual perturbation run. As shown in Fig.2, there is a similarity between the AC skill scores and the time series of persistence from Day (5-11) to Day (8-14). Therefore, a low persistency criterion can be used to catch an inferior forecast. The bottom panel presents the result. Eight out of 129 forecasts were objectively identified as inferior forecasts (the red squares, and the red horizontal line represent their average), among which only one turns out to be a false alarm. The blue line represents the average of 129 original ensemble-mean forecasts. The green lines represent the remainders and their average. The red and blue average lines are about 40 points apart. It is evidently clear that this objective way of catching the inferior forecasts is effective.

2) The persistency between the current week.2 forecast and the forecast made one day earlier. As shown in the upper panel of Fig.3, there is also a similarity between these two time series. This feature provides us with another tool to catch an inferior forecast. As shown in the lower panel, nine out of 129 forecasts were identified as inferior, among which only one turns out to be a false alarm. Again, the red and blue horizontal lines are

about 30 points apart. This objective criterion also works nicely.

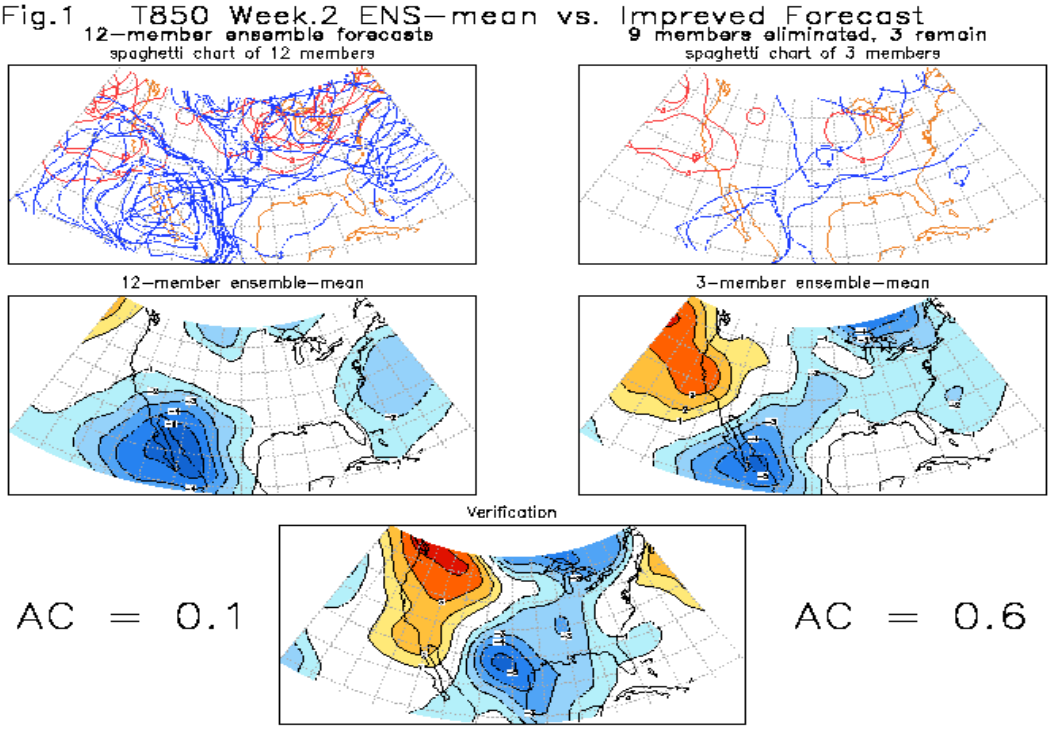
3) Similar to the above method, except for resorting now to the persistency between the current week.2 forecast and the one made two days earlier. Seven out of 129 forecasts were identified as inferior, among which only one turns out to be a false alarm. Again, the red and blue horizontal lines are about 40 points apart, indicating another effective criterion for catching an inferior forecast.

Figure 5 is a repeat, provided for the purpose of judging how much independence these three methods offer for identifying the inferior forecasts. Although there are some over-lapping, the figure does show an additional value from application of each method. If we apply all three methods on each individual perturbation run, there are 16 out of 129 forecasts being identified as inferior, as shown in the top-left panel of Fig.6. There are only two false alarms versus 14 correct calls. Most importantly, the average of red squares (the red horizontal line) is way below the average of the original 129 forecasts (the blue line). They are about 40 points apart, implying a very effective objective scheme for catching and eliminating the inferior forecasts among members of an ensemble forecast.

Figure 6 displays the results of a test on each of the 12 perturbation runs, which are carried out operationally every morning at NCEP. Although some of the distances between the red and blue lines are shorter than 40 points, it is nevertheless evidently clear that the objective techniques are effective in forecasting the inferior Week.2 forecasts.

Similar tests on ensemble runs for other NDJF winters (2000/01, 2001/02, and 2002/03) have also been investigated. Although not shown here to save the valuable space, they are all showing similar results like Fig. 6, indicating an effectiveness of this forecasting forecast skill scheme.

Fig.1 T850 Week.2 ENS—mean vs. Impreved Forecast



AC = 0.1

AC = 0.6

Fig.2

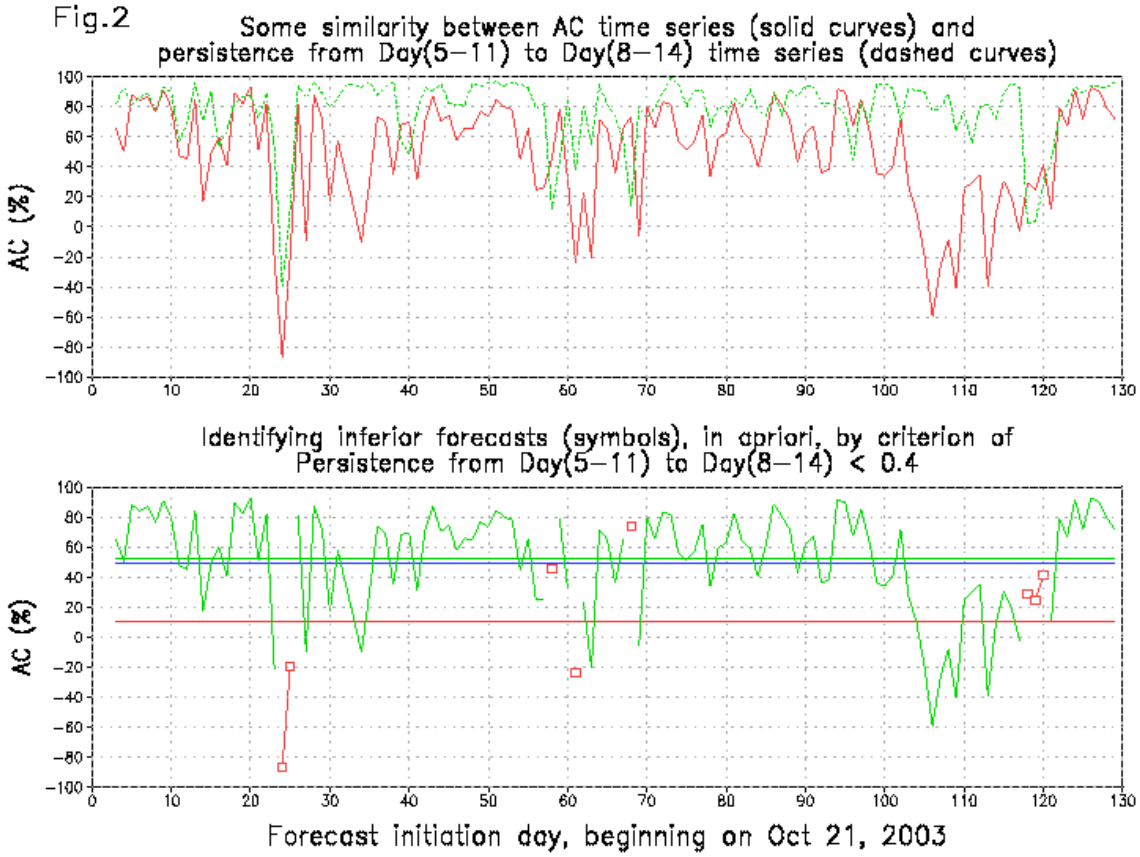

Some similarity between AC time series (solid curves) and
persistence from Day(5−11) to Day(8−14) time series (dashed curves)
AC (%)
Identifying inferior forecasts (symbols), in apriori, by criterion of
Persistence from Day(5−11) to Day(8−14) < 0.4
AC (%)
Forecast initiation day, beginning on Oct 21, 2003

Fig.3

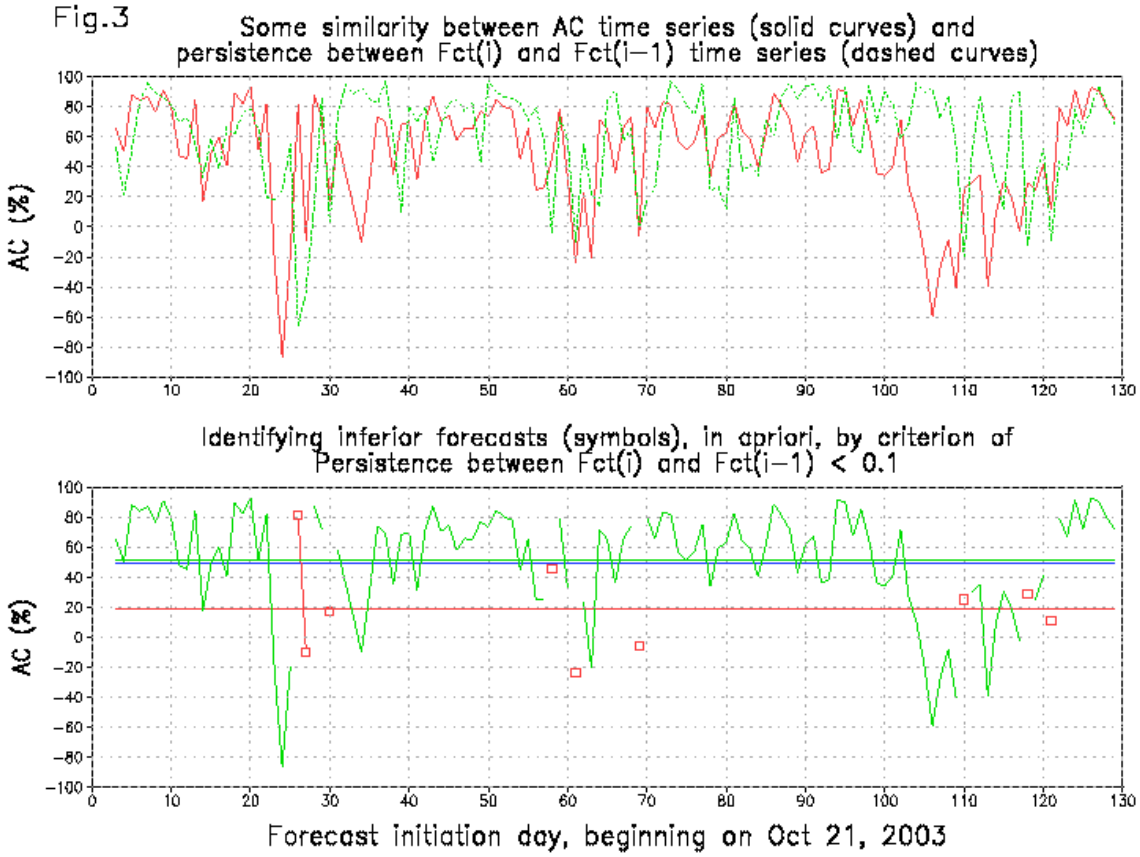

Some similarity between AC time series (solid curves) and
persistence between Fct(i) and Fct(i−1) time series (dashed curves)
AC (%)
Identifying inferior forecasts (symbols), in apriori, by criterion of
Persistence between Fct(i) and Fct(i−1) < 0.1
AC (%)
Forecast initiation day, beginning on Oct 21, 2003

Fig.4

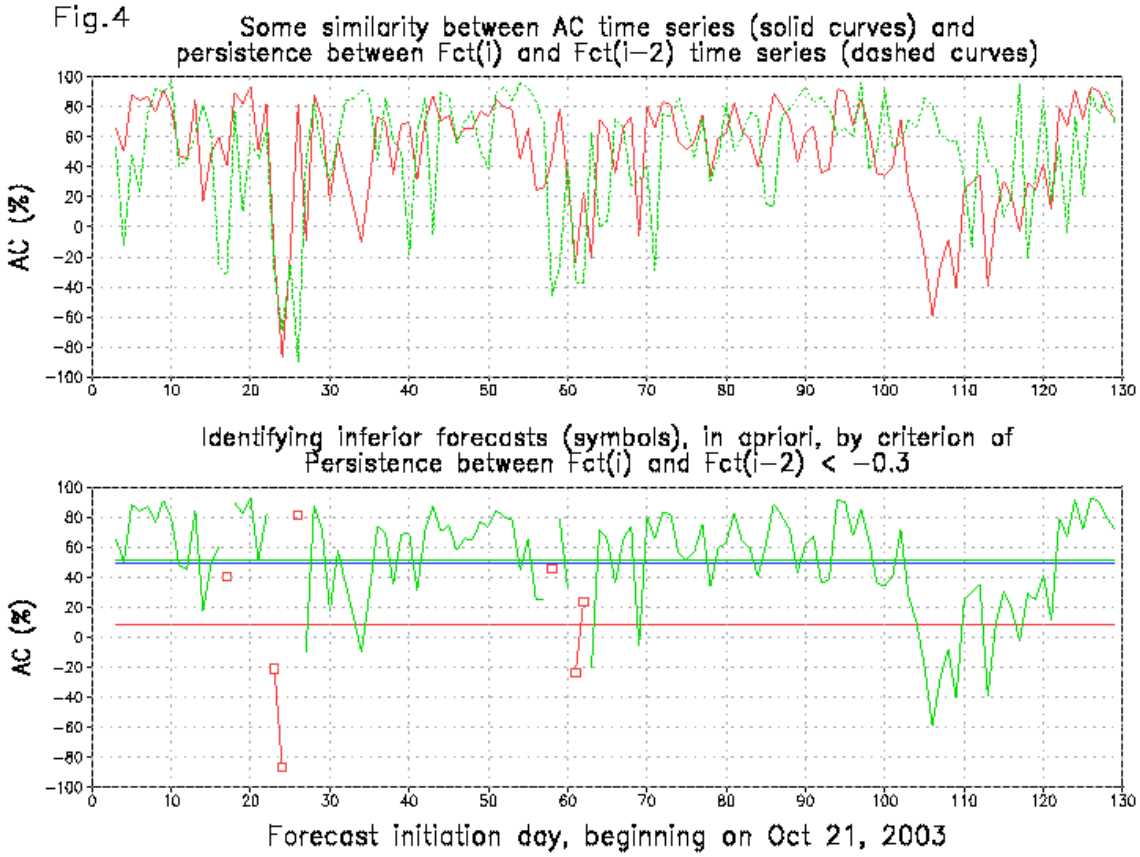

Some similarity between AC time series (solid curves) and
persistence between Fct(i) and Fct(i−2) time series (dashed curves)
AC (%)
Identifying inferior forecasts (symbols), in apriori, by criterion of
Persistence between Fct(i) and Fct(i−2) < −0.3
AC (%)
Forecast initiation day, beginning on Oct 21, 2003

Fig.5

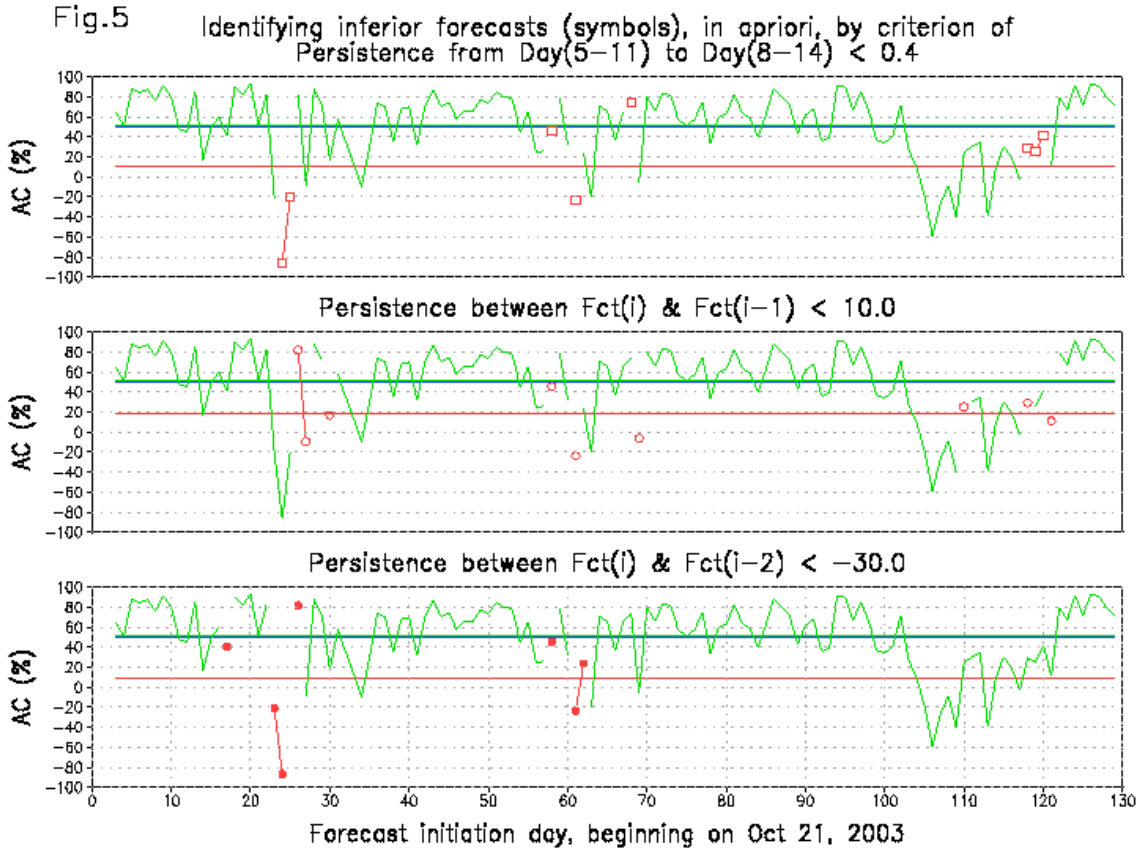
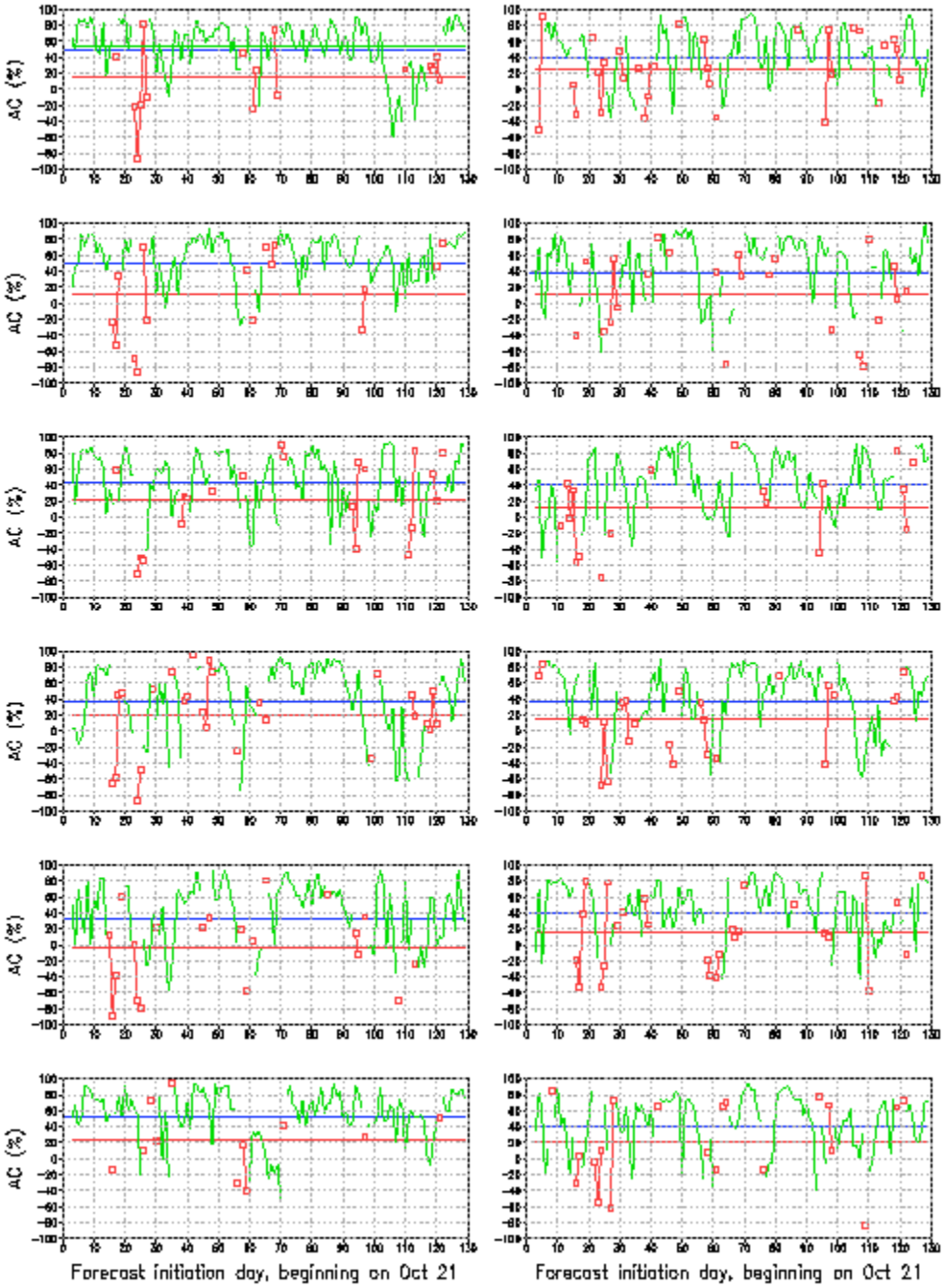

Identifying inferior forecasts (symbols), in apriori, by criterion of
Persistence from Day(5−11) to Day(8−14) < 0.4
Persistence between Fct(i) & Fct(i−1) < 10.0
Persistence between Fct(i) & Fct(i−2) < −30.0
AC (%)
Forecast initiation day, beginning on Oct 21, 2003

Fig.6 Identifying inferior Week.2 forecasts (squares) in apriori

blue=ave of 129 fcsts; red=bad fcsts identified, NDJF 2003/04