NCEP NOTES

ᵊA Consensus Forecast for Tropical Cyclone Gale Wind Radii

CHARLES R. SAMPSON

Naval Research Laboratory, Monterey, California

JOHN A. KNAFF

NOAA/Center for Satellite Applications and Research, Fort Collins, Colorado

ABSTRACT

The National Hurricane Center (NHC) has been forecasting gale force wind radii for many years, and more recently (starting in 2004) began routine postanalysis or "best tracking" of the maximum radial extent of gale [34 knots (kt; $1 \, kt = 0.514 \, m \, s^{-1}$)] force winds in compass quadrants surrounding the tropical cyclone (wind radii). At approximately the same time, a statistical wind radii forecast, based solely on climatology and persistence, was implemented so that wind radii forecasts could be evaluated for skill. If the best-track gale radii are used as ground truth (even accounting for random errors in the analyses), the skill of the NHC forecasts appears to be improving at 2- and 3-day lead times, suggesting that the guidance has also improved. In this paper several NWP models are evaluated for their skill, an equally weighted average or "consensus" of the model forecasts is constructed, and finally the consensus skill is evaluated. The results are similar to what is found with tropical cyclone track and intensity in that the consensus skill is comparable to or better than that of the individual models. Furthermore, the consensus skill is high enough to be of potential use as forecast guidance or as a proxy for official gale force wind radii forecasts at the longer lead times.

1. Introduction

Surface winds associated with tropical cyclones (TCs; see Table 1 for a list of acronyms used in this note) are critical to many public, private, and governmental stakeholders. The National Hurricane Center (NHC) makes 6-hourly analyses and forecasts of TC tracks, intensities, and structures for all active TCs in the Atlantic and eastern North Pacific basins. Initial and forecast TC wind structures are provided in terms of the maximum radial extent of gale [34 knots (kt; $1 \, kt = 0.514 \, m \, s^{-1}$)], damaging (50 kt), and hurricane (64 kt) force winds in compass quadrants surrounding the TC. These are collectively referred to as wind radii. NHC forecasts

hurricane force wind radii through 36 h, damaging and gale force wind radii through 72 h, and intensity (1-min mean maximum wind speed near the center) and track through 120 h. These forecasts are used for the official NHC watch and warning decision process and are employed as inputs to other decision aids designed to estimate wind probabilities (DeMaria et al. 2013), storm surge (NHC 2015), wave forecasts (Sampson et al. 2010), infrastructure damages (e.g., Quiring et al. 2014), Department of Defense conditions of readiness (Sampson et al. 2012), etc.

To provide an evaluation database for wind radii forecasts, NHC began best tracking wind radii in 2004 and runs a purely statistical model based on climatology and persistence or the radii-CLIPER (DRCL; Knaff et al. 2007) for every forecast. These best tracks and forecasts are saved in the databases of the Automated Tropical Cyclone Forecast System (ATCF; Sampson and Schrader 2000). The best track provides ground truth for forecasts while the DRCL provides a baseline forecast from which skill can be determined. The

---

ᵊ Denotes Open Access content.

Corresponding author address: Charles R. Sampson, NRL, 7 Grace Hopper Ave., Stop 2, Monterey, CA 93943-5502.
E-mail: sampson@nrlmry.navy.mil

TABLE 1. List of acronyms and descriptions.

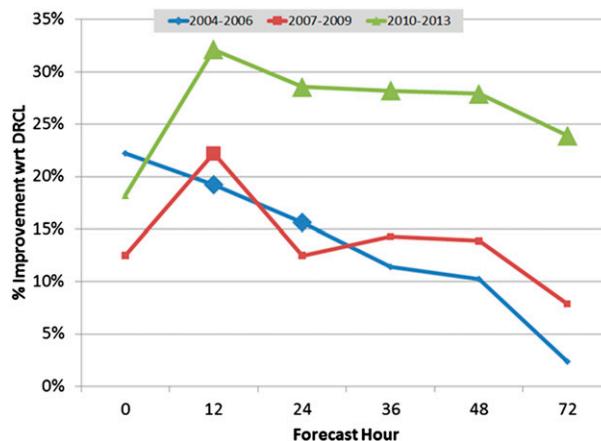|	
AHNI	GFS model radii with bias correction phased out at 36 h or removed entirely
AVNI	GFS model radii, bias corrected
DRCL	Wind radii climatology and persistence model used for investigating skill
ECMWF	European Centre for Medium-Range Weather Forecasts
EHXI	ECMWF model radii with bias correction removed
EMXI	ECMWF model radii, bias corrected
FAR	False alarm rate
GHTI	GFDL model radii with bias correction phased out at 36 h or removed entirely
GFTI	GFDL model radii, bias corrected
HHFI	HWRF model radii with bias correction phased out at 36 h or removed entirely
HWFI	HWRF model radii, bias corrected
MAE	Mean absolute error
NHC	National Hurricane Center
OFCL	Official forecasts issued by the NHC
R34	Radii of 34-kt ($15\,\mathrm{m\,s^{-1}}$) winds; also known as gale force wind radii
RVCN	R34 consensus forecast
TC	Tropical cyclone

FIG. 1. Percent improvement (skill) of MAE with respect to DRCL forecasts for the periods 2004–06 (blue), 2007–09 (red), and 2010–13 (green). Statistical significance, accounting for 30-h serial correlation, is indicated by the larger line markers [after Knaff and Sampson (2015)].

accuracy of the wind radii in the best tracks, which are estimated to have errors as high as 10%–40%, are discussed in greater detail in both Knaff and Harper (2010) and Knaff and Sampson (2015).

Despite the forecast requirements, no skillful radii tool or model aids existed as recently as 2005 and the only skillful forecast came from the NHC whose 2005 Atlantic gale force wind radii forecasts were more skillful than those of DRCL out to 36 h (Knaff et al. 2006). However, more recent evaluations of NHC's gale force wind radii forecasts revealed that gale force wind radii forecasts have improved quite dramatically in the past four years and that forecasts are now skillful (better than DRCL) through 72 h (Knaff and Sampson 2015), as shown in Fig. 1. In addition, Cangialosi and Landsea (2014) showed that several forecast aids or models provided skillful forecasts of gale force winds when compared with the highest quality wind radii estimates (aircraft co-incident). These studies taken collectively imply 1) that forecast aids of gale force wind radii may now possess skill beyond 36 h and 2) that forecasters may have used these aids to improve their forecasts in the last four years.

With this evidence, it is now time to explore whether a gale force wind radii forecast consensus could provide increased skill relative to its members, similar to results found for track (Goerss et al. 2004) and intensity (Sampson et al. 2008). First, gale force wind radii guidance from four NWP models discussed in Cangialosi and Landsea (2014) will be reevaluated. Then, similar to what was done in Goerss et al. (2004) and Sampson et al.

(2008), we will construct a consensus and evaluate it against the input NWP models. Our evaluation is done in terms of mean absolute errors, bias, probability of detection, and false detection (false alarms). These results are presented followed by conclusions and recommendations related to consensus forecasts of gale force wind radii.

2. Data and methods

The verification of maximum extent of gale force winds (R34) is based on best-track data and operational forecasts made by DRCL during the period 2012–14 (2014 is pending final analysis and approval). The R34 is estimated and forecasted in compass quadrants (northeast, southeast, southwest, and northwest) surrounding TCs that have intensities of 34 kt ($17\,\mathrm{m\,s^{-1}}$) or greater. As stated above, NHC makes official (OFCL) forecasts of intensity and track through 120 h and OFCL forecasts of R34 through 72 h. The input data for each DRCL forecast is the corresponding OFCL track and intensity forecast and results in DRCL forecasts being available for all OFCL forecasts, which ensures that a fair baseline forecast can be constructed.

This study will concentrate on R34 verification statistics since R34 is likely best observed or estimated because of the larger spatial coverage area and the availability of more platforms suited to observe these winds (e.g., ships, buoys, land stations, scatterometer winds, etc.). The authors are keenly aware of the R34 quality and dependency issues noted in Knaff and Sampson (2015) and Knaff et al. (2006), as well as those identified by forecasters at NHC, but no special provisions are made here to account for those errors. The
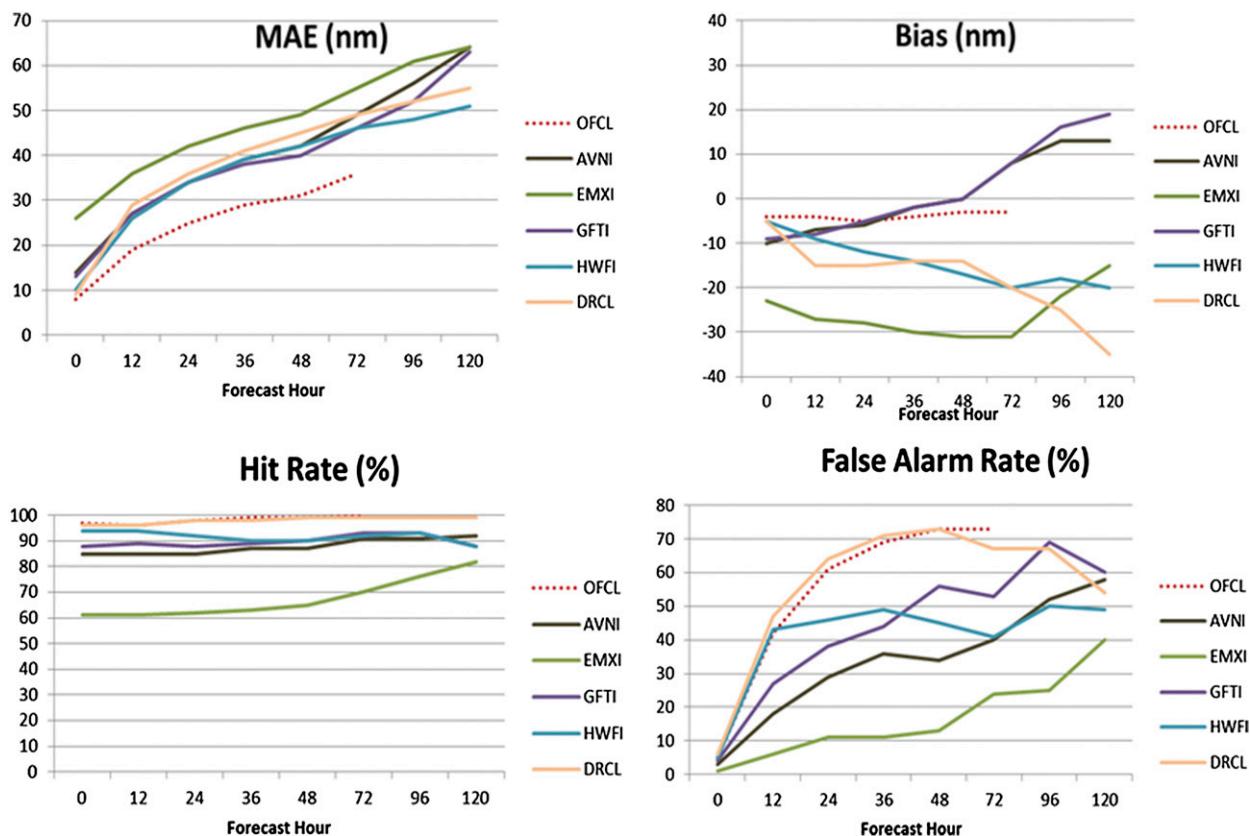
FIG. 2. For the Atlantic basin during 2012–14, the (top left) R34 MAE of individual NWP model aids and OFCL, where all aid R34 forecasts are bias corrected to match current analysis; (top right) R34 mean forecast bias; (bottom left) hit rate; and (bottom right) FAR. The sample is homogeneous and the numbers of forecasts are 2580, 2404, 2136, 1848, 1616, 1192, 896, and 628 at 0, 12, 24, 36, 48, 72, 96, and 120 h, respectively.

dataset used in this study is contained in the ATCF databases and is freely available from NHC.

To calculate verification statistics, forecast values of R34 in each quadrant and at each forecast lead time are compared to the final best-track values. The occurrence of zero-valued wind radii introduces an added complication when verifying wind radii. The zero-valued wind radii typically occur when storms are near the 34-kt intensity or when storm translation speeds are large (i.e., $>8\,\mathrm{m\,s^{-1}}$). For this study, the following verification strategy is adopted. If any of the quadrants in the best track have nonzero wind radii, all quadrants for that case are verified. This strategy allows the individual quadrant statistics to be combined to form a single measurement of mean absolute error and bias for each forecast lead time and also results in an approximately 20%–25% increase in the number of cases. Since the forecast of R34 is in units of nautical miles (n mi; 1 n mi = 1.852 km) and of intensity in units of knots, these units will be used throughout.

To evaluate the ability of the forecasts to discriminate the occurrence of R34 and to complement the MAE and bias statistics, the probability of detection or "hit rate"

and probability of false alarm or "false alarm rate" are also presented. The hit rate and false alarm rate are based on whether or not a quadrant had a nonzero wind radii value.

To keep the verification brief, we present the statistics for combined quadrants (i.e., all the errors in the different quadrants are averaged). We also only evaluate tropical cyclones (e.g., no subtropical, extratropical, or posttropical cases). Errors are also calculated in homogeneous sets (i.e., they all include the same cases). Statistical significance discussed in this paper is assessed using a Student's t test assuming two tails and the 95% level with serial correlation removed (see Leith 1973). The results of these analyses will be presented in the next section.

3. Results

Figure 2 shows an evaluation of a 3-yr sample (2012–14) of DRCL, OFCL, and four NWP model R34 forecasts for the Atlantic basin. Since the NWP model forecasts are considered "late models" by the

operational centers (i.e., their forecasts are not available until approximately 6 h after the initial time), they are "interpolated" for 6 or 12 h (Goerss et al. 2004) to produce guidance that is relabeled at the current time. The wind radii are bias corrected so that the initial wind radii match the current analysis, and this bias correction is applied at all forecast hours. When the interpolation software was written, this appeared to be a reasonable way to process wind radii guidance that was not skillful; however, some of the interpolated NWP model aids now show skill relative to DRCL, as seen in Fig. 2.

At this point it is worth mentioning that a consequence of using zero-valued wind radius forecasts in our evaluation is higher mean errors and generally more negative biases, especially for the European Centre for Medium-Range Weather Forecasts (ECMWF) model aid (EMXI). The hit rate of the EMXI is very low relative to the other aids, so it is penalized in the MAE (higher) and the bias (more negative). Another way to do the evaluation is to only evaluate each wind radius when both the forecast radius and verification radius are both nonzero. Evaluation of our dataset this way (not shown) reduces the MAE of EMXI and three other NWP model aids below that of DRCL, though differences between the NWP model MAE and that of DRCL are not significant. The evaluation where cases with zero-valued wind radius forecasts are removed also shows that the NWP model aid biases generally move toward zero, especially for the EMXI. The NWP model aid biases are also generally closer to zero than DRCL. DRCL performs well in the evaluation where we count the zeros because it is designed to predict an R34 anytime the intensity is 35 kt or greater. Finally, we can see that OFCL has the lowest errors, near-zero bias, and a nearly 100% hit rate (a desirable feature for some applications and end users). The false alarm rates for most of the aids and OFCL are high, but these only represent on the order of 10% of the forecasts and the authors consider this issue to be less detrimental than a low hit rate.

As described above in the results presented in Fig. 2, the NWP model aids are all bias corrected to the current analysis via interpolation. But is the bias correction really adding value to the forecast and at what time does the bias correction stop improving the wind radii forecast? Figure 3 shows the effects of the bias correction to the analyzed radii on other forecast lengths. We did this by adjusting the forecast to the current time with and without the bias correction. The aids with "H" as the second character indicate that the bias correction was removed when the adjustment to the current time was made. The first thing to note is that the bias correction seems to have a positive impact on the ECMWF aid
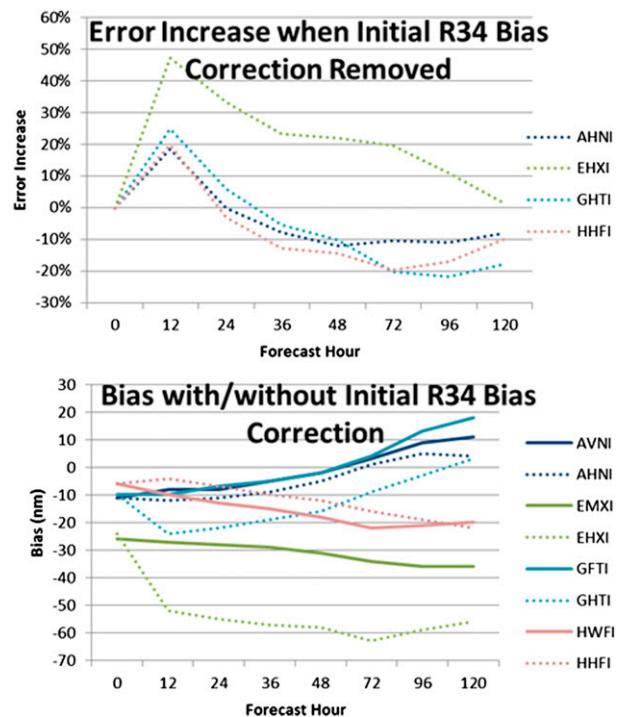

FIG. 3. (top) Forecast MAE increase from removing R34 bias correction from the interpolator. (bottom) Mean forecast bias for aids with (solid) and without (dotted) R34 bias correction. The sample is homogeneous from the Atlantic during 2012–14 and the numbers of forecasts are 2696, 2532, 2268, 2008, 1732, 1272, 932, and 668 at 0, 12, 24, 36, 48, 72, 96, and 120 h, respectively.

(EMXI vs EHXI), reducing MAE through 96 h and bias through 120 h. For the other three NWP models, the results are mixed. The MAE is reduced only to about 24 h and the effects on bias are mixed, but one could possibly remove the bias correction at about 36–48 h without ill effect. This result is notably similar to the 32-h persistence phase-out period in the DRCL model (Knaff et al. 2007).

For the current effort we will now prescribe a linear phase out of the bias correction between 12 and 36 h for all NWP aids except the ECMWF aid (for which we apply the bias correction out to 120 h since that appears to reduce the MAE), and we then compute the R34 consensus forecasts for each quadrant (the average of nonzero radii available in each quadrant). For comparison we also computed a consensus with all of the guidance receiving bias correction at all forecast times, but that consensus underperforms (not significantly) our consensus with the phase outs, so we do not include its results.

Figure 4 shows an evaluation of the consensus (RVCN) and its members (all but the ECMWF with phase outs of the bias correction) for the Atlantic 2012–14 dataset. For comparison, OFCL is also included. As
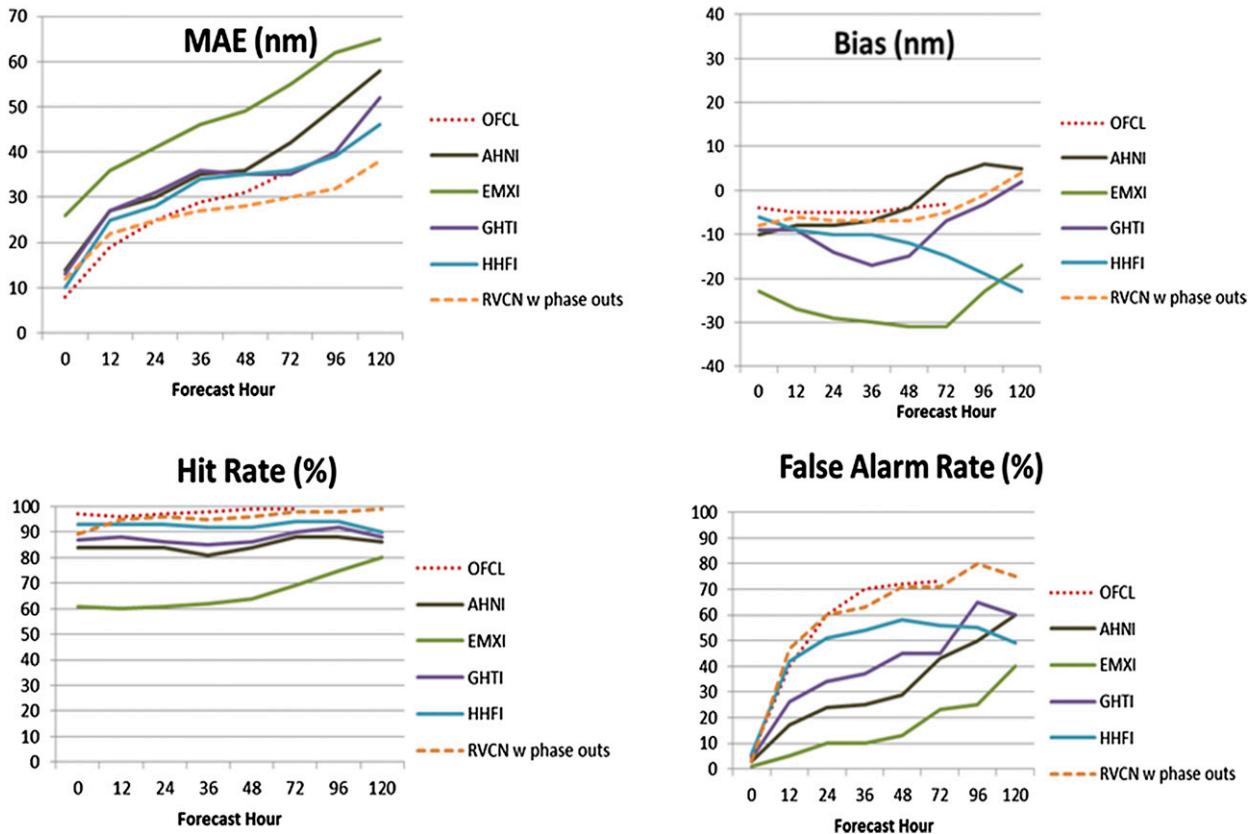
FIG. 4. For the Atlantic basin during 2012–14, the (top left) R34 MAE of individual NWP model aids and OFCL, (top right) R34 mean bias, (bottom left) hit rate, and (bottom right) FAR. The homogeneous set is considered to be dependent. The numbers of forecasts are 2628, 2436, 2172, 1892, 1644, 1212, 904, and 636 at 0, 12, 24, 36, 48, 72, 96, and 120 h, respectively.

seen in consensus studies of track and intensity, the consensus ranks among the leaders in MAE performance (in this case it is the top aid). Also, the consensus MAE is significantly less than that of DRCL (not included in Fig. 4) out to 72 h. The consensus has reasonable bias (slightly negative is acceptable since we are evaluating the zero forecasts), a very high hit rate, and a high false alarm rate. For forecasting and downstream algorithms we will assume that a low hit rate is probably more detrimental than a high false alarm rate since it is better to have the guidance available when R34 is not going to verify than to have no guidance when R34 verifies. Again, the less skillful EMXI performance in MAE and bias is largely a function of its low hit rate.

Finally, we could not consider the Atlantic evaluation truly independent since the development was done on this dataset. We made an effort not to tune or weight the guidance since that fits our consensus philosophy as well as that of many others (e.g., Kharin and Zwiers 2002; Weigel et al. 2010; DelSole et al. 2013), but there is no substitute for completely independent data and so we chose the eastern North Pacific during 2012–14 as our

test data. There is some risk in this since the Atlantic TC R34 climatology is on average about a third larger than those in the eastern North Pacific (Knaff et al. 2007), but we at least have the same NWP models available. In addition, the eastern North Pacific is the only basin other than the Atlantic for which postseason R34 reanalysis is performed, which further limits our options. Figure 5 shows the results of the evaluation with 2012–14 eastern North Pacific data. The results are surprisingly similar to those in the Atlantic, especially considering the differences in climatology. The MAEs are obviously smaller for this basin, which produces generally smaller tropical cyclones (Knaff et al. 2014). The consensus is still among the best performers among all of the metrics save the false alarm rate; however, the consensus MAE is not significantly better than that of DRCL at any forecast time. Still, this independent verification demonstrates that we can construct a consensus radii forecast with reasonable performance from a defined set of moderately skillful aids in one basin, then apply it to a basin with different climatology, and still get fairly consistent results.
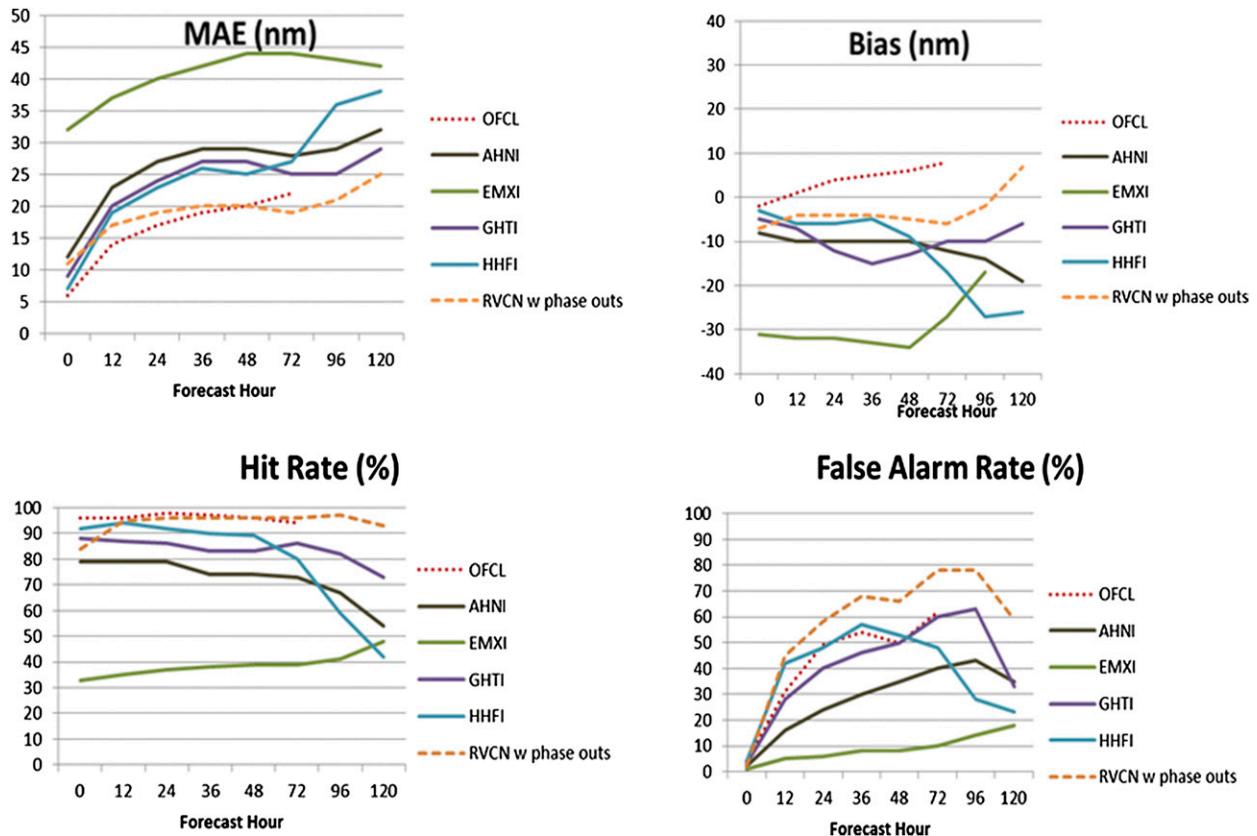
FIG. 5. As in Fig. 4, but for the eastern North Pacific. The homogeneous set is considered to be independent. The numbers of forecasts are 2096, 1944, 1684, 1420, 1168, 767, 464, and 252 at 0, 12, 24, 36, 48, 72, 96, and 120 h, respectively.

4. Conclusions and recommendations

The results presented above indicate that R34 forecasts from NWP models are competitive with DRCL, especially in the Atlantic. An equally weighted consensus forecast has been constructed and is proposed as forecast guidance and for potential use in other applications. The proposed R34 consensus extends to 120 h and its error characteristics are similar to those of the NHC official forecast in that it does not suffer from large negative biases like the DRCL model in the Atlantic. Since DRCL and other proxies for NHC radii are used in applications requiring wind radii that extend beyond 72 h [e.g., the wind probabilities of DeMaria et al. (2013) and the wave forecasts in Sampson et al. (2010)], the authors believe that the new R34 wind radii consensus could be explored as possible replacements for these proxies.

Finally, there is some debate about whether the gale force wind radii in the best tracks can serve as ground truth for evaluation because of concerns with sparse, intermittent, and poor quality observations. Cangialosi and Landsea (2014) attempted to address this using only the highest quality best-track data, and found similar

skill to that of our work. Knaff and Sampson (2015) ran some experiments introducing random error to the best track, and found that both the aid and official forecast skill signals remain in the evaluation. The R34 consensus skill in the Atlantic is also probably real and we now have an algorithm for making R34 forecasts with performance characteristics comparable to those of OFCL and that extends to 120 h. At a minimum, the longer-lead forecasts can provide extra information that can be leveraged for forecasting and other applications, which may prove to be critical to some public, private, and governmental stakeholders.

tropical cyclone intensity forecasting. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. government position, policy, or decision.

REFERENCES

Cangialosi, J. P., and C. W. Landsea, 2014: National Hurricane Center forecast wind radii verification. *Proc. 31st Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., 56. [Available online at https://ams.confex.com/ams/31Hurr/webprogram/Paper244740.html.]

DelSole, T., X. Yang, and M. K. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.*, **139**, 176–183, doi:10.1002/qj.1961.

DeMaria, M., and Coauthors, 2013: Improvements to the operational tropical cyclone wind speed probability model. *Wea. Forecasting*, **28**, 586–602, doi:10.1175/WAF-D-12-00116.1.

Goerss, J., C. Sampson, and J. Gross, 2004: A history of western North Pacific tropical cyclone track forecast skill. *Wea. Forecasting*, **19**, 633–638, doi:10.1175/1520-0434(2004)019<0633:AHOWNP>2.0.CO;2.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15**, 793–799, doi:10.1175/1520-0442(2002)015<0793:CPWME>2.0.CO;2.

Knaff, J. A., and B. A. Harper, 2010: Tropical cyclone surface wind structure and wind–pressure relationships. *Proc. WMO Int. Workshop on Tropical Cyclones—VII*, La Reunion, France, WMO, KN1–KN35. [Available online at http://www.wmo.int/pages/prog/arep/wwrp/tmr/otherfileformats/documents/KN1.pdf.]

——, and C. R. Sampson, 2015: After a decade are Atlantic tropical cyclone gale force wind radii forecasts now skillful? *Wea. Forecasting*, **30**, 702–709, doi:10.1175/WAF-D-14-00149.1.

——, C. Guard, J. Kossin, T. Marchok, C. Sampson, T. Smith, and N. Surgi, 2006: Operational guidance and skill in forecasting structure change. *Proc. WMO Int. Workshop on Tropical Cyclones—VI*, San Juan, Costa Rica, WMO, 160–184. [Available online at http://severe.worldweather.org/iwtc/document/Topic_1_5_John_Knaff.pdf.]

——, C. R. Sampson, M. DeMaria, T. P. Marchok, J. M. Gross, and C. J. McAdie, 2007: Statistical tropical cyclone wind radii prediction using climatology and persistence. *Wea. Forecasting*, **22**, 781–791, doi:10.1175/WAF1026.1.

——, S. P. Longmore, and D. A. Molenar, 2014: An objective satellite-based tropical cyclone size climatology. *J. Climate*, **27**, 455–476, doi:10.1175/JCLI-D-13-00096.1.

Leith, C. E., 1973: The standard error of time-average estimates of climate means. *J. Appl. Meteor.*, **12**, 1066–1069, doi:10.1175/1520-0450(1973)012<1066:TSEOTA>2.0.CO;2.

NHC, 2015, Introduction to storm surge. National Hurricane Center/Storm Surge Unit, 5 pp. [Available online at http://www.nhc.noaa.gov/surge/surge_intro.pdf.]

Quiring, S., A. Schumacher, and S. Guikema, 2014: Incorporating hurricane forecast uncertainty into decision support applications. *Bull. Amer. Meteor. Soc.*, **95**, 47–58, doi:10.1175/BAMS-D-12-00012.1.

Sampson, C. R., and A. J. Schrader, 2000: The Automated Tropical Cyclone Forecasting system (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, doi:10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2.

——, J. L. Franklin, J. A. Knaff, and M. DeMaria, 2008: Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecasting*, **23**, 304–312, doi:10.1175/2007WAF2007028.1.

——, P. A. Wittmann, and H. L. Tolman, 2010: Consistent tropical cyclone wind and wave forecasts for the U.S. Navy. *Wea. Forecasting*, **25**, 1293–1306, doi:10.1175/2010WAF2222376.1.

——, and Coauthors, 2012: Objective guidance for use in setting tropical cyclone conditions of readiness. *Wea. Forecasting*, **27**, 1052–1060, doi:10.1175/WAF-D-12-00008.1.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *J. Climate*, **23**, 4175–4191, doi:10.1175/2010JCLI3594.1.