Section 10

Forecast verification: methods and studies.

An update on linear regression and error correlation: Exploration of baseline climate change impacts on Arctic and North Atlantic fog

Richard E. Danielson[1,2] (rickedanielson@gmail.com), William A. Perrie[2], and Minghong Zhang[2]
[1]Danielson Associates Office, Inc., Halifax, Nova Scotia
[2]Fisheries and Oceans Canada, Bedford Institute of Oceanography, Dartmouth, Nova Scotia

Introduction

The experience of fog in nature is as ethereal as it is challenging to capture in observations and numerical forecast models. Visibility is a measure of fog that is readily diagnosed from surface humidity, but seemingly not without bias. Gultepe and Milbrandt (2010) provide one parameterization that Danielson et al. (2020) use to explore 21st century regional trends in marine visibility. Following an approach that can be described as conventional, all linear adjustments developed in that study assume that the forecast model is further removed from reality than observations (i.e., in every way). By contrast, one may also question whether an analysis (specifically, parameterizations of visibility applied to an analysis that benefits from a forecast model and observations) can be considered in various ways better than, equivalent to, as well as worse than an observational proxy of reality. An emphasis on forecast model strengths that are *complementary* to the strengths of observations is an important proposition of Parker (2016). Oreskes et al. (1994) and Beven (2019) provide further motivation for the philosophical challenge of whether to accommodate an error of representation associated with the forecast model (or analysis) that is equivalent to an error of representation associated with the observations (cf. Daley 1991). It is important to acknowledge that representation error is a misnomer here, insofar as its inclusion provides a better representation of visibility.

An Updated Linear Regression

It is only by longstanding convention that historical marine visibility observations are recorded (i.e., as one of ten categories). Further standardization of these measures is not anticipated (cf. World Meteorological Organization 2017), but at least one may take them as a good indication of the presence or absence of fog. The same might be said of estimates of visibility derived from numerical forecast data. Although both estimates are exploratory, comparisons are also instructive. A nascent approach to comparing such measures (with no one dataset assumed to be uniformly better) is given by Danielson (2018). Our update focuses on a canonical (yet imperfect) measurement model, or linear regression framework (Danielson et al. 2020). If any dataset is a partial measure of truth with error, then a numerical translation from uncalibrated (U) to calibrated data (C) can be represented by an additive (α_U) and multiplicative (β_U) adjustment, where

$$\begin{aligned} C &= t + \epsilon + \epsilon_C \\ U &= \alpha_U + \beta_U t + \epsilon + \epsilon_U. \end{aligned} \tag{1}$$

We describe t as a *linear association* that is only partially shared by both datasets, and hence, linear calibration can be only partial as well. An interesting consequence of equation (1) is that the covariance between C and U (e.g., between in situ and gridded estimates of visibility) also involves *nonlinear association*. From a metrological point of view, equation (1) is a canonical expression of errors-in-variables linear regression (Fuller 2006; Dunn 2011), except that it allows Fuller's equation error (ϵ, the *nonlinear association* term) to be shared between two different datasets. The interpretive consequences of adding such a term are not yet fully understood, but with the benefit of well sampled data (Danielson 2018), relatively direct numerical solutions of (1) are available.

Conclusions

A linear calibration of forecast model output helps to reveal consistent decreasing trends in 21st century marine visibility (Danielson et al. 2020). Given the use of a conventional visibility parameterization, these

trend estimates take in situ observations as a *perfect* reference. While this is consistent neither with equation (1) nor with the subjective nature of marine visibility observations, it provides a useful baseline and is easy to interpret. However, there seems to be a gap in our ability to interpret a seemingly simple linear calibration when this involves measurement error in both visibility datasets. Thus, we have begun to explore more than just linear calibration and to accommodate an interpretation of more than just measurement error among ϵ, ϵ_C, and ϵ_U. Idealized control experiments are also being explored.

Acknowledgements

This work has been supported by the Ocean Frontier Institute, the Belmont Forum, the Office of Energy Research and Development, and Fisheries and Oceans Canada.

References

Beven, K., 2019: Towards a methodology for testing models as hypotheses in the inexact sciences. *Proc. Roy. Soc. A*, **475**, 1–19, doi:10.1098/rspa.2018.0862.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, New York, New York, 457 pp.

Danielson, R. E., 2018: On retrieving parameters of a linear regression model that accommodates error correlation in well sampled data. Working Group on Numerical Experimentation Research Activities (Blue Book) accessed May 2020 at http://bluebook.meteoinfo.ru/uploads/2018/sections/BB_18_S10.pdf.

Danielson, R. E., M. Zhang, and W. A. Perrie, 2020: Possible impacts of climate change on fog in the Arctic and subpolar North Atlantic. *Adv. Statist. Clim. Meteor. Ocean.*, **475**, 1–19, doi:10.5194/ascmo–6–31–2020.

Dunn, G., 2011: Method Comparison Studies, *Int. Encyclopedia of Statistical Science*, M. Lovric, Ed., Springer, 815–816, doi:10.1007/978-3-642-04898-2_36.

Fuller, W. A., 2006: Errors in variables. *Encyclopedia of Statistical Sciences*, S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic and N. L. Johnson, Eds., doi:10.1002/0471667196.ess1036.pub2.

Gultepe, I., and J. A. Milbrandt, 2010: Probabilistic parameterizations of visibility using observations of rain precipitation rate, relative humidity, and visibility. *J. Appl. Meteor. Clim.*, **49**, 36–46, doi:10.1175/2009JAMC1927.1.

Oreskes, N., K. Shrader-Frechette, and K. Belitz, 1994: Verification, validation, and confirmation of numerical models in the Earth sciences. *Science*, **263**, 641–646, doi:10.1126/science.263.5147.641.

Parker, W. S., 2016: Reanalyses and observations: What's the difference? *Bull. Amer. Meteor. Soc.*, **97**, 1565–1572, doi:10.1175/BAMS–D–14–00226.1.

World Meteorological Organization, 2017: Guide to Meteorological Instruments and Methods of Observation, Chapter 9, Measurement of visibility, WMO-No. 8, Geneva.

SCANTEC: A Community System for Evaluation of Numerical Weather and Climate Prediction Models

João Gerd Zell de Mattos*, Luiz Fernando Sapucci, Carlos Frederico Bastarz, Ariane Frassoni, Wanderson Santos, Arletis Carrasco

*Center for Weather Forecasting and Climate Studies, Brazilian National Institute for Space Research, Cachoeira Paulista, SP, Brazil

1. Introduction

The Brazilian meteorological community has few model verification tools following the World Meteorological Organization (WMO) recommendations. The National Institute for Space Research (INPE), Center for Weather Forecasting and Climate Studies (CPTEC) operational Numerical Weather Prediction (NWP) center is providing operationally the Community System for Evaluation of Numerical Weather and Climate Prediction Models – SCANTEC (from the acronym in Portuguese to ***S**istema **C**omunitário de **A**valiação de modelos **N**uméricos de **T**empo **E** **C**lima* – de Mattos and Sapucci 2017) in order to contribute with the improvement of model quality assessment. SCANTEC project aims to offer for the community and operations a unified, standardized, and flexible tool for forecast verification.

The SCANTEC project is under management in a flexible institutional project management web application at CPTEC. SCANTEC Version 1.0 includes traditional measures for categorical and continuous variables, like the Root Mean Square Error (RMSE), bias, and Anomaly Correlation (AC). Besides, the package provides advanced spatial forecast evaluation techniques in research mode, like the Method for Object-based Diagnostic Evaluation (MODE). SCANTEC advantages span in the flexible integration of modeling systems employed by different institutions in Brazil, or different versions of the same modeling system. SCANTEC is also flexible to receive new statistical metrics (de Mattos and Sapucci 2017).

This paper aims to describe the main components of SCANTEC and its potential as an open-source, community-based development software for model verification. Section 2 presents the basic structure of SCANTEC Version 1.0, in which the main features are described. Section 3 presents the statistical metrics available, and finally, Section 5 describes the planning of future developments, including the capability for broad use in supercomputing environments and future applications.

2. Basic structure of SCANTEC

SCANTEC is a system based on open-source tools and can be run on different operating systems, such as UNIX, Linux, Windows, and macOS. The system structure includes a kernel fully modular to facilitate the implementation of new features. The kernel is developed in Fortran 2003 programming language, following the ANSI standard. Besides, SCANTEC is userfriendly and configurable through American Standard Code for Information Interchange (ASCII) files. The main components of SCANTEC are illustrated in Figure 1.

SCANTEC was designed by software development practices that encourage the reuse and community sharing of algorithms among the scientific community. The components were designed as functional abstractions using flexible object-oriented programming paradigms to facilitate reuse and the development of future implementations. Interoperable features in SCANTEC also include reuse and joint development with other numerical modeling groups. Similar to the nature of object-based in structured programming, SCANTEC provides standard functionalities for model evaluation and allowing the user to fill variable functionalities according to their needs. The number of variable functionalities in SCANTEC includes interfaces to facilitate the incorporation of (1) domains, (2) numerical models interfaces, (3) types of observations, and (4) statistical methods for evaluation. A set of abstract functions are incorporated to represent the variable functionalities. These interfaces called "plugins" (model-plugins, obs-plugins, stat-plugins), contain access points or extensible interfaces to incorporate routines to read new models and statistical metrics not included in the stat-plugins component. The model-plugin component is responsible for model data reading; obs-plugin is responsible for the reference dataset reformatting and access a dataset server, and stat-plugin performs the computation of statistical metrics; visualization and post-processing tools provide iterative access to SCANTEC products.

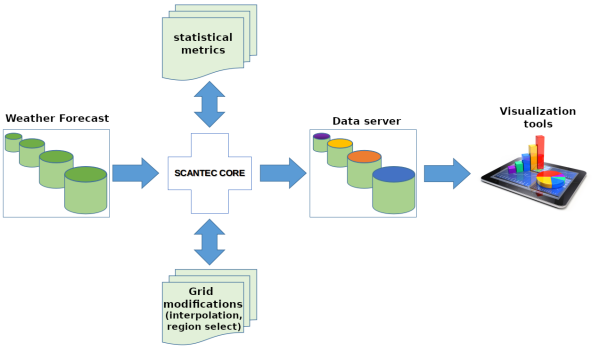

Fig. 1. The main components of SCANTEC.

3. Statistics and visualization

SCANTEC includes two modules that perform statistical computations: Method For Object-Based Diagnostic Evaluation (MODE) and basic-statistic. Both are components of stat-plugin. The basic-statistic module includes standard statistical metrics for comparing forecasts and grid point references. This functionality is suitable for comparing model outputs with its analysis, and also perform intercomparison between different models or experiments performed with the same model. In addition to MODE, SCANTEC also offers methods for dichotomous forecast evaluation, extracting information from the contingency table after considering, for the precipitation field, cer-

* *Corresponding author address:* João Gerd Zell de Mattos, Center for Weather Forecasting and Climate Studies, Rodovia Presidente Dutra Km 40, Cachoeira Paulista, SP, Brazil

E-mail: joao.gerd@inpe.br

tain thresholds commonly used by the community. The metrics available in Version 1.0 are: Anomaly Correlation (CC), Root Mean Square Error (RMSE) and Mean Error (ME), metrics for specific precipitation assessment such as frequency histogram and contingency table.

SCANTEC produces an output file in ASCII, which contains the average statistical results on the selected domain as a function of lead time or time of day. A file in a sequential binary format is produced, containing the statistical results for each grid point of the domain and period. SCANPLOT tool performs the visualization of statistics tables provided by SCANTEC. This tool consists of a set of scripts written in Python, where graphical outputs include visualization of statistics in the scorecard and Taylor Diagram format. Figure 2 shows an example of a scorecard provided by SCANTEC, highlighting the RMSE improvement in different variables and levels evaluated (Sapucci et al. 2016).

4. Spatial verification method

Traditional metrics are not sufficiently informative to evaluate numerical models, especially those with high horizontal resolution. MODE is an object-based method to verify properties of spatial forecasts of entities, where an entity is anything that can be defined by a closed contour (Ebert and McBride 2000). This technique emulates the visual identification of a forecaster analyzing the meteorological field, identifying matched objects, and then comparing each other (Davis et al. 2006, 2009). MODE has been implemented by (Carrasco 2017) and was applied to evaluate precipitation forecasts and intercompare the Brazilian developments on the Regional Atmospheric Modeling System (BRAMS) and the Weather Research and Forecasting (WRF) Model (Carrasco et al. 2020). Recent work at CPTEC has used MODE to identify and evaluate forecasts of heatwaves predicted by two versions of BRAMS model using ECMWF ERA5 reanalysis and GFS analysis as the reference database (Garcia 2020).

5. Future developments

SCANTEC is a project in progress, and future developments consider community needs and contributions. A graphical web interface based on the Python Jupiter notebook tool is under development, which should be migrated to a stand-alone graphical interface. NetCDF (network Common Data Form) format will be included as one of the data formats supported in SCANTEC. Observational datasets provided in PrepBuffer and ASCII, which include conventional data such as SYNOP, SHIP, METAR, among others, will also be supported and included in the list of references database available in obs-plugin. As many centers work in a high-performance computing environment, parallel processing is desired for SCANTEC to provide faster and more efficient statistical computation. To meet this aim, the parallelization of SCANTEC is required and will be available in future versions. Scientific applications restricted to meteorological variables are considered to be extended to air quality variables, as CPTEC is a producer of operational air quality forecasting for South America, and recognizes the need for an operational procedure for air quality forecasting verification.

6. Summary and Conclusions

SCANTEC has been developed at INPE/CPTEC for use by the internal community in NWP assessment. SCANTEC Version 2.0 is under release in the operational CPTEC NWP environment. The tool is applied to evaluate meteorological variables and offers a flexible environment for user needs under a userfriendly configuration to another modeling system other than those currently available. SCANTEC has been applied over the last year in scientific studies and is under the GNU General Public License. As a community tool, SCANTEC is open-source software that will be available to community contributions to enhance the tool and keep it relevant for scientific applications.

Acknowledgments. SCANTEC has been sponsored by The Brazilian National Council for Scientific and Technological Development (CNPq) and Foundation for Research Support of the State of São Paulo (FAPESP). We acknowledge Dr. Chou Sin Chan, head of Modeling and Development Branch at CPTEC, for her support and effort to release the operational version of SCANTEC for the INPE/CPTEC NWP community.

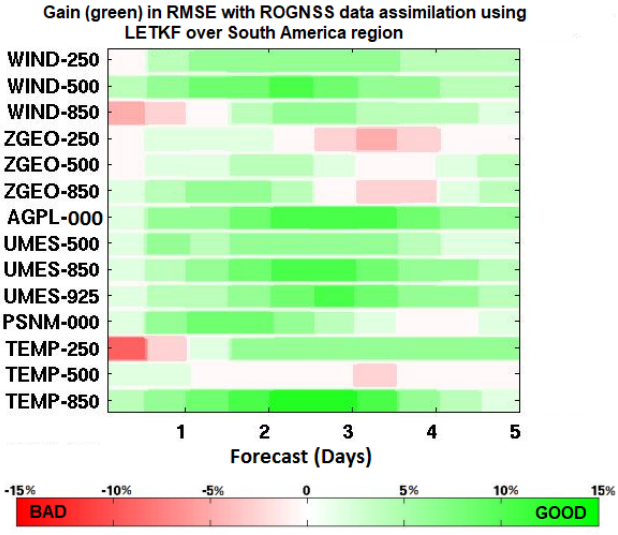

FIG. 2. Type of analysis provided by SCANTEC exploring scorecard shows the gain in RMSE over South America after the assimilation of radio occultation data. More detail about this study is available in Sapucci et al. (2016).

References

Carrasco, A. R., 2017: Método de avaliação orientada a objeto aplicado às previsões de precipitação sobre a américa do sul. M.S. thesis, Instituto Nacional de Pesquisas Espaciais (INPE), 116 pp., São José dos Campos, URL http://urlib.net/rep/8JMKD3MGP3W34P/3NH9KMB.

Carrasco, A. R. C., L. F. Sapucci, J. G. Z. de Mattos, M. S. Lorenzo, and I. B. Monteiro, 2020: Exploring the particularities of the method object-based in the precipitation forecast evaluation, In publication process. *Revista Brasileira de Meteorologia*.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, **134 (7)**, 1772–1784.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (mode) applied to numerical forecasts from the 2005 nssl/spc spring program. *Weather and Forecasting*, **24 (5)**, 1252–1267.

de Mattos, J. G. Z., and L. F. Sapucci, 2017: BR 51 2017 000576-1. Scantec - sistema comunitÁrio de avaliaÇÃo de modelos numÉricos de tempo e clima. 13-junho-2017.

Ebert, E., and J. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology*, **239 (1–4)**, 179–202.

Garcia, G. R., 2020: Object-based evaluation of the impact of burning aerosols on heat waves forecast in south america. Master dissertation: Meteorology, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, document in Portuguese.

Sapucci, L. F., F. L. R. Diniz, C. F. Bastarz, and L. A. Avanço, 2016: Inclusion of GNSS radio occultation data into CPTEC Local Ensemble Transform Kalman Filter (LETKF) using the ROPP as an observation operator. *Meteorological Applications*, **23 (2)**, 328–338, doi:10.1002/met.1559.

Verification of the NCEP/EMC Unified Forecast System for Subseasonal to Seasonal Timescales

Lydia Stefanova[1*], Jessica Meixner[2], Avichal Mehra[2], Partha Bhattacharjee[1], Robert Grumbine[2], Bin Li[1], Shrinivas Moorthi[2], Jiande Wang[1], Denise Worthen[1]
[1]IM Systems Group at NOAA/NWS/NCEP/EMC, [2]NOAA/NWS/NCEP/EMC
[*]email: Lydia.B.Stefanova@noaa.gov

1. The UFS system for seasonal to subseasonal prediction

The NCEP Environmental Modeling Center Unified Forecast System (UFS) is a community-based modeling system designed for weather and climate forecasting on global or regional scales. The configuration of UFS for seasonal to subseasonal timescales is currently under development; at present it consists of atmosphere, ocean, and sea ice component models, coupled through a NEMS mediator. The addition of a coupled global wave model (WAVEWATCH III) is planned in the near future. The atmospheric model in this system is composed of the Finite Volume Cubed Sphere (FV3) dynamical core with GFS physics and GFDL microphysics parameterization. The oceanic model is the Modular Ocean Model (MOM6), and the sea ice model is the Los Alamos Sea Ice Model (CICE5). Upgrades and bug fixes to the system components are constantly incorporated as model components are updated by community effort.

2. UFS prototypes and benchmark framework

As the system grows in maturity and complexity, a systematic monitoring of performance is needed to ensure its quality. As part of this monitoring, sequential system prototypes (identified by specific components, settings, and initial conditions) specified in the course of development are validated and verified within a fixed “benchmark” framework. This benchmark framework is designed to test system performance for each new prototype with a consistent structure and fixed metrics.

The consistent structure is provided by requiring each prototype to produce a set of 35-day coupled forecasts initialized on the first and fifteenth day of every month between April 2011 and March 2018 for a total of 168 forecasts. The length of this dataset is a balance between providing a sufficient length for statistical analysis and limiting the strain on computing resources. The chosen period for benchmark verification spans varying climate conditions, as it includes several El Niño and La Niña events, as well as recent years of both high and low Arctic ice extent.

Since the model components are constantly being upgraded, it is not feasible to conduct a benchmark evaluation after every change. Instead, benchmark testing is performed at specific milestones of system development. To date, four prototypes, primarily targeting the impact of changing the source of initial conditions, have been defined and fully evaluated. The first prototype, UFS_p1, consisted of model components as described above, with the component versions current as of Oct 2018, and CFSR initial conditions for atmosphere, ocean, and sea ice initialization. For Prototype 2 (UFS_p2), the model components were updated to their then-current Mar-2019 states, and the initial conditions for the ocean were replaced with the 3Dvar from the NCEP Climate Prediction Center (CPC) GODAS. For Prototype 3 (UFS_p3), the model components were updated to their Jun-2019 states, and additionally the sea ice initial conditions were replaced with an ice analysis developed by CPC. An additional intermediary Prototype 3.1 (UFS_p3.1) was created to assess the impact of the coding changes implemented between Jun 2019 and Jan 2020, with the same initial conditions as in UFS_p3. A separate prototype (UFS_p3.2) is in the process of being run for tests to document the impact of atmospheric initial conditions only, while holding the code base the same as in UFS_p3.1.

3. Metrics

The main benchmark verification metrics consist of bias, RMS errors and anomaly correlations (AC) for a set of surface and upper air fields by lead week. Anomalies are calculated with respect to a smoothly interpolated climatology calculated by fitting the 7-year time series to a sine wave of period 365.25 days plus three harmonics. The smoothly interpolated climatology is calculated in the same way for both forecast and verification fields, separately for each grid point and lead time. Verification is performed against the CPC global 0.5-degree Unified Rain Gauge data (for precipitation over land), 6-hourly analysis guess 6-hr predictions from operational CFSv2 CDAS (for precipitation over ocean and upper air fields), CPC global 0.5 degree daily 2-meter temperatures, daily 0.25-degree OSTIA SST analysis, 500-hPa geopotential 6-hourly analyses from the operational CFSv2 CDAS. Model and verification data sets are interpolated to a common resolution prior to anomaly calculations. In addition,

MJO index RMM1 & RMM2, and bivariate correlation skill are calculated following Wheeler and Hendon (2004) and Lin et al. (2008).

4. Results

For brevity, we focus here on week 3 and 4 AC scores, as this is the lead time for which subseasonal forecasts hold the most unrealized potential. Skill at shorter lead times is larger for all benchmark comparisons, but the conclusions regarding relative performances are similar. Beginning with the first prototype, the UFS system offers an improvement over the operational CFSv2 in terms of the week 3 and 4 AC scores for most fields (Fig. 1, left panel). The replacement of ocean initial conditions between UFS_p1 and UFS_p2 provided an additional skill improvement. The subsequent replacement of sea ice initial conditions between UFS_p2 and UFS_p3 did not have a beneficial impact for these scores for the fields shown here; it did however result in more accurate ice concentration threat scores (not shown). Little change was seen between UFS_p3 and UFS_p3.1. The lead time until the MJO bivariate correlation falls to 60% in the benchmark comparison went from 12 to 16.5 days between CFSv2 and UFS_p1, 19 days in UFS_p2, and 18 days in both UFS_p3 and UFS_p3.1 (Fig. 1, right panel, colored bars). These results are encouraging when compared to the individual models from the WWRP/WCRP sub-seasonal to seasonal prediction (S2S) project (Fig. 1, right panel, grey bars; Vitart 2017)

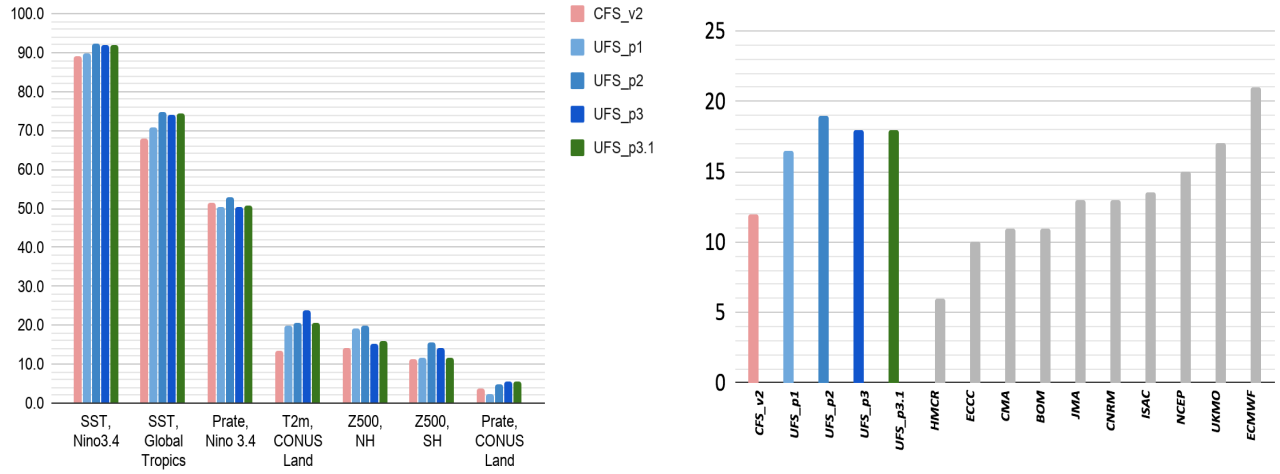

Figure 1. *Left*: Weeks 3 and 4 anomaly correlation (%) for select fields from benchmark runs (Apr 2011-Mar 2018) of CFSv2 and UFS prototypes 1 through 3.1. *Right*: Forecast lead time (days) at which the MJO bivariate correlation falls to 60%. Colored bars represent benchmark runs (Apr 2011-Mar 2018) of CFSv2 and UFS prototypes 1 through 3.1. Grey bars represent control runs (i.e., not ensembles) from various S2S models for 1999-2010 (based on Vitart, 2017).

Additional evaluations across prototypes demonstrate that ongoing developments have not altered the overall pattern of biases for most fields, and the prototypes are generally biased warm and wet. Across the lineup of prototypes, the largest boost in AC skill was associated with changing the ocean initial condition from CFSR to the 3Dvar CPC. AC scores for subsequent prototypes are comparable and remain an improvement over the operational CFSv2. This provides confidence in the system as components are refined; as system complexity increases, the reduction of biases and further skill improvement remain a target. It is likely that the greatest benefit for future performance is to be gained from planned component physics improvements and tuning, and advances in initializations, e.g., via land DA.

References

Lin, H., G. Brunet, and J. Derome, 2008: Forecast skill of the Madden–Julian Oscillation in two Canadian atmospheric models. *Mon. Wea. Rev.*, **136**, 4130–4149.

Saha, S., S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y. Hou, H. Chuang, M. Iredell, M. Ek, J. Meng, R. Yang, M.P. Mendez, H. van den Dool, Q. Zhang, W. Wang, M. Chen, and E. Becker, 2014: The NCEP Climate Forecast System Version 2. *J. Climate*, **27**, 2185–2208.

Vitart, F., 2017: Madden–Julian Oscillation prediction and teleconnections in the S2S database. *Q.J.R. Meteorol. Soc*, **143**, 2210-2220.

Wheeler, M.C. and H.H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932.

Ten-year Performance of HWRF Model in RI Forecasts -- A New Metric

Weiguo Wang, Bin Liu, Zhan Zhang, Lin Zhu, Avichal Mehra* and Vijay Tallapragada*
IMSG@EMC/NCEP/NWS,*EMC/NCEP/NWS, College Park, MD 20740
Email: Weiguo.Wang@noaa.gov

1. Introduction

Forecasts of rapid intensification (RI) of tropical cyclones (TC) are still a challenge, in spite of improvements in track and intensity forecasts in the past decade. RI is a scenario where the intensity of a TC increases dramatically in a very short period of time. In practice, RI is defined as an increase in the maximum sustained winds of a TC equal to or greater than 30 knots (55 km/h) in a 24-hour period (Kaplan and DeMaria, 2003). Improving the ability of NCEP hurricane models to forecast RI events is a top priority for EMC developers. Currently, the probability of detection (POD) and false alarm ratio (FAR) are routinely used to measure the performance of RI forecasts. POD of RI is quantified as a percentage of the total number of observed individual RI events which are correctly forecasted, while FAR of RI is a percentage of RI forecasts that were not RI events based on observations. While this method is effective in assessing the overall RI forecast performance of a model, it is not straightforward in revealing how well individual forecasts over a period of time (e.g., typically 5 days for mesoscale models) perform in capturing RI events. In other words, a POD may not be able to reflect how many 5-day forecasts successfully capture RI events. This is because there may be multiple RI events during a 5-day period. To this end, we proposed a new metric, which is based on the total number of RI events forecasted during the whole integration time in a model. This gives modelers a direct assessment of the number or percentage (i.e., success rate) of 5-day forecast cycles capturing some or all of RI events. With the new metric, we calculated success rates of RI forecasts in a 5-day period by NCEP HWRF in the past decade, showing the model is improving RI forecasts.

2. Methodology

The question we would like to answer is how many 5-day model integrations (cycles) can successfully capture one or more observed RI events based on best-track data. For a threshold of wind speed increase, RI events can be identified as binary (yes or no) every six hours during a 5-day integration period. The same procedure is applied to observational data (e.g., NHC best-track data). Then one can compare the results from the model with observational data, and determine how many observed RI events have been captured by the 5-day forecast. To illustrate the method, Figure 1 presents an example of a 5-day time series of the maximum 10-m wind speed of Hurricane Lorenzo (2019) forecasted by the operational HWRF model initialized at 18 UTC, September 24, 2019. RI events (with the threshold of 30 kt) are identified every 6 hours, denoted by blue crosses for observations and triangles for HWRF. The best-track data suggests RI cases occurred at 10 lead times (hours). RI cases predicted by HWRF occurred at 4 lead times. Triangles in red indicate that the occurrence times of RI events simulated by HWRF exactly match those of observations. In this example, RI cases at the 36th, 42nd, and 48th hours are successfully captured by HWRF. HWRF produced one false alarm prediction at the 54th hour and failed to capture RI at seven lead times (hours).

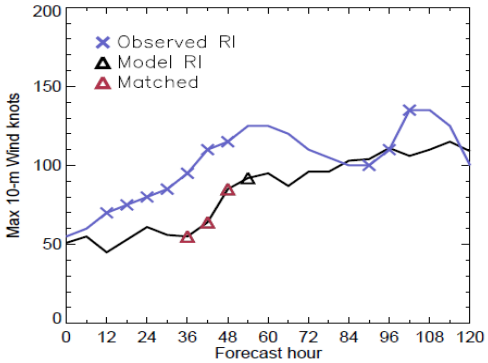

FIG 1 Time series of maximum 10-m wind speed of 5-day periods from a HWRF forecast (black line) and from observation data (blue line).

Given the uncertainties in numerical models and errors in observations, multiple criteria or thresholds can be used to determine whether a 5-day forecast by HWRF is a success or not. Three criteria were tested, depending on what percentage of observed RI cases have been produced in a 5-day forecast by HWRF. The first requires that all observed RI cases must be forecasted by HWRF with matching RI occurrence times. This is a very strict criterion, particularly for multiple RI events occurring in a 5-day forecast period. The second is somewhat relaxed, and requires that half of observed RI cases are produced (with time

matching). The third is most relaxed criterion, requiring that at least one observed RI case is produced (with time matching). A false alarm RI cycle is defined as a 5-day forecast during which HWRF predicts RI events but there are no observed RI events in that period. A success rate is a percentage of total 5-day periods with observed RI events successfully produced by HWRF. In addition, due to the fact that observation error in intensity can be greater than 5 knots, we used 20 and 30 knots as the thresholds of the increase in wind speed when determining RI cases for comparisons.

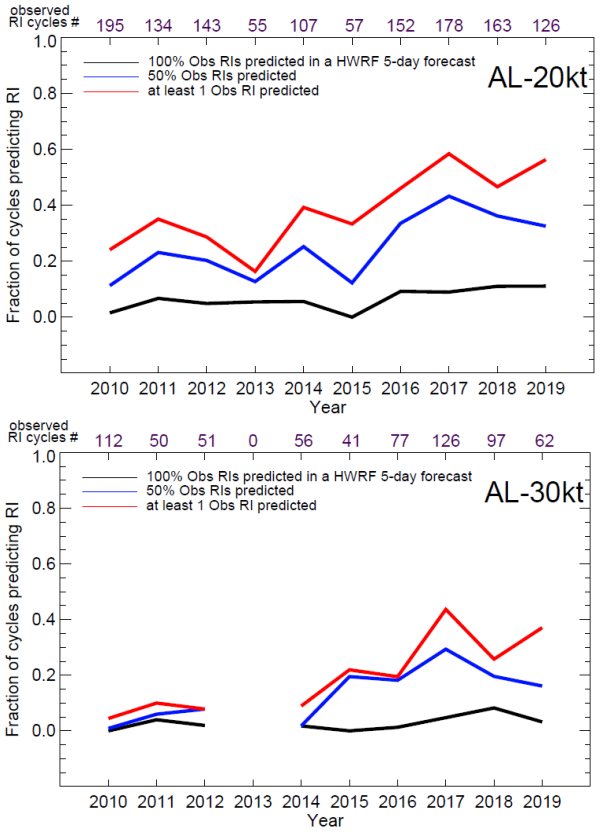

FIG. 2 Success rates of 5-day forecasts predicting observed RI events of TCs in NATL basin by operational HWRF model from 2010 to 2019. top: 20-kt as RI threshold , bottom: 30 kt.

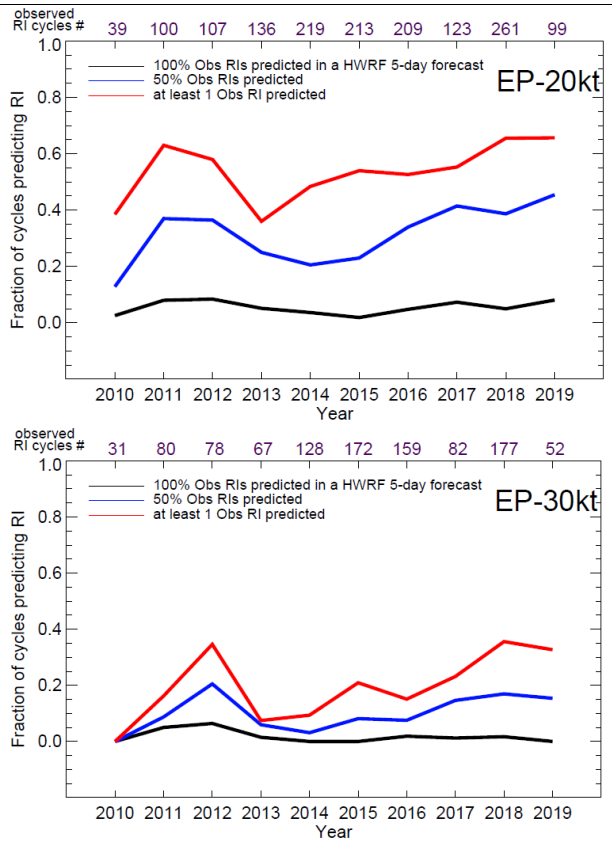

FIG. 3 as FIG. 2 except for TCs of EPAC basin.

3. Results and discussion

With the method described above, we computed success rates of 5-day forecasts for the North Atlantic (NATL) basin (Fig. 2) and East Pacific (EPAC) basin (Fig. 3) by the operational HWRF model over the past 10 years. Depending on the criterion, the fraction of the number of cycles producing observed RI events increases with years. The success rate is higher with a lower RI threshold. For the 30-kt RI threshold, as much as 40% of the cycles can predict at least one RI event during a 5-day forecast. This number increases to 60-70% for the 20-kt RI threshold. However, the success rate of cycles in predicting all observed RI events in a 5-day forecast (with the time matching requirement) is still very low, though there is an increasing trend in the NATL basin. Out of the total 5-day periods when observations did not show RI events, approximately 20% of forecasts (in the same period) predict at least one RI event, giving false alarms. This number is not changed much over the years. The improvement in success rate is attributed to yearly upgrades of the HWRF model, especially model horizontal resolution increases from 9 to 1.5 km, and in the number of vertical levels from 43 to 75 km. Tuning and calibrations of parameters in the model physics schemes, PBL and convection schemes for example, also play a key role in the improvement of the intensity and intensity change forecasts. We also calculated the intensity RMS error and bias during RI events each year for the NATL basin, showing that mean bias is significantly reduced with time from –18kt (2009) to –8 kt (2019), though RMS error curve is flat at approximately 20 kt.

A Statistical Analysis of High Frequency Track and Intensity Forecasts from NOAA's Operational Hurricane Weather Research and Forecast (HWRF) Modeling System

Zhan Zhang[1,2], Jun Zhang[3,4], Keqin Wu[1,2], Ghassan Alaka[3,] Avichal Mehra[2], Vijay Tallapragada[2]
[1]IMSG at EMC/NCEP/NWS/NOAA, College Park, MD 20740, [2]EMC/NCEP/NWS/NOAA, College Park, MD 20740, [3]NOAA/AOML/Hurricane Research Division, Miami, Florida, [4]University of Miami/CIMAS, Miami, Florida

Email: Zhan.Zhang@noaa.gov

1. Introduction

In addition to the tropical cycle (TC) track and intensity forecast guidance at 6-hourly synoptic times valid at 00Z, 06Z, 12Z, and 18Z provided in Automated Tropical Cyclone Forecasting (ATCF) format, the operational Hurricane Weather Research and Forecast **(**HWRF) model also provides high-frequency tropical cyclone forecast (HTCF) output at every model time step of the innermost domain (10/3 seconds). The variables in the HTCF output include magnitude and location (latitude/longitude) of 10-meter maximum wind speed (Vmax), minimum sea level pressure (Pmin), and radius of maximum wind (RMW). In this study, a statistical analysis is performed on the high-frequency output from the operational HWRF forecasts of track and intensity for all TCs in the North Atlantic basin for a 3-year period (2017-2019). The results show that there are large temporal fluctuations and uncertainties in the high-resolution TC track and intensity that is not captured by the conventional (six-hourly) forecast guidance provided to the TC forecast centers. Running means at various time windows are applied to the high frequency track and intensity forecast data from the model output to study their statistical characteristics. The analysis demonstrates for the first time that the operational HWRF model is capable of producing the high frequency trochoidal TC motion seen in observations. The TC track and intensity verification indicates that the +/- 3-hour and 4.5-hour running means of the high-frequency intensity outputs are ~5% more skillful than the standard 6-hourly HWRF intensity forecasts, while the skill for the track forecast is comparable between the two methods.

2. High-Frequency Track and Intensity Analysis

Track forecasts using the high-frequency internal tracker and the 6-hourly external tracker are compared with one and another and with observations for Hurricane Florence initialized at 0000 UTC 09 September 2018, Figure 1a. Observations include the observed best track and high-frequency (2-min) track observations. Fig. 1a represents typical characteristics of high-frequency HWRF track output and clearly shows the small-scale oscillation of trochoidal motion, which rotates counterclockwise around the 6-hourly TC tracks. Trochoidal TC motions have been previously observed in radar observations (Marks et al. 2008). Fig. 1b shows the temporal fluctuations of the TC intensity forecasts for Hurricane Florence. TC intensity is compared between the HWRF external tracker (every 6 h), the HWRF internal tracker at every model time step (10/3 s), plus 10-, 60-, and 360-minute running averages, and the best track.

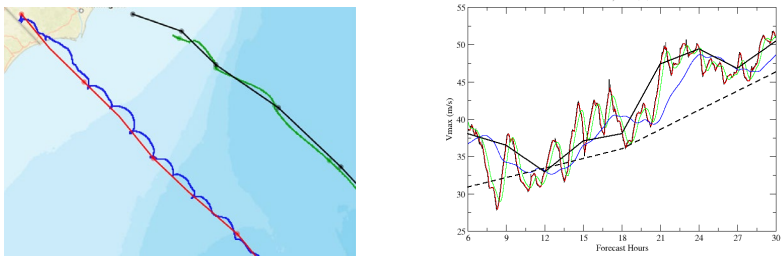

Fig. 1 a) HWRF track forecasts for Hurricane Florence 06L, 20180909 00UTC (left). Two forecast tracks and two observed tracks are displayed the operational HWRF track forecasts (red), HWRF high-frequency track (blue), observed. b). HWRF intensity forecasts for Hurricane Florence 06L, 20180909 00UTC (right) panel.

3. Verification

The track and intensity forecast skill are compared between the operational HWRF, and 6h- and 9h-running means of high frequency. It is found that track forecast skill was similar at all forecast lead times for the operational HWRF, 360M, and 540M (not shown). This result was expected because small temporal scale trochoidal motion uncertainties are removed by using sufficiently long running mean windows. At early forecast lead times, 540M shows a 3% track skill improvement over the operational HWRF. On the other hand, intensity verification shows quite different results. The intensity forecast skill for both 360M and 540M show at least a 3% improvement over the operational HWRF Vmax forecasts at all lead times.

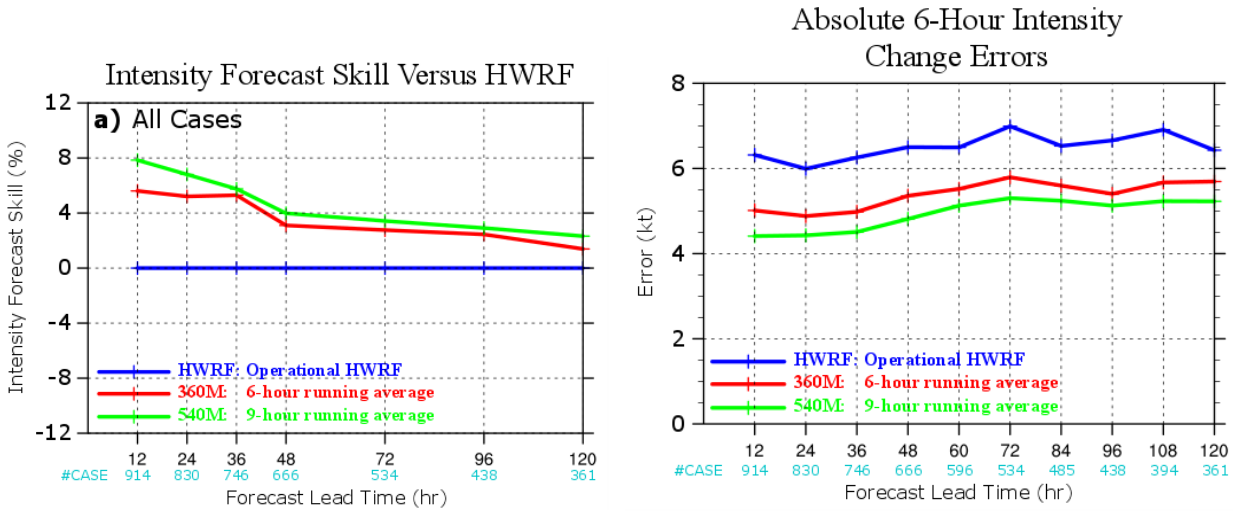

FIG. 2 Comparisons of intensity (left) and 6h intensity change (right) forecast skill for all cycles, the operational HWRF (HWRF, blue line), running mean of high frequency over +/-3-hour time window (360M, red line), and running mean of high frequency over +/-4.5 hour time window (540M, green line).

4. Conclusions

The analysis demonstrates for the first time that the operational HWRF is capable of producing the trochoidal TC motion in the high-frequency TC tracks seen in observations. The high frequency fluctuations of predicted TC tracks show trochoidal motion with a rotational period of ~1-hour in the temporal scale and ~20-100 kilometers in the spatial scale. The fluctuations that lead to large uncertainties in the 6-hourly model track and intensity forecast guidance are estimated. It is found that the uncertainties in model track forecasts are small enough to have no impact on the verification against the best track. On the other hand, the high frequency fluctuations of Vmax present some uncertainties in the 6-hourly intensity forecasts. Removing temporal intensity uncertainties results in a ~3-5% improvement in TC intensity forecasts, compared to the standard 6-hourly HWRF forecasts.

References:

Marks, F. D, P. G. Black, M.I T. Montgomery, Robert W. Burpee, 2008: Structure of the Eye and Eyewall of Hurricane Hugo (1989). *Mon. Wea. Rev., 136,* 1237-1259, *https://doi.org/10.1175/2007MWR2073.1*

Zhang Z., Vijay Tallapragada, C. Kieu, S. Trahan, and W. Wang, 2014: HWRF based Ensemble Prediction System Using Perturbations from GEFS and Stochastic Convective Trigger Function. Tropical cyclone Research and Review, Vol. 3, 145-161