Measuring the quality of updating high resolution time-lagged ensemble probability forecasts using spatial verification techniques.

Tressa L. Fowler, Tara Jensen, John Halley Gotway, Randy Bullock

1. Introduction

Numerical models are producing updated forecasts more frequently as computing resources increase. Some updates are plagued by large changes from one time to the next, causing users to lose confidence in the forecasts. For forecasts that are revised, especially those with frequent updates, the magnitude and randomness of the revision series is an important aspect of forecast quality. Similar problems exist in economics and other fields, and many types of metrics are in place for simple updating time series. Unfortunately, though everyone knows forecast jumpiness when they see it, it is rarely measured objectively in weather forecasting (Roebber, 1990; Lashley *et al*, 2008). Users may examine dprog/dt and even calculate the total difference in the forecast from one time to the next. However, this measure suffers from the same double penalty issues as traditional verification measures, namely that a small displacement may be measured as a large change at multiple locations. McLay (2010) and Zsoter *et al* (2009) apply consistency metrics to ensemble forecasts. In this presentation, assessments of forecast revision magnitude and randomness are applied to attributes of forecast objects using spatial verification techniques, thus incorporating temporal and spatial information into the assessment. Examples of revision assessments from probability forecasts from a high resolution time-lagged ensemble that is updated hourly are presented.

Users of forecasts are often interested in the consistency of updating forecasts for a terminal event time. In this instance, users are evaluating an experimental time-lagged ensemble that produces a hazard probability forecast, updated hourly. They desire to have objective measures of forecast consistency to either confirm or refute the perception that time-lagged ensembles provide greater consistency than do traditional ensemble forecasts.

Because the focus of the hazard probability forecasts is precipitation, all of the usual issues with traditional verification arise. Basic measures of consistency are typically computed for fixed locations, and just like root mean square error (RMSE) and probability of detection (POD), these can be affected drastically by displacements, etc.

Both the magnitude and randomness of the revision series impact forecast consistency. Small, random revisions are typically of little interest. Large, consistent updates usually point to a forecast reacting appropriately to recent information. Thus, the goal is to identify forecasts with both large and random revisions. For this work, assessments of forecast revision magnitude and randomness are applied to attributes of forecast objects using a spatial verification technique, thus incorporating both temporal and spatial information into the assessment. The metrics have all been previously proposed and used, the novelty here is their application to revisions series of object attributes.

2. Methods

The suite of metrics presented here

measure forecast consistency based on object attributes. This effort does not define a new metric, strictly speaking. Rather, it provides new information by combining existing methods and metrics in a novel way. Characteristics of probability forecasts are determined by features-based spatial verification software, MODE-TD (Method for Object-based Diagnostic Evaluation – Time Domain), that tracks objects through their series of updates. Then the revision in various attributes from each time compared with the previous time are calculated. These revisions are assessed via existing statistical tests for randomness and magnitude. In this way, the changes in a feature can be evaluated through time without regard to specific locations. This small suite of metrics provides more detailed information to users than a single metric or index. Further, they allow some choices to the user in determining the most appropriate way to measure magnitude and randomness for their data.

The MODE-TD software identifies and tracks objects in a succession of forecasts. Predominantly, it has been used on individual model runs with increasing lead time, thus tracing the evolution of forecast systems as time goes on. However, it functions equally well on a series of forecasts with the same valid time from different model runs and thus, with decreasing lead times. Hereafter, we refer to these forecasts as 'updating' forecasts. The change in the forecast attribute from each time to the prior is the 'revision'. Thus, increases in the forecast are positive revisions and decreases in the forecast are negative revisions. The examples here focus on consistency in areal coverage of the probability of snowfall. Use of other object attributes is possible, depending on the forecast type.

A consistency measure can certainly be applied to time-series from individual grid points. However, many of the same issues apply as when standard verification statistics are used in this manner. If a feature moves, but remains otherwise similar, it may be considered relatively consistent. However, a time series from an individual grid point may show that the feature disappears completely, which is definitely not consistent.

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 the 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. These tests have been demonstrated previously on updating forecast data (Fowler, 2010). Additionally, they were extended to hurricane track and intensity forecasts (Fowler *et al.*, 2015).

The **autocorrelation** is the same as the Pearson correlation, but using the revision 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 (Clements, 1997). It is sensitive to outliers and lack of stationarity (a change in location and/or variability) in the time series. 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.

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 analysing the

revisions, the positive and negative values indicate the direction of change of the forecast. The test cannot be applied unless the series 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 series of positives and negatives. 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 revision series tell us that the forecast changes are consistent through time.

3. Examples

Examples of revision assessments from probability forecasts from a high resolution time-lagged ensemble that is updated hourly are discussed. In particular, it is desirable to evaluate changes in POP forecasts without restricting them to specific locations. The experimental model HRRR-TLE (High Resolution Rapid Refresh – Time Lagged Ensemble) produces hourly updates of probability of snowfall forecasts over the CONUS. These examples use probability for snowfall rates > 0.5". All forecasts are valid at January 23, 2016 0Z, with lead times beginning 13 hours ahead and decreasing hourly. The MODE-TD software was used to track four forecast objects through a series of updates. Additionally, the total domain forecast is included here as object 5. Attributes of two-dimensional probability objects were derived, though MODE-TD can also calculate three-dimensional attributes. By taking differences at each time step, a revision series is derived. For probabilities, object area may be an interesting attribute.

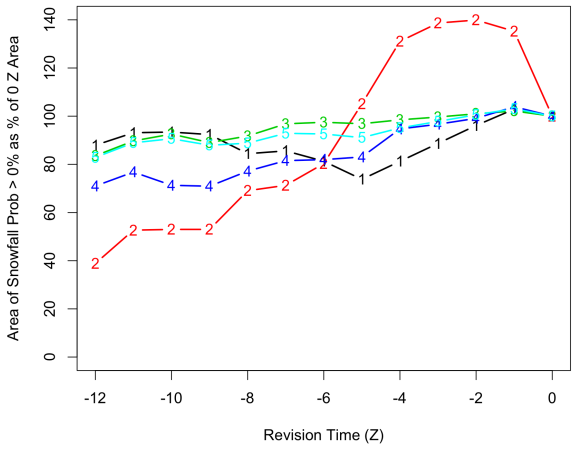

Figure 1: Line plot showing area of each object as a percent of final object area for each individual object plus the entire domain (object 5).

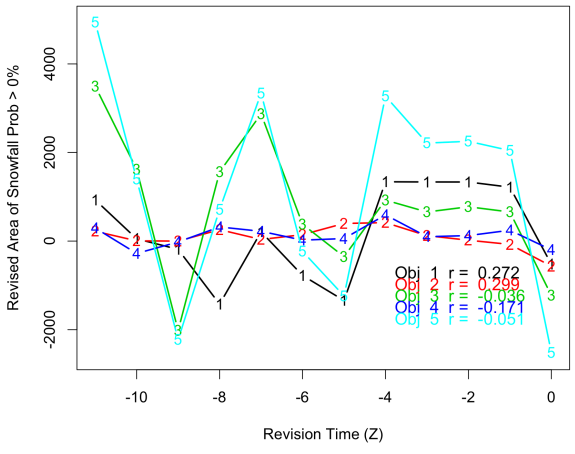

Figure 2: Line plot showing area revisions for each individual object plus the entire domain. The autocorrelation value of each revisions series is shown in the legend. None are statistically significant, possibly due to the sample size typical of a single case study.

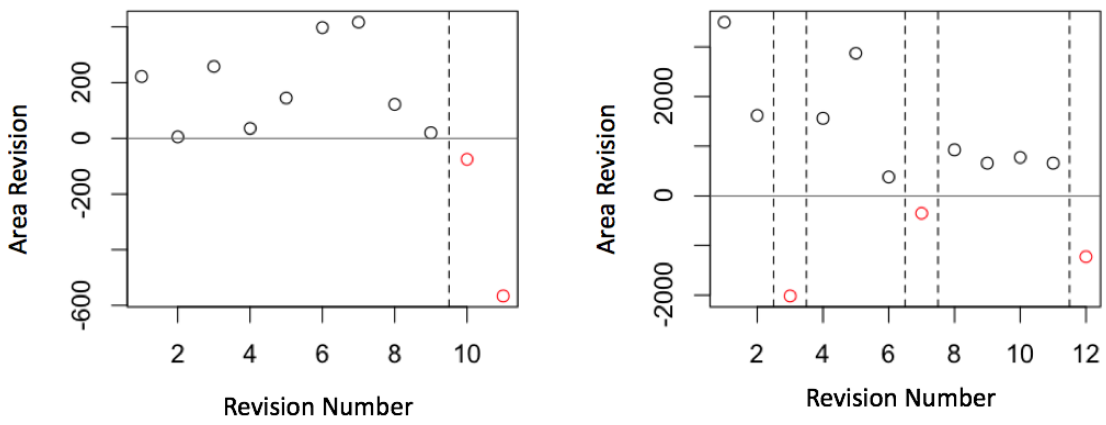

Figure 3: Runs test graphic showing increases and decreases in forecast area revision series for objects 2 (left) and object 3 (right). Dashed lines delineate each run. Object 2 has only 2 runs and a p-value = 0.036, indicating a trend in the revisions. Object 3 has 6 runs with a p-value = 0.75, indicating randomness.

Figure 1 shows the area of each object with decreasing lead time as a percent of the object area at the valid time. Object 2 stands out as having a distinct increasing trend. Figure 2 shows the revisions in the area of each object from as the forecast updated. Visually, object 3 stands out as having some large, oscillating revisions. The autocorrelation and Wald Wolfowitz tests may not give the same answer. In particular, for object 2, the autocorrelation (0.299) is not statistically significant (probably due to small sample size, though it is larger than the values for the other objects), while the runs test (Figure 3) shows a significant trend in the revisions (p-value 0.036). (This is distinct from a trend in the forecast, as overall a forecast area can increase while the revisions oscillate.) Meanwhile, object 3 shows no significant autocorrelation (-0.036) and the runs test agrees (6 runs, p-value 0.75).

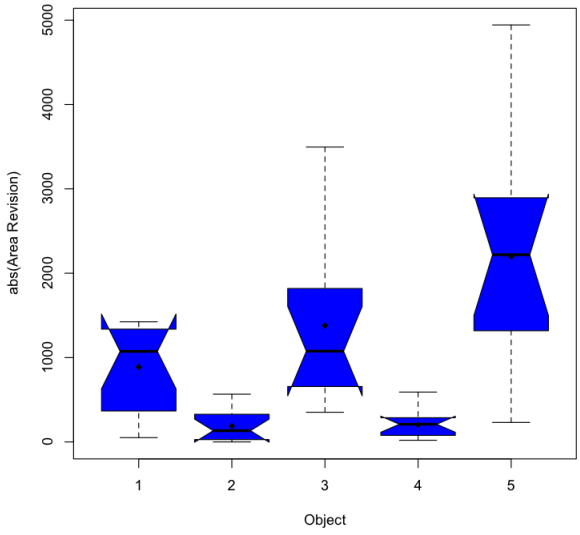

Figure 4: Boxplots showing absolute area revision for each object (1-4) plus the overall domain (5). The mean absolute revision is indicated by the filled diamond in each box, while the median is the center of the 'waist' of each box.

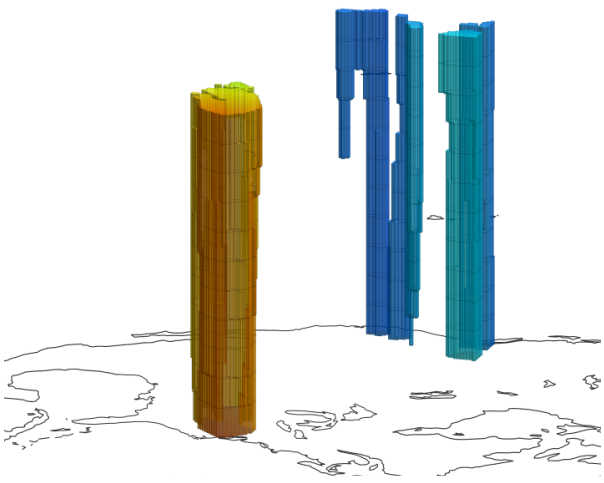

Figure 5: Graphic showing MODE-TD tracked area for forecast objects (probability of ½" snowfall > 0). An object that is completely consistent across updates would appear with straight sides. Objects 'grow up' with decreasing lead time. In dark blue, two object are shown merging as the valid time approaches (top of the graphic).

Figure 4 shows boxplots of the magnitude of the revisions for each object (1-4) plus the overall domain (object 5). Object 2 and 4 show much smaller revisions than object 1, 3, and 5. Users may also assess magnitude of the revisions via the mean absolute revision or the root-mean square of the revision. Object 2 showed a trend in the forecast area revisions and the magnitude of the change is relatively small. The other objects, especially object 3, and the overall domain have larger, random revisions, indicating a lack of consistency in these updates. Of course, ideally these metrics should be applied to a large number of cases, and the magnitude should be judged via comparison to some reference or competing forecast.

Figure 5 shows a MODE-TD object graphic. Two-dimensional object areas are stacked vertically, with the longest lead times at the bottom and shortest at the top. An object with perfect consistency in the updates would have 'smooth' sides. The movements and growth of some objects are easily seen, as is the merging of two objects near the US west coast as the valid time approaches.

4. Conclusions

The examples here are from a single case study, for ease of interpretation. However, these metrics are easily extended to larger sets of cases. In this example, we grouped the four individual objects into object 5, covering the whole domain. This accumulation can also be accomplished across a set of forecasts to determine overall consistency and guide the forecaster on interpretation of forecast revisions. Additionally, forecasters are encouraged to provide consistent forecasts to the public unless large changes are warranted. Therefore, guidance with optimal consistency measures are more likely to be adopted by the forecast community.

There are several metrics related to forecast consistency. Dprog/dt is often examined, but in many cases the determination is subjective (Hamill, 2003). The Ruth-Glahn Forecast Convergence Score (2009; Pappenberger *et al*, 2011) and Griffith's Flip Flop Index combine a magnitude threshold with a ratio of flip-flops at fixed locations to determine consistency. Ehret's (2010) convergence index also uses a threshold, but weights short lead times more than far lead times. Further, the concern is primarily with convergence (decreasing error) rather than random oscillation. All of these measures are typically used at a fixed location, at either a station or on a grid. They also operate on the series of forecasts directly, rather than the series of revisions.

The statistics used here are all standard measures that are well documented in the statistics literature. Further, the software packages required for this analysis are supported open source and well documented. MODE-TD has been released with the Model Evaluation Tools (MET) software, and the statistical tests are available in R. These methods allow for a trend in the forecasts but detect trends in the revisions, which threshold-based "flip-flop" ratios do not handle automatically (though trends can be manually removed prior to calculation of the index). Used together, separate tests for magnitude and randomness of the revisions provide more detailed information to users than those metrics that attempt to combine the two, thus masking the contribution of each.

5. Future Work

These metrics were demonstrated at the National Weather Service / Weather Prediction Center (NWS/WPC) to assess their utility. The forecasters and product development team thought the metric would be especially useful if a user-defined threshold is applied to the runs to identify changes that would impact the forecaster's decisions. The suite of metrics will be included in the WPC

Winter Weather Experiment in January-February 2017. Additional object attributes will be incorporated in future tests and demonstrations.

Acknowledgements

NCAR is sponsored by the National Science Foundation.

References

Clements, M. P., 1997: Evaluating the rationality of fixed-event forecasts, *J. Forecast*., **16**, pp. 225–239.

Clements, M. P. and Taylor, N., 2001: Robustness of fixed-event forecast rationality, *J. Forecast*., **20**(4), pp. 285–295.

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

Fowler, T. L., 2010: Is Change Good? Measuring the quality of updating forecasts. *20th Conference on Numerical Weather Prediction*. American Meteorological Society, Boston.

Fowler, T.L., B. G. Brown, J. Halley Gotway and P. Kucera, 2015: Spare Change : Evaluating Revised Forecasts. Mausam, **66**(3) (July 2015), pp. 635-644.

Hamill, T. M., 2003: Evaluating forecasters' rules of thumb: a study of D(Prog)/Dt. *Weather Forecast*., **18**, pp. 933–937.

Lashley, S. L., Fisher, L., Simpson, B. J. , Taylor, J., Weisser, S., Logsdon, J. A., and Lammers, A. M., 2008: Observing verification trends and applying a methodology to probabilistic precipitation forecasts at a National Weather Service Forecast Office. Preprints, 19th Conf. on Probability and Statistics, New Orleans, LA, *Am. Meteorol. Soc*., 9.4. available at : http://ams.confex.com/ams/pdfpapers/134204.pdf

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.

Pappenberger, F., Bogner, K., Wetterhall, F., He, Y., Cloke, H. L., and Thielen, J., 2011: Forecast convergence score: a forecaster's approach to analysing hydro meteorological forecast systems, *Adv. Geosci*., **29**, pp. 27–32, doi:10.5194/adgeo-29-27-2011.

Roebber, P. J., 1990: Variability in successive operational model forecasts of maritime cyclogenesis, *Weather Forecast*., **5**, 586–595.

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., and Richardson, D., 2009: "Jumpiness" of ECMWF and Met Office EPS Control and Ensemble-Mean Forecast. *Mon. Weather Rev*., **137,** pp. 3823–3826.