8A.6 Is Change Good? Measuring the quality of updating forecasts.

Tressa L. Fowler [*]
National Center for Atmospheric Research, Boulder, Colorado

1 INTRODUCTION

The question of consistency of updated forecasts through time often comes up in the context of weather events such as hurricanes or high wind days. For a single event, forecasts are made and updated as the time of the event nears. The consistency of these updates is important to many users, though some users find this quality desirable while others do not. Historically, the consistency of weather forecasts through time has not been considered or measured. Recently, several authors have constructed consistency measures for specific forecast types, including ensembles (Zsoter *et al*., 2009), precipitation (Ehret, 2010), operational forecasts (Ruth *et al*., 2009), and Markov chains (McLay, 2010). The forecast verification community may also find it useful to have simple and widely applicable measures of forecast continuity. Such measures are already in use in other areas of forecasting, such as economics (Clements, 1997). This paper considers the use of two such measures for weather forecasts.

For many users, consistency in forecasts through time is a desirable quality. If updating forecasts change much or often, a user may believe they are of low quality, possibly even random. This is particularly an issue for decision makers who create plans based on early forecasts, then must change their plans repeatedly as new forecasts arrive. “The consistent high expense of the volatile sequences is evidence that the run-to-run volatility or ‘jumpiness’ (Zsoter *et al*., 2009) that is so disliked by forecasters can have a quantitatively meaningful impact on the decision process.” (McLay, 2010).

However, many users see consistency in forecasts as evidence of a poor forecast. In both statistical and numerical weather modeling, the errors will ideally be noise, with no structure. Consistency in the forecasts indicates structure in the errors and thus suggests room for forecast improvement. The question of whether forecast consistency is desirable or not will not be addressed further here. However, it is clear that this quality should be measured.

For forecasts that get updated, the consistency in the updates is an important aspect of forecast quality. Unfortunately, though everyone knows forecast consistency or jumpiness when they see it, the use of objective measures of this quality in forecast verification is very limited.

A similar problem exists in economic forecasting, where inconsistent forecasts are referred to as rational or efficient. Rational or efficient forecasts are deemed to contain all information available at the time of issuance, a desirable quality. Any relationship of the forecasts through time is evidence of hedging, or holding back some information to include later, a form of “cheating”. In economics, new information, perhaps in the form of rate or policy changes, happen all at once. “Useful information on the terminal event is assumed to arrive in one lump sometime during the n periods before the terminal event.” (Nordhaus, 1987). For weather forecasts, new information may trickle in over time. If this is true, then consistency in weather forecasts may not be as undesirable as consistency in economic forecasts.

Two measures used by economists to determine rationality are tested on example weather forecasts. Tests of market efficiency include serial correlation tests and runs tests. It is typical in such tests to allow for a linear trend. The utility and sensitivity of these tests for evaluating the quality of updated weather forecasts are discussed.

Generally, consistency is a property of the forecasts only, though observations can be incorporated into the measures of consistency. The accuracy of a forecast is unrelated to its consistency. Thus, a measure of consistency should be considered an addition to accuracy measures.

2 DATA

Forecasts with decreasing lead times of a single terminal event (i.e. with equal valid time) were selected for four surface locations (Boston, Chicago, Denver, and Los Angeles). The North American Model (NAM) forecasts used have lead times out to 84 hours with updates each 6 hours. Thus, series of 14 forecasts are examined. Ideally, longer series would be available. However, it appears common for weather forecasts to have a short series of updates. Thus, a useful measure must be able to detect consistency in short series in at least some cases. Further, in order to be useful, a measure of consistency must work on a variety of forecast variables, whether they are symmetric, skewed, Gaussian, or some other distribution. Therefore, temperature, pressure, and wind speed forecasts are included in this analysis. Precipitation, since it is only conditionally continuous, is not included in this work, but will be examined in future analyses, as will other meteorological variables.

[*] Corresponding author address: Tressa L. Fowler, NCAR, P.O. Box 3000, Boulder, CO 80307. tressa@ucar.edu

For each series of forecasts, two types of series can be derived for analysis, series of errors and series of revisions. Revisions are commonly analyzed in economics, to analyze forecast changes while allowing for drift (i.e. a change in location) in the series. Error series are examined here as well, though forecast drift will show up as consistent behavior.

2.1 Forecast Revisions

For each forecast series, f_i, a forecast revision, R_i, is the change in the event forecast between two adjacent time steps, so it is the update.

$$R_i = f_{i+1} - f_i$$

By subtracting the earlier forecast from the later, increases in the forecast will have a positive sign and decreases in the forecast will have a negative sign.

2.2 Forecast Errors

For each forecast series, f_i, there is a single observation or realization, o. Thus, the error series is just a centered version of the forecast series. Forecast errors, e_i, are the differences between each forecast and the actual observation of the event.

$$e_i = f_i - o$$

Overforecasts will have a positive sign while underforecasts will have a negative sign.

3 METHODS

The autocorrelation and Wald Wolfowitz tests are used to measure the association of forecasts through time. Two tests are included because each has different types of sensitivity and robustness, similar to use of the mean and median. The autocorrelation uses continuous measures, so it is sensitive but not robust. The Wald Wolfowitz uses categorical information, making it robust to outliers but less sensitive.

Each test is shown with some example numerical weather forecast data. A thorough testing of these measures is not necessary since both have seen extensive use and documentation in both statistics and economics. The goal here is to demonstrate the potential utility of these measures for assessing the consistency of weather forecasts through time. The primary concerns for this application are short time series and weather variables with non-Gaussian distributions.

3.1 Autocorrelation

Autocorrelation measures the association (correlation) of values in a series to those that precede them in time (Box *et al*, 1994). Autocorrelation is generally calculated for several different 'lag' values, where the lag value is the number of time steps by which one value precedes the other. Lag one autocorrelation measures the association of each measure with that immediately preceding it, etc.

The autocorrelation is the same as the Pearson correlation, but using the lagged series. Thus, it is familiar to the weather forecasting community and simple to interpret. The distribution of the autocorrelation is known, allowing for simple determination of statistical significance (i.e. calculation of hypothesis tests and confidence intervals). However, the autocorrelation calculation is not robust. It is sensitive to outliers and lack of stationarity (a change in location and/or variability) in the time series.

Autocorrelation of errors may tell us that our forecast is non-stationary. Autocorrelation of revisions can tell us if the forecast is stepping toward some new forecast value or zigzagging. This is not a measure of convergence, as both series may converge.

3.2 Wald Wolfowitz (Runs) Test

The Wald Wolfowitz test (1943) tests for the random distribution of 'runs', or series of the same value, of two discrete categories. As an example, in this series of positive and negative values, +++++----++, there are three runs. For this analysis, the two categories are positive or negative. When analyzing the revisions, the positive and negative values indicate the direction of change of the forecast. For the errors, the bias must first be removed from the series. Then, the positive and negative are the direction of the bias-adjusted errors. Without the bias adjustment, several series are completely positive or negative, and the Wald Wolfowitz statistic is undefined. Thus, the test cannot be run unless the series to be analyzed has at least two runs.

We can calculate the expected number of runs if the two categories are arranged with respect to time at random. The two categories need not have equal probability. Then, a one sided test for too few runs will conclude if the series has fewer changes between negative and positive than would be expected from a random distribution of changes. A series with more changes than a random series is not consistent through time, so there is no need to have a two sided test.

The runs test is very robust to outliers and to lack of stationarity in the time series, because the data are comprised only of two categories. However, a threshold for dividing the series into positive and negative values must be chosen. When series values lie very close to this threshold value, the test can be quite sensitive to the choice of threshold.

Too few runs in the error series tell us that forecast mistakes are consistent. Too few runs in the revision series tell us that the forecast changes are consistent through time.

4 RESULTS

4.1 Forecast error assessment

The bias corrected forecast errors display much consistency through time. An example showing the pressure error in updating forecasts for the four cities is shown in Figure 1. The forecast biases for each city are shown by the dashed lines. Both Denver and LA have 'flat' error series, while Boston and Chicago have smaller errors at longer lead times (the opposite of what one would expect). The Boston and Chicago series show "drift", or a change of location over time. The autocorrelation calculation detects the drift, and thus both Boston and Chicago have statistically significant values of the autocorrelation, *r*. Meanwhile, the autocorrelation values for Denver and LA are not significantly different from 0.

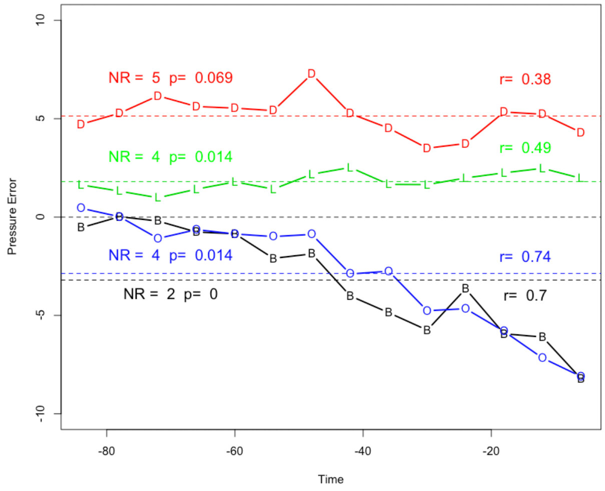

Figure 1: Error series from updating pressure forecasts for Boston (B, black), Chicago (O, blue), Denver (D, red), and Los Angeles (L, green) by lead time.

The runs test tells a slightly different story. The number of runs (NR) for each of the four cities is shown on the plot, along with the probability (*p*) of seeing that many or fewer runs in the series given a random distribution of errors above and below the bias. LA and Chicago, with 4 runs each, and Boston with 2 runs, all have so few runs that the probability that the errors are randomly distributed about the mean error (bias) is less than 5% for each. For Denver, with 5 runs, this probability is about 7%, not statistically significant when a customary error level of 5% is used.

The Chicago error series demonstrates the sensitivity of the runs test to the threshold. Though the Chicago and Boston series are very similar looking, Boston has 2 runs while Chicago has 4. For Chicago, two values near the center of the series fall very near the bias, one on each side. This causes additional changes between positive and negative errors, resulting in 4 runs. A slight change in that bias value would result in only 2 runs.

4.2 Forecast revision assessment

Figure 2 shows an example of wind speed forecast revisions for the four cities. For all cities, the number of runs exceeds the expected number based on random fluctuation. Thus, none of the series is consistent through time with respect to the Wald Wolfowitz test. Similarly, all revision series lack positive first order autocorrelation, indicating a lack of consistency through time. Los Angeles has a statistically significant negative autocorrelation, indicating more "zigzagging" than random. This statistic confirms a similar conclusion that might be drawn by visual inspection.

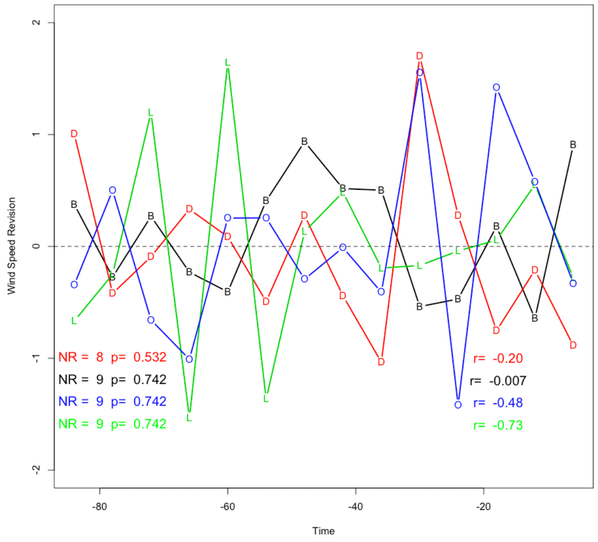

Figure 2: Series of forecast revisions for Boston (B), Chicago (O), Denver (D), and Los Angeles (L) by lead time.

The revision series for all other weather variables and cities (not shown) had no significant association through time, as measured by either the Wald Wolfowitz test or by the autocorrelation. Thus, the revisions in those series could be considered to be noise. In particular, the series of pressure forecast revisions for Los Angeles and Chicago shows no association through time. The error series (and thus the forecast series) for those cities have drift, but the revision series does not. This demonstrates that tests on the revision series ignore drift while tests on the error series detect drift as association in the series.

5 CONCLUSIONS AND FUTURE WORK

Both the autocorrelation and the runs tests can measure association of forecasts through time, in complementary ways. Both are simple to calculate and understand, thoroughly documented, and have known distributions (useful for determining significance of results). They can be used on any forecast series of a continuous variable. The two tests have different sensitivities and robustness, so users should consider which makes the most sense for each application.

The runs (Wald Wolfowitz) test is robust to outliers and changes in variability. Use of this test on error series may require bias removal first. Further, since it is discrete, the runs test is sensitive to small changes near the “transition” line.

Autocorrelation is the most common method of examining association of measurements through time. It is insensitive to bias, but sensitive to changes in location or variability of the series.

A considerable amount of future work remains. Assessment of other types of forecasts, such as hurricane intensity and location, should be tested. Precipitation is typically only a conditionally continuous variable, and thus is unlikely to work well with either of these measures. It may be desirable to include several series of forecasts in a single test rather than examining only individual forecast series, so some other test statistics may be required.

ACKNOWLEDGEMENTS

NCAR is sponsored by the National Science Foundation.

REFERENCES

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel, 1994: *Time Series Analysis: Forecasting and Control*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Clements, M. P., 1997: Evaluating the Rationality of Fixed-event Forecasts. *Journal of Forecasting*, **16**, pp. 225-239.

Ehret, U., 2010: Convergence Index: a new performance measure for the temporal stability of operational rainfall forecasts. *Meteorologische Zeitschrift* **19**, pp. 441-451.

McLay, J., 2010: Diagnosing the relative impact of "sneaks", "phantoms", and volatility in sequences of lagged ensemble probability forecasts with a simple dynamic decision model. *Mon. Wea. Rev.* doi: 10.1175/2010MWR3449.

Nordhaus, W. D., 1987: Forecasting Efficiency: Concepts and Applications. *The Review of Economics and Statistics*, **69**, pp. 667-674

Ruth, D. P., B. Glahn, V. Dagostaro, and K. Gilbert, 2009: The Performance of MOS in the Digital Age. *Weather and Forecasting*, **24**, 504-519.

Wald, A. and J. Wolfowitz, 1943: An exact test for randomness in the non-parametric case based on serial correlation. *Ann. Math. Statist.*, **14**, 378–388.

Zsoter, E., Buizza, R., & Richardson, D., 2009: 'Jumpiness' of the ECMWF and UK Met Office EPS control and ensemble-mean forecasts'. *Mon. Wea. Rev.*, **137**, 3823-3836.