MAUSAM, **66**, 3 (July 2015), 635-644

551.509.5

Spare Change : Evaluating revised forecasts

TRESSA L. FOWLER, BARBARA G. BROWN, JOHN HALLEY GOTWAY and PAUL KUCERA

National Center for Atmospheric Research/Research Applications Laboratory Boulder, CO USA

e mail : tressa@ucar.edu

सार – प्रचंड चक्रवात अथवा तेज हवा के दिनों जैसी मौसम घटनाओं के संबंध में प्रायः समय पर संशोधित पूर्वानुमान के सुसंगत होने का प्रश्न उठता रहता है। किसी एक घटना के लिए पूर्वानुमान तैयार किया जाता है और जैसे-जैसे वो घटना घटने वाली होती है उसमें संशोधन किया जाता है। पूर्वानुमान में संशोधन से नवीन और बेहतर सूचना प्राप्त होती है जिससे बेहतर पूर्वानुमान दिया जा सकता है। फिर भी, यदि पूर्वानुमान बार-बार बदलता है और बड़ी मात्रा में बदलता है, उपयोगकर्ता मान सकता है कि वे खराब है या अनिश्चित है। समय के साथ उपयोगकर्ता पूर्वानुमान पर विश्वास खो सकता है जो सुसंगत नहीं है। यह विशेष रूप से निर्णयकर्ताओं के लिए एक मुद्दा है जो पूर्वानुमान (उदाहरणतः आपातकालीन प्रबंधक) के आधार पर योजनाएँ बनाते हैं और नए पूर्वानुमान मिलने पर बार बार अपनी योजनाएँ बदलते हैं।

अतः संशोधित पूर्वानुमानों के संशोधन में सुसंगतता पूर्वानुमान की गुणवत्ता का महत्वपूर्ण पहलू है। दुर्भाग्यवश, यद्यपि प्रत्येक व्यक्ति पूर्वानुमान की सुसंगतता को जानता है जब वह उसे देखता है, तथापि पूर्वानुमान के सत्यापन में सुसंगतता का मूल्यांकन करने के लिए उद्देश्यात्मक उपायों का उपयोग बहुत सीमित है। आर्थिक पूर्वानुमान में भी इसी प्रकार की समस्या रहती है, जहाँ सुसंगतता को मापने के लिए एकल समय श्रृंखला पर कुछ साधारण परीक्षण किए गए हैं। तथापि, ये उपाय मौसम पूर्वानुमान पर आसानी से लागू नहीं होते हैं जो बहुविमीय अथवा बहुत सी समय श्रृखलाओं का संग्रहण हो सकता है। उदाहरण के लिए, प्रचंड चक्रवात के मार्ग में पीछे और आगे (अथवा विंडशील्ड वाइपर) के पूर्वानुमान के प्रभाव को मापने का कोई सामान्य तरीका नहीं है। इस शोध पत्र में, पूर्वानुमान संशोधन समय श्रृंखला के कुछ सुसंगत उपायों पर चर्चा की गई है। प्रचंड चक्रवात पूर्वानुमान मार्ग और बहुसंख्यक समय श्रृंखला के संचयन से लिए गए कुछ प्रारंभिक उदाहरणों का उपयोग करके उपायों को अधिक जटिल पूर्वानुमानों पर लागू करके जाँच की गई। विभिन्न पूर्वानुमान में सुसंगतता के उपयों के मध्य तुलना पर विशेष ध्यान दिया गया है।

ABSTRACT. The question of consistency of revised 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 then revised as the time of the event nears. Hopefully, the revision reflects new and better information that will yield a better forecast. Nonetheless, if forecasts change frequently or by large amounts, a user may believe they are poor or uncertain. Over time, a user may lose trust in forecasts that are not consistent. This is particularly an issue for decision makers who create plans based on early forecasts (*e.g*., emergency managers), then must change their plans repeatedly as new forecasts arrive.

Thus, for forecasts that are revised, the consistency in the revisions is an important aspect of forecast quality. Unfortunately, though everyone knows forecast consistency when they see it, the use of objective measures to evaluate consistency in forecast verification is very limited. A similar problem exists in economic forecasting, where some simple tests are applied to a single time series to measure the consistency. However, these measures do not easily extend to weather forecasts that may be multi-dimensional or a collection of many time series. For example, there is no simple way to measure the back and forth (or 'windshield wiper') effect of changing hurricane track forecasts. In this paper, some consistency measures of forecast revision time series are discussed. Extensions of these measures to more complex forecasts are examined using some preliminary examples from hurricane forecast tracks and accumulations of multiple time series. Particular attention is paid to comparisons of consistency measures between competing forecasts.

Key words – Wald-wolfowitz test, TC forecast, Surface station forecast, Revised forecasts.

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 find this quality desirable while others do not. Historically, the consistency of weather forecasts through time has not frequently been

considered or measured. Recently, several authors have constructed consistency measures for specific forecast types, including ensembles (Zsoter *et al*., 2009), precipitation (Ehret, 2010 and Lashley *et al*., 2008), operational forecasts (Ruth *et al*., 2009) and Markov chains (McLay, 2011). 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 consistency 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, 2011).

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.

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 here 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. However, these measures do not easily extend to weather forecasts that may be multi-dimensional or a collection of many time series. There is no simple way to measure the 'windshield wiper' effect of changing hurricane track forecasts. In this paper, some consistency assessments of forecast revision time series are discussed. Extensions of these measures to more complex forecasts are proposed and tested on examples from hurricane forecast tracks and accumulations of multiple time series. Particular attention is paid to comparisons of consistency measures between competing forecasts.

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

Though many types of forecasts are incorporated into this evaluation, the quantity used to evaluate them is the forecast revision. The revision is the change in the forecast for a certain (valid) time that has occurred due to a forecast update. A simple example is maximum temperature for Friday in Denver. On Monday, the four-day forecast may be for a high on Friday of 16 °C while on Tuesday the three-day forecast for Friday's high temperature may be for 22 °C. In this example, the forecast revision is 6 °C. Thus, for each forecast series, fi, a forecast revision, Ri, is the change in the event forecast between two adjacent time steps, so it is the magnitude of 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. 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 in the economic literature as well, though forecast drift will show up there as consistent behavior. For this evaluation, errors are not examined, but they may be included in future work.

Revisions are a quality of the forecast only and are independent of the observations. Thus, they are not a measure of the forecast error. Revision assessments can provide users with information about the forecasts, particularly in comparison to some reference forecast. However, they are not a measure of goodness of a forecast. They may be able to provide users with some measure of the uncertainty in the forecast.

(a). *Surface station forecasts*

Forecasts with decreasing lead times of a single terminal event (*i.e.*, with equal valid time) were selected for four surface station locations (Boston, Chicago, Denver, and Los Angeles). The North American Mesoscale (NAM) model predictions used here 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 is 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.

(b). *Tropical cyclone forecasts*

Tropical cyclones (TC) present a unique evaluation challenge requiring very different metrics than traditional forecast verification. The number of forecasts is relatively small and varies for different valid times over the lifetime of the cyclone. Though many quantities are forecast (*e.g.*, mean sea level pressure, radius of maximum winds) the primary evaluations are carried out on the track and intensity (*i.e.*, maximum wind speed) (Sampson and Schrader, 2000). Thus, the evaluations in this paper will focus only on these two quantities. The official TC forecasts are used for comparison to numerical weather prediction (NWP) forecasts. These official forecasts are issued by the National Hurricane Center for the Atlantic basin. They are human generated and based on a variety of information sources (Rappaport *et al.*, 2009).

The collection of forecasts used in the TC intensity examples is for a single tropical cyclone (Gabrielle). A set of revision series from an NWP model provides forecasts for 11 different valid times with one to six revisions for those times. The official forecasts for the same storm are used for comparison. However, only seven valid times with one to five revisions are available. Though using matched samples is often recommended, the examples here demonstrate the utility of consistency measures on mismatched samples, which are quite common.

The TC track forecasts from several NWP models are compared with the operational forecasts for the entire 2013 season in the Atlantic basin. Tropical cyclone Gabrielle is also singled out for some examples. Different numbers of valid times and revisions are available from each forecasting system and no effort was made to unify the set of cases across forecasts for this preliminary assessment. For Gabrielle, the total number of forecasts issued varied from 23 to 26, with varying combinations of lead and valid times. For the entire 2013 season, the total number of forecasts varied from about 1300 to 2500, covering 39 tropical cyclones. Not all forecasting systems produced forecasts for every tropical cyclone.

Although observations are not used to determine revision series, for track forecasts the actual TC best track is used as a reference. The best track is the best estimate of the tropical storm location based on all available data sources, including those that become available following the valid time of the storm (Torn and Snyder, 2012).

3. Methodology

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, analogous to use of both the mean and median to characterize average behavior. 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.

Additionally, several summary statistics are examined for TC forecast revision series with the primary goal of comparison rather than determination of randomness. These include measures of magnitude, spread, and sign. Together, the measures give the user a sense of whether one forecast has larger revisions and / or more random revisions than some standard.

Each test is shown using some example NWP forecast data. The goal 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, sets of time series, and weather variables with (possibly multi-dimensional) non-Gaussian distributions.

(a). *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. 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 revisions can tell us if the forecast is stepping consistently toward some new forecast value or zigzagging.

(b). *Wald-Wolfowitz 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. If the series is completely positive or negative, then the Wald-Wolfowitz statistic is undefined. Thus, the test cannot be run unless the series to be analyzed has at least two runs.

The expected number of runs can be calculated if the two categories are arranged randomly with respect to time. 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 revision series indicate that the forecast changes are consistent through time.

(c). *Summary statistics for revision series*

Typical weather forecast data require a different sort of assessment than the simple autocorrelation or runs tests. Some forecasts are multi-dimensional and nearly all have several sets of shorter time series (*e.g.*, short series of forecasts for many valid times) that must be evaluated rather than a single long series. Missing data may also be encountered. In these cases, the assumptions of the autocorrelation and runs tests are not met so other evaluation methods are necessary. By examining simple statistics of the set of revisions, an assessment of consistency and magnitude can still be made, particularly via comparisons of forecasts for the same event. Statistical inference based on these measures will be more difficult, but future work will investigate the possibility of confidence assessment via bootstrapping.

The focus also changes somewhat for evaluation of weather forecasts. The goal of the autocorrelation and Wald-Wolfowitz tests is to identify randomness. For weather forecasts, it is also of interest to examine the magnitude of the revisions, usually with respect to some standard. So in this case, consistent behavior is not only a lack of randomness but also a set of smaller revisions. The lack of a long time series makes identifying randomness more difficult than in the textbook case, but determining whether the typical magnitude of the revisions is large or small is quite straightforward. In fact, many users may be less interested in the random *vs* consistent nature of a forecast so long as the revisions are small.

The TC track and intensity revisions are used to demonstrate the simple summary statistics. The intensity revisions (*e.g.*, wind speed forecast changes) are evaluated using averages, quartiles, frequencies, and distributional plots. These statistics can be compared to a reference forecast using paired or two sample tests of significance (Lanzante, 2005; Snedecor and Cochran, 1989).

The "windshield wiper" effect of the track revisions is of particular interest to users (Elsberry and Dobos, 1990). Forecasts often have a back and forth track adjustment that is of concern to forecast users. This quality can be examined in several ways. Elsberry and Dobos evaluated successive cross track errors (the error component perpendicular to the storm best track). Here, the average or total path length of the revision track measures the magnitude of the revisions over the storm lifetime and can be compared to operational or other reference forecasts. The area inside the convex hull of the revision track is another measure of magnitude. The analog to the Wald-Wolfowitz test is to count the forecast crossovers of a reference track, fewer crossovers than randomly expected indicate consistent behavior. The hurricane best track is the simplest such reference track, but if the forecast is biased to one side then the number of crossovers will be too low. Use of the average forecast track unbiases the number of crossovers. This test is essentially the Wald-Wolfowitz, but applied to a set of revision series with discontinuities rather than a single series.

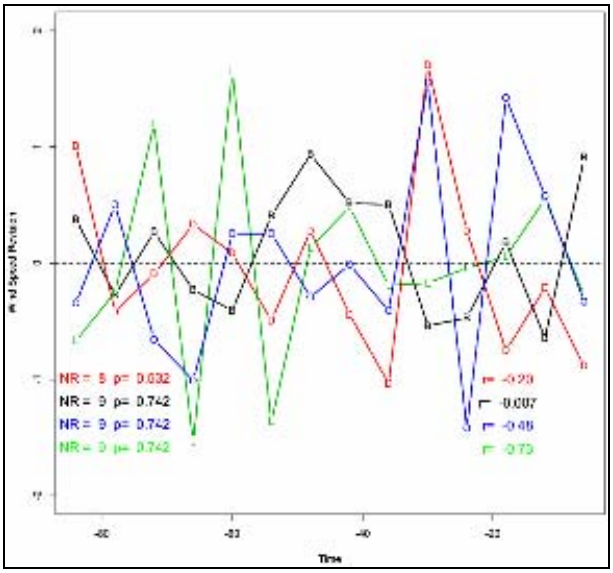


Fig. 1. Revision series for surface wind speed forecasts in Boston (B in black), Denver (D in red), Chicago (O in blue), and Los Angeles (L in green). Autocorrelation values (r) and number of runs (NR) along with associated p-values (p) are noted on the figure. Time is listed in hours prior to the event

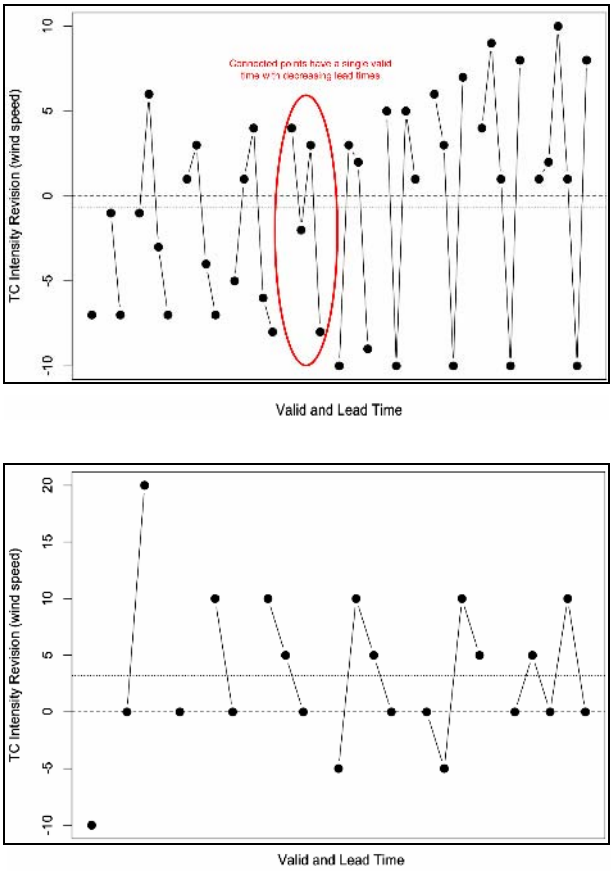


Fig. 2. Sets of revision series from forecasts of tropical cyclone intensity. Connected points share a valid time and have decreasing lead times. The latest valid times are to the far right of the figure, and the longest lead times begin each series of points. For tropical cyclones, more lead times are generally available for later valid times. The dotted lines show the bias (*e.g.*, average revision) for each series (-0.63 and 3.2, respectively)

The assessment of crossovers is similar to the categorization of the successive cross track errors undertaken by Elsberry and Dobos, 1990. The result for this work is a single measure of randomness with an associated statistical significance value. Further, the reference track may be either the best track or the unbiased average forecast track. In contrast, the EAD analysis provides tables of conditional proportions for each combination of left, right and center with reference to the best track. For both measures, track forecast biases may appear to be consistency since the forecast storm location may zigzag without ever crossing the best track.

4. Results

(a). *Forecast revision assessment on simple time series*

Fig. 1 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. 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.

(b). *Forecast revision assessment on TC intensity forecasts*

Fig. 2 shows sets of revision series by increasing valid time for a model forecast (panel a) and the official

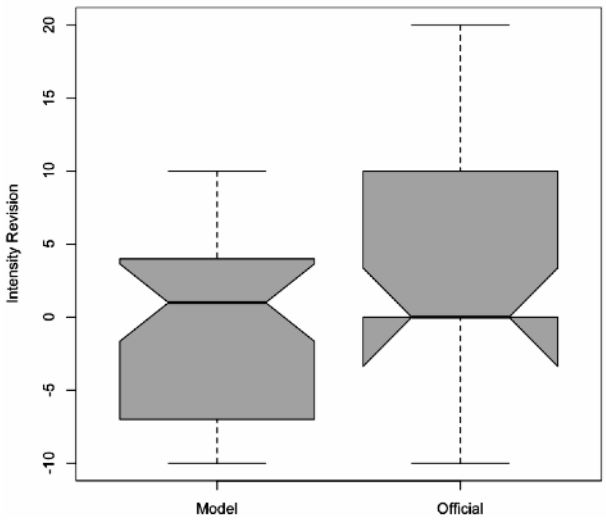


Fig. 3. Boxplots showing distributions of intensity revision values for tropical cyclone forecasts for an NWP model and the official forecasts

forecasts (panel b). Connected points share the same valid time, so these represent several revision series on one plot. Note that the number of revisions increases as the valid times increase, as is typical with TC forecasts. Statistical tests will not work well on these revision series individually, as most have too few points. However, when these series are all strung them all together, with 'missing values' between the series, it is possible to examine if the revisions are random or not. This violates some assumptions about statistical time series, but if we are comparing different forecasts to each other, it should still be possible to get a relative sense of how consistent the revisions are. In this example, there are 32 line segments (*i.e.*, connected points) and 18 of them cross the 0 line which has a probability of 0.84 under a null hypothesis of random (*i.e.*, inconsistent) behaviour. Thus, this set of revisions lacks structure. The bias line is shown here (-0.63) and the test can be done with bias removal by looking at how many lines intersect this line rather than 0, but in this case it makes no difference. This is in contrast to the official forecasts, which have only two positive to negative transitions (p-value = 0.02). So, the official forecasts fairly consistently update by increasing the forecast wind speed. By unbiasing the forecast (*e.g.*, counting the transitions across the 3.2 line), the consistent behavior disappears and the revisions are random (p-value = 0.34).

The autocorrelation of the forecast entire series, with breaks treated as missing values and removed, is -0.26 and is not statistically significant. Similarly for the official forecasts, the autocorrelation value is -0.3. Since the autocorrelation values are not significant, the revisions are random.

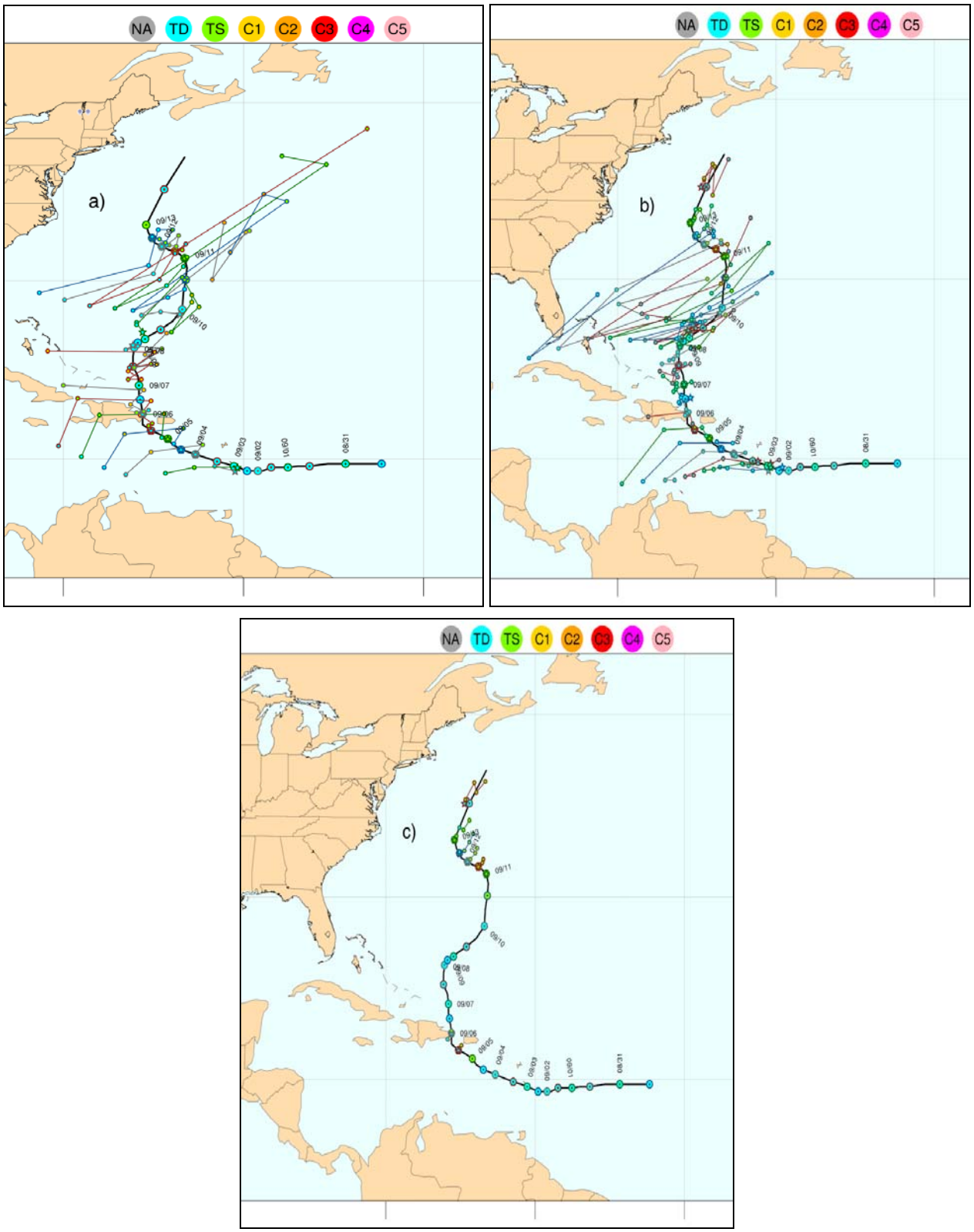
In Fig. 3, the magnitude of the entire distribution of revisions is examined *via* boxplots. Larger revisions, paired with randomness, indicate a high level of uncertainty in the forecasts. In this figure, the human versus computer-generated nature of the forecasts become evident. The official forecasts are human generated, and very often no change is made from one time to the next. However, when changes are made, they can be large (*e.g.*, maximum revision of 20). Further, revisions are in even increments of 5. The computer generated forecast revisions are slightly biased and small revisions are seen at most time steps. Few large revisions are noted and the magnitude is continuous, not necessarily in any specific increment.

The average magnitudes (absolute values) of the revisions for the two examples are quite similar (5.2 *vs* 5) and a two sample t-test shows no statistically significant difference (p-value = 0.88). Further, the median values of the official revisions are similar to those from the numerical weather forecast (0 and 1, respectively). Since both forecasts are also random in nature, the user may conclude that the uncertainty associated with the official forecasts is similar to the uncertainty associated with the NWP forecasts. However, the official forecast is more likely to increase the forecast wind speed with an update while the NWP model is about equally likely to increase or decrease the forecast wind speed (*e.g.*, the official forecasts have a revision bias).

(c). *Forecast revision assessment on TC track forecasts*

Hurricane track revisions are two dimensional, and thus existing univariate statistics may prove insufficient. Some other metrics to quantify "randomness" include the area of revisions, the average path length of revisions, and the number of 'crossovers' with reference to either the average path or the best track.

Figs. 4(a-c), panels a and b, show maps for two NWP model track forecast revision series in alternating colors, along with the best track shown by the black line. For the earlier valid times, the forecasts changed consistently towards the east-northeast. For later valid times, the revisions are more erratic and were larger. The same plot for the official forecasts is shown in panel c. This figure shows very little adjustment to the official forecasts, with somewhat larger revisions only at the latest valid times. The visual representation shows how different the updates can be from different forecasting systems, and how different NWP forecasts are from the very consistent



Figs. 4(a-c). Plots showing geographic location of hurricane track forecasts from two different NWP models (a and b) and the official forecasts (c). Connected points share the same event (valid) time, so each represents a revisions series. The bold black line shows the 'best track' or actual location (estimated) of the tropical cyclone. Colored tracks are used to distinguish the individual revision series

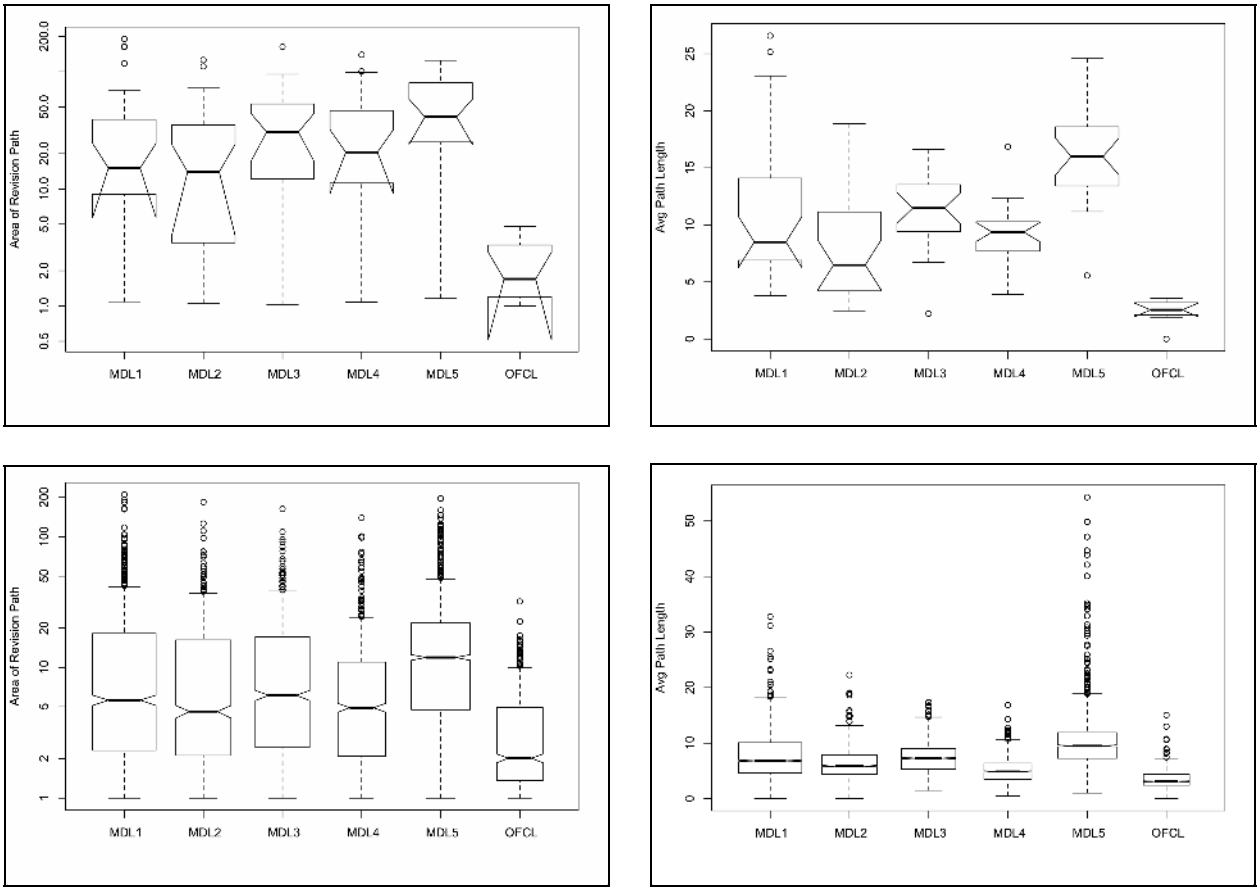


Fig. 5. Four panels with boxplots showing forecast revision path area and average path length for Gabrielle and all 2013 Atlantic Basin tropical cyclones for several NWP model forecasts and the official forecasts

official forecasts. The 'windshield wiper' or zigzagging update is apparent, along with more consistent changes in some of the forecasts.

Wald-Wolfowitz can be applied to cross track errors (switch between positive and negative). Although the number of best track crossovers here (72 out of 251 forecast points) is sufficient to indicate random behavior, it can be seen in the Figure that the randomness probably changed over the storm lifetime. Early tracks show consistent movement towards the east or northeast and convergence on the best track. Later revisions appear more random and tend to cross the best track repeatedly. This conditional behavior is not detected by this test (or any of the other tests considered here), but may be of interest to users. Since track forecasts are usually biased (at least in the Atlantic), the test may need to use average revision location (unbiased) rather than best track.

Fig. 5 shows boxplots (McGill *et al.*, 1978) of revision path area (left) and average revision path length (right) for a single storm (GABRIELLE, top) and 2013 season (bottom). These plots show the distributions of these values that measure consistency. These graphics make it easy to see differences between models, particularly with reference to the official (OFCL) forecasts (which are quite consistent). The revision path area may be insensitive to addition of more or larger revisions. However, it gives the user a sense of the total geographic extent where a storm was forecast to travel over its lifetime. The average path length is very sensitive to larger or additional revisions. It gives the user a sense of how much the forecast typically changes with each update. These particular examples show that the forecasting systems vary pretty drastically for Gabrielle, but taken over the entire 2013 season the forecasting systems are more similar. As expected, the official forecasts are more consistent than any of the NWP forecasts.

5. Summary

Examination of the revisions series and associated statistics can provide forecast users with diagnostic

information about forecasts. In particular, comparison of forecast changes relative to some reference forecast provides information about relative uncertainty, consistency, and magnitude. This work presents some examples of metrics for assessing sets of revision series on weather forecast data.

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 applied to any forecast series of a continuous variable. The two traditional statistical 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, whether applied to a traditional time series or TC track revisions. 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.

TC forecasts are complex in format, making it somewhat difficult to measure consistency. Comparisons between models seem more straightforward than statistical tests of random behavior. This is probably the only respect in which weather forecasts are simpler to evaluate than other time series forecasts. Magnitude and consistency need not be compared to a known statistical distribution, but only to a comparable forecast. For the analyst, this eases the burden of missing data, sets of short time series, etc.

Comparison of univariate sets of revision series via summary statistics provides information about magnitude and consistency of forecast updates. These statistics are both understood and accepted by the weather forecasting community, and this work simply applies these metrics to a different forecast measure, the revision. The interpretation will of course be different than when the same statistics are applied to the errors or other forecast measures.

Prior work on consistency in weather forecasts has typically focused on the magnitude of forecast revisions for a few specific meteorological quantities. Lashley *et al.* (2008) and Ruth *et al.* (2009) both examine proportions of revisions that are large according to some expert criteria specific to the meteorological quantity being forecast. In some cases, large magnitude revisions are weighted less when they occur with very long lead times. An exception to the magnitude focus is the paper by Pappenberger *et al.* (2011), which assesses the randomness through time of flood forecasts via anomaly correlations of time-lagged pairs of forecast errors.

The TC track forecast presents the greatest challenge for measuring consistent forecast updates, but many of the proposed measures show promise for providing useful information to forecast users. The number of best track crossovers is simple to understand and compare, and can be tested for randomness. The total path length gives an assessment of the magnitude of forecast changes, particularly when compared to a reference forecast. The total path area is insensitive to additional revisions within the forecast area boundary, but can give users an excellent sense of the total spatial extent of the forecasts.

The crosstrack error analysis of Elsberry and Dobos (1990) offers some similar assessment of crossovers, but only with reference to the TC best track. Further, they use crosstabulation tables of each type of error (left, center, or right) with each type of subsequent error (left, center, or right). These tables provide users with some level of information about both consistency and magnitude of changes, but without separating these from the errors. This level of detail is probably of interest to sophisticated users of TC track forecasts. Other users may prefer the single values provided by the crossover count test or the revision area. Additionally, a very biased forecast will display "consistent" behavior with reference to the best track since it may rarely cross over, while it may be in fact be quite inconsistent when compared with its own average track which it may cross repeatedly.

Measures of consistency should be examined along with error-only measures to provide a full suite of diagnostic information, as a consistent but poor forecast has little value. As with any type of evaluation, it is essential to examine several complementary measures to ensure that users have access to a complete picture. None of the examined consistency measures is without issues, and there may be pitfalls that are not yet obvious. All measures need to be tested operationally and refined according to weather forecast users' needs.

Revision size is likely to be very dependent on the lead time, with larger (smaller) revisions occurring more often at longer (shorter) lead times. For single time series, which have but a single forecast at each lead time, this difference must be ignored and the time series as a whole evaluated as one entity. However, this work accumulates many time series together for evaluation, which allows the possibility to evaluate similar lead times together if sample sizes allow. In future work, evaluation of the revision magnitude separately for each lead time may prove informative.

A considerable amount of other future work remains. Assessment of other types of forecasts, such as precipitation, is necessary. Since precipitation is typically only a conditionally continuous variable (*e.g.*, the value is 0 much of the time so most revisions will also be 0), the measures described here are unlikely to work well since each test assumes a continuous distribution of values for the revisions. Ehret's (2010) Convergence Index handles precipitation by eliminating any cases with low or no precipitation, and a similar approach may prove useful with the measures proposed here. Assessment of statistical significance via bootstrapping may be possible and should be tested. Refinement of information based on users' interests would make revision assessments more relevant.

Acknowledgements

The authors thank the World Meteorological Organization, the United States Hurricane Forecast Improvement Project, several reviewers, and the staff of the National Hurricane Center. NCAR is sponsored by NSF.

References

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C., 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**, 225-239.

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

Elsberry, R. L. and Dobos, P. H., 1990, "Time consistency of track prediction aids for western North Pacific tropical cyclones", *Mon. Wea. Rev*., **118**, 746-754.

Lanzante, J. R., 2005, "A cautionary note on the use of error bars", *J. Climate*, **18**, 3699-3703.

Lashley, S., Lammers, A., Fisher, L., Simpson, R., Taylor, J., Weisser, S. and Logsdon, J., 2008, "Observing verification trends and applying a methodology to probabilistic precipitation forecasts at a National Weather Service forecast office", 19th conference on Probability and Statistics, New Orleans, LA. American Meteorological Society.

McGill, R., Tukey, J. W. and Larsen, W. A., 1978, "Variations of box plots", *The American Statistician*, **32**, 12-16.

McLay, J., 2011, "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*., **139**, 2, 387-402. doi: 10.1175/2010MWR3449.

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

Pappenberger, F., Cloke, H. L., Persson, A. and Demeritt, D., 2011, HESS Opinions – "On forecast (in) consistency in a hydro-meteorological chain: curse or blessing?", *Hydrol. Earth Syst. Sci*., **15**, 2391-2400, doi:10.5194/hess-15-2391-2011.

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

Sampson, C. R. and Schrader, A. J., 2000, "The Automated Tropical Cyclone Forecasting System (Version 3.2)", *Bull. Amer. Meteor. Soc*., **81**, 1231-1240. doi: http://dx.doi.org/10.1175/1520-477(2000)081<1231:TATCFS> 2.3.CO;2

Snedecor, G. W. and Cochran, W. G., 1989, "Statistical methods", Iowa State University Press., 99-100.

Torn, R. D. and C. Snyder, 2012: Uncertainty of tropical cyclone best track information. *Wea. Forecasting*, 27, 715-729.

Wald, A. and Wolfowitz, J., 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 the ECMWF and UK Met Office EPS control and ensemble-mean forecasts", *Mon. Wea. Rev*., **137**, 3823-3836.