NOAA Technical Memorandum, OAR-AOML-107

https://doi.org/10.25923/m0ah-bh98

Serial Correlation of Tropical Cyclone Track and Intensity Forecasts

S. Aberson

Atlantic Oceanographic and Meteorological Laboratory
Miami, Florida

September 2021

U.S. DEPARTMENT OF COMMERCE

noaa

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
OFFICE OF OCEANIC AND ATMOSPHERIC RESEARCH

Suggested Citation

Aberson, S.D., 2021: Serial correlation of tropical cyclone track and intensity forecasts. NOAA Technical Memorandum, OAR-AOML-107, https://doi.org/10.25923/m0ah-bh98, 6 pp.

Acknowledgments

This study was aided by discussions with Dr. Ghassan Alaka of NOAA's AOML-Hurricane Research Division, and originally stimulated by discussions with Dr. Mark DeMaria and the late Charles Neumann. Mike Jankulak carefully edited the manuscript and greatly improved the presentation, along with three anonymous reviewers who helped to clarify a number of important points. The forecast and best-track data used in this study are available via anonymous ftp from NOAA's National Hurricane Center.

Disclaimer

NOAA Technical Memorandum OAR-AOML-107

https://doi.org/10.25923/m0ah-bh98

Serial Correlation of Tropical Cyclone Track and Intensity Forecasts

Sim D. Aberson

NOAA–Atlantic Oceanographic and Meteorological Laboratory
Hurricane Research Division
Miami, Florida

September 2021

UNITED STATES DEPARTMENT OF COMMERCE
Ms. Gina M. Raimondo, Secretary

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
Dr. Richard W. Spinrad, Under Secretary for Oceans and Atmosphere and NOAA Administrator

OFFICE OF OCEANIC AND ATMOSPHERIC RESEARCH
Mr. Craig N. McLean, Assistant Administrator

This page intentionally left blank.

Table of Contents

This page intentionally left blank.

Abstract

Statistical significance tests can inform whether differences between two samples are real or due to sampling error. Forecasts are serially correlated because the first guess for each model cycle is a forecast from the previous one, and this must be accounted for in statistical tests on the impact of new observing systems or model system techniques. Prior studies showed that tropical cyclone track forecasts created every 24 h or 12 h were serially correlated so that only every other forecast was independent of the others. Forecasts are now initialized more frequently than those used in the earlier studies (every 6 h), requiring a reassessment of the serial correlation. The current study calculates the effective time between independent samples based on two distinct techniques for both tropical cyclone track and intensity forecasts. The calculated effective time varies by storm, forecast, and technique, though it appears that the separation times for both track and intensity are about 12 h/18 h/24 h from lead times 12-36 h/42-96 h/102-120 h, respectively. These calculations may be used to best calculate whether differences between tropical cyclone track and intensity forecasts from various models are statistically significant, and to inform the efficient design of tests of new systems.

This page intentionally left blank.

1. Introduction

An important research thread in the literature during the last few decades is whether various types of observations improve forecasts. Such studies can be done within the framework of an Observing System Experiment (OSE) for data types that already exist, or an Observing System Simulation Experiment (OSSE) for proposed data types. In both types, model forecasts are run both with and without the observations in question. After a sufficient number of forecasts are made, they are verified versus the truth (the real atmosphere for OSEs and the nature run for OSSEs).

If forecasts are improved by the assimilation of the new data types, regular observations from these new platforms are then added to the mix. For example, Burpee *et al.* (1996) showed that dropwindsonde observations in the environments of tropical cyclones improve forecasts of their ultimate track in an OSE; Atlas *et al.* (2001) performed an OSSE for scatterometer data, and this study informed the requirements and characteristics for new scatterometers, and also techniques to assimilate the data into numerical models. Potential issues with these studies are the number of cases utilized and whether the impacts shown are real or based on sampling error. Statistical significance tests can inform whether these impacts are important. Since, in modern data assimilation systems, model forecasts serve as the first guess for subsequent analyses, series of forecasts may be correlated, limiting the usefulness of individual elements of the sample. This must be taken into account for the best assessment of the impact of observations from OSEs and OSSEs.

Hurricane forecasting centers such as the National Hurricane Center, the Central Pacific Hurricane Center, and the Joint Typhoon Warning Center all provide via their websites annual verification statistics for their forecasts and for the numerical modeling systems which help in their preparation. The Hurricane Forecast Improvement Program tracks the progress of numerical tropical cyclone guidance to assess whether its goals have been met. Declarative statements as to how various models perform compared to others, whether new systems are better than older ones, or whether forecasting metrics have been reached must be based on careful and correct statistical calculations.

The literature on how to account for serial correlation in tropical cyclone forecasts began with Neumann *et al.* (1977), who adapted a general procedure from Siegel (1956), and found that track forecasts with their initial conditions separated by less than 30 h are serially correlated. The forecast models tested were statistical, using predictors from a dynamical model and a linear regression technique to forecast the subsequent track. Aberson and DeMaria (1994) used that technique to calculate the serial correlation of a barotropic track forecast model (VICBAR) initialized every 12 h and found that forecasts whose initializations were separated by more than 15-21 h were uncorrelated. These results suggested that calculating statistical significance using no more than one forecast per day is optimal, which cuts the sample size by three quarters for models that are run four times daily. Such calculations have not since been updated in the literature since forecasts became regularly available four times daily and from full physics dynamical models, nor have they been attempted for other forecasts such as for intensity. This study discusses techniques to calculate serial correlation and separation times between uncorrelated forecasts of tropical cyclone track and intensity from the best models currently available and from official forecasts provided by the National Hurricane Center. Values for the serial correlation and separations times are updated for the first time in the literature since 1994.

2. Data

A large sample size is necessary for accurate assessment of the serial correlation between forecasts. Tropical cyclones generally have short lifetimes. Since 2012, only three tropical cyclones have lasted longer than three weeks globally: Hurricanes Nadine (2012), Kilo[1] (2015), and Jose (2017), and the next longest-lived tropical cyclone had a lifetime of just more than two weeks. Forecasts from these tropical cyclones are used in this study.

[1]Kilo moved from the Central Pacific into the Western Pacific on 1 September, and was redesignated a typhoon.

Forecasts from the two National Oceanic and Atmospheric Administration (NOAA) operational dynamical forecast models that predicted tropical cyclone track and intensity at the time (the Hurricane Weather Research and Forecast [HWRF] and the Global Forecast System [AVNO]), and the Official (OFCL) forecasts, are gathered; intensity forecasts from the Statistical Hurricane/Typhoon Intensity Prediction Systems (DSHP) are also collected[2]. Forecasts are verified against best tracks produced by NOAA's National Hurricane and the Central Pacific Hurricane centers. OFCL forecasts were only available during the time that Kilo was east of the dateline; therefore, the sample size is small, and these results are not presented in this study.

3. Techniques

3.1 Laurmann and Gates (1977)

Laurmann and Gates (1977; hereafter LG) devised a technique to check whether an adequate amount of time had been used to calculate a climatic mean from a time series, assuming the data were stationary. Thiébaux and Zwiers (1984) and Zwiers and von Storch (1995) derived and used the same equations, though with different naming conventions. The equation for the standard deviation (σ) of the mean of a variable (x_i)

$$\sigma^2(\bar{x}_i) = \sigma^2(x_i) / N$$

is valid only if all the values are statistically independent. If the values are not statistically independent, the sample size N must be replaced with N', the number of independent samples in the time series. The LG technique calculates N' or the effective time (T_0) between independent samples. Though the time series of errors from short tropical cyclone tracks are unlikely to be stationary, their technique may provide some guidance as to the serial correlation in those errors.

[2]Model designations used herein are consistent with those used at the National Hurricane Center. Descriptions of these designations and models, and official verifications back to 1970, are available at https://www.nhc.noaa.gov/verification/.

They start by calculating the serial correlation coefficient (ρ) for various lag time counts (v),

$$\rho_v = \frac{1}{(N-v)} \sum_{i=1}^{N-v} \frac{(\bar{x}_i - x_i)(x_{i+v} - x_i)}{\sigma^2(x)}$$

where x_i are the sample variables, N is the total sample size, σ^2 represents the variance, and the overbar denotes the mean. If the samples are equally spaced in time, as in the current datasets, then the effective time is related to the serial correlation by

$$T_0 = \frac{T}{N} \sum_{v=-(N-1)}^{N-1} \frac{1-|v|}{N} \rho_v$$

where T is the sampling time, and thus the effective sample size is given by

$$N' \approx N\left((1-\rho_1) / (1+\rho_1)\right)$$

and

$$T_0 \approx T/N\left((1+\rho_1) / (1-\rho_1)\right).$$

3.2 Aberson and DeMaria (1984)

The calculation of stable ρ values from small samples (the result of short tropical-cyclone lifetimes) is difficult. A different technique that overcomes this problem by using a *Runs Test* was taken from Neumann *et al.* (1977) and based on Siegel (1956). Aberson and DeMaria (1984, hereafter AD) used this technique to calculate the effective sample size. Instead of calculating the serial correlation directly, this technique uses the number of times that the ordered values change from one class to another; in this case, the two classes are cases in which the error is above or below the sample mean. The lower the number of runs of cases with values either above or below the sample mean is, the higher the serial correlation will be.

The number of independent observations (N_e) depends on the number of runs of occurrences of a particular class, in this case whether the forecast errors are above or below the sample mean. In this case

$$N_e = N(R_a + R_b) / E,$$

where R_a and R_b are the numbers of groups of consecutive forecasts in which the errors are above or below the sample mean, respectively, and E is the expected sum of R_a and R_b given a random sample. In large samples, this can be approximated by

$$E = 1 + 2\, N_a N_b / (N_a + N_b),$$

where N_a and N_b are the numbers of cases in which the forecast errors are above or below the sample mean, respectively. Therefore,

$$T_0 = TN / N_e.$$

4. Results

Values of T_0 for track and intensity forecasts from the two techniques are provided in **Table 1**. Despite the long lifetimes of the three hurricanes, the total dataset for each is small, and therefore the values are not consistent between different hurricanes. The AD technique suggests that the effective time between independent track and intensity forecasts is about 12 h for 12-h forecasts, increasing to about 24 h by 120 h; T_0 for the Jose AVNO intensity forecasts is a notable outlier. Jose was more intense and remained intense longer than the other two tropical cyclones; because AVNO was run at a lower resolution than currently, all intensity forecasts had a low bias that was serially correlated.

The LG values vary more than those from the AD method, though the general trend is also for the effective time between independent forecasts to increase with lead time. The large variations are likely due to the difficulty of calculating accurate serial correlation coefficients with small samples.

There is a tendency for the effective time for OFCL forecasts to be larger than those from individual models except AVNO intensity forecasts which are serially correlated due to model bias. This is expected since operational protocol is to slowly nudge forecasts toward new model solutions so as to avoid the so-called windshield-wiper effect in which forecasts shift rapidly between multiple solutions; numerical models themselves have no such artificial constraint.

The values shown here are applicable to these particular models and tropical cyclones. A large sample size would be ideal to get reliable statistics across multiple systems, but tropical cyclones have short lifetimes. Similar calculations with shorter-lived tropical cyclones show large variations due to small sample sizes and are therefore not considered reliable estimates of serial correlation.

5. Implications

Due to serial correlation between tropical cyclone track and intensity forecasts, the behavior of any particular model run is related to that of the immediately previous one. For example, in predicting the performance of tropical cyclone track forecasts, the most effective predictor of the model performance is its performance in the previous forecast (Aberson, 1997). Statistical tests must take this into account for the greatest accuracy.

When testing the impact of model system changes or of new observing platforms on those model systems, the models are generally run to mimic operational procedures. Most global and regional hurricane models are now run at least four times daily (every 6 h or more often). The results presented here suggest that model runs during such tests could be completed less frequently than in operations. If serially correlated forecasts are not run, but a similar sample size of runs is conducted, the variety of forecasts in the dataset will be increased. This would assume that the cycled data assimilation systems continue as before, but without the running of the deterministic forecasts from each analysis. If the test is over a set period of time, then less frequent model forecasts would allow such tests to be conducted faster and more efficiently than is currently done. This suggestion necessarily neglects other reasons that serially correlated forecasts should be run, such as the search for outlying forecasts, bugs in computer code, or other issues related to the running of operational models.

Table 1. Effective time (T_0) between independent samples in h for the two techniques, for the two parameters (i is intensity, t is track), tropical cyclone name (TC name), model, and forecast lead time.

T	P	TC name	model	12-h	24-h	36-h	48-h	60-h	72-h	84-h	96-h	108-h	120-h
AD	i	Jose	AVNO	28.4	20.2	29.5	12.2	78.7	21.4	23.7	27.7	16.3	20.0
AD	i	Jose	DSHP	6.8	15.7	16.7	24.8	18.7	43.3	33.9	33.5	34.7	18.4
AD	i	Jose	HWRF	5.3	8.8	15.0	10.2	10.9	11.7	10.9	9.2	11.1	9.9
AD	i	Jose	OFCL	7.8	9.7	15.4	34.4		36.7		17.2		14.7
AD	i	Kilo	AVNO	12.9	12.3	12.6	17.6	13.4	14.7	14.5	13.7	19.8	20.2
AD	i	Kilo	DSHP	7.4	8.9	10.2	12.6	12.0	13.9	11.9	13.2	21.3	14.8
AD	i	Kilo	HWRF	6.4	10.1	13.7	12.5	11.2	13.2	18.7	18.4	21.0	24.6
AD	i	Nadine	AVNO	11.2	8.7	8.8	12.3	11.0	9.1	14.3	8.4	8.6	12.1
AD	i	Nadine	DSHP	8.3	10.4	11.4	15.9	15.2	18.1	10.4	7.8	11.3	8.1
AD	i	Nadine	HWRF	6.4	10.3	11.6	10.7	11.5	13.0	10.4	14.6	16.6	14.8
AD	i	Nadine	OFCL	8.1	10.1	10.9	18.3		18.2		20.1		16.6
LG	i	Jose	AVNO	69.0	120.3	209.4	167.7	224.4	225.1	183.9	120.1	95.7	93.0
LG	i	Jose	DSHP	18.2	50.9	88.3	173.5	293.3	263.9	195.3	121.9	118.1	101.4
LG	i	Jose	HWRF	12.1	12.8	26.1	23.0	26.8	31.8	36.0	31.1	28.2	25.4
LG	i	Jose	OFCL	9.9	23.1	50.5	102.8		116.1		113.6		78.7
LG	i	Kilo	AVNO	44.4	84.1	100.4	122.8	196.1	132.7	161.3	160.9	132.2	170.4
LG	i	Kilo	DSHP	18.1	28.9	48.0	93.3	128.9	157.2	171.9	198.5	178.2	197.0
LG	i	Kilo	HWRF	13.5	26.5	33.1	50.8	49.3	49.6	56.1	86.2	61.9	89.6
LG	i	Nadine	AVNO	30.0	40.9	57.6	55.5	64.9	46.7	52.6	45.7	48.9	57.5
LG	i	Nadine	DSHP	30.9	46.5	70.8	95.0	82.7	63.8	54.6	50.5	38.3	33.5
LG	i	Nadine	HWRF	15.9	40.3	55.4	27.0	21.1	20.3	23.4	42.5	36.8	53.6
LG	i	Nadine	OFCL	23.9	45.7	50.7	80.5		108.7		121.9		148.5
AD	t	Jose	AVNO	6.8	8.2	7.1	10.8	11.1	12.3	10.3	14.2	17.9	17.4
AD	t	Jose	HWRF	7.6	11.0	8.3	9.1	7.5	8.9	10.3	9.3	12.7	11.5
AD	t	Jose	OFCL	8.7	9.1	10.7	9.8		12.9		13.4		11.6
AD	t	Kilo	AVNO	9.7	7.4	9.8	12.9	12.9	18.4	15.5	19.0	24.9	14.8
AD	t	Kilo	HWRF	9.3	8.7	8.5	16.3	20.5	13.3	16.7	17.5	25.7	16.4
AD	t	Nadine	AVNO	7.0	7.5	10.1	10.1	10.5	11.1	11.9	11.2	11.5	10.9
AD	t	Nadine	HWRF	9.4	13.0	10.4	12.9	16.2	13.1	10.2	9.4	8.6	8.4
AD	t	Nadine	OFCL	10.1	13.3	14.9	15.9		11.2		18.1		17.1
LG	t	Jose	AVNO	10.4	13.2	10.1	11.3	14.1	16.1	17.1	27.8	32.6	33.7
LG	t	Jose	HWRF	12.6	18.3	16.5	19.6	26.3	41.4	36.1	34.2	26.8	18.1
LG	t	Jose	OFCL	9.8	13.1	17.7	23.7		22.8		25.6		38.3
LG	t	Kilo	AVNO	28.2	18.6	19.7	18.2	21.0	24.1	31.1	31.5	28.5	31.1
LG	t	Kilo	HWRF	18.3	17.6	29.3	29.4	21.8	22.0	33.2	50.8	73.0	68.7
LG	t	Nadine	AVNO	12.8	10.3	16.2	21.4	21.2	18.8	19.1	17.3	18.2	18.5
LG	t	Nadine	HWRF	19.2	35.3	46.6	49.8	50.0	42.1	38.0	23.9	17.6	14.0
LG	t	Nadine	OFCL	16.2	41.6	53.6	87.6		56.4		53.5		58.1

Study of the serial correlation of tropical cyclone forecasts is limited by the short lifetimes of individual tropical cyclones. Similar calculations using global or regional weather patterns will not have this problem (e.g., Hering and Genton, 2011; Gilleland *et al.,* 2018), and they may show different levels of serial correlation for such phenomena or regions. Other techniques to deal with serial correlation between forecasts, largely from economics, include a variance inflation factor (Kutner *et al.,* 2004). Diebold and Mariano (1995) proposed a series of tests of whether two forecasts are statistically different while accounting for serially correlated errors. Their tests are different in that they do not calculate an effective time between independent samples or an effective sample size, but directly allow forecast errors to be non-Gaussian and serially and contemporaneously correlated. These techniques may be usefully applied to tropical cyclones forecasts, but this is beyond the scope of this study. The current study was done to update the effective sample size or time between independent forecasts so that current operational verification techniques can be updated.

6. References

Aberson, S.D., 1997: The prediction of the performance of a nested barotropic hurricane track forecast model. *Wea. Forecasting*, 12, 24–30.

Aberson, S.D., and M. DeMaria, 1994: Verification of a nested barotropic hurricane track forecast model (VICBAR). *Mon. Wea. Rev.*, 122, 2804-2815.

Atlas, R., and Coauthors, 2001: The effects of marine winds from scatterometer data on weather analysis and forecasting. *Bull. Amer. Meteor. Soc.*, 82, 1965-1990.

Burpee, R.W., J.L. Franklin, S.J. Lord, R.E. Tuleya, and S.D. Aberson, 1996: The impact of Omega dropwindsondes on operational hurricane track forecast models. *Bull. Amer. Meteor. Soc.*, 77, 925-993.

Diebold, F.X., and R.S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*,13, 253-263, http://doi.org/10.1080/07350015.1995.10524599.

Gilleland, E., A.S. Hering, T.L. Fowler, and B.G. Brown, 2018: Testing the tests: What are the impacts of incorrect assumptions when applying confidence intervals or hypothesis tests to compare competing forecasts? *Mon. Wea. Rev.*, 146, 1685-1703, https://doi.org/10.1175/MWR-D-17-0295.1.

Hering, A.S., and M.G. Genton, 2011: Comparing spatial predictions. *Technometrics,* 53, 414-425, https://doi.org/10.1198/TECH.2011.10136.

Kutner, M.H., C.J. Nachtsheim, and J. Neter, 2004: *Applied Linear Regression Models,* 4th ed. McGraw-Hill, 701 pp.

Laurmann, J.A., and W.L. Gates, 1977: Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. *J. Atmos. Sci.*, 34, 1187-1199.

Neumann, C.J., M.B. Lawrence, and E.L. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone models. *J. Appl. Meteor.*, 16, 1165-1174.

Siegel, S.I., 1956: *Nonparametric Statistics for the Behavioral Science*. McGraw-Hill, 312 pp.

Thiébaux, H.J., and F.W. Zwiers, 1984: The interpretation and estimation of effective sample size. *J. Climate Appl. Meteor.*, 23, 800-811.

Zwiers, F.W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate*, 8, 336-351.

National Oceanic and Atmospheric Administration
Office of Oceanic and Atmospheric Research
Atlantic Oceanographic and Meteorological Laboratory
4301 Rickenbacker Causeway
Miami, FL 33149
https://www.aoml.noaa.gov