Evaluating Medium-Range Tropical Cyclone Forecasts in Uniform- and Variable-Resolution Global Models

CHRISTOPHER A. DAVIS, DAVID A. AHIJEVYCH, WEI WANG, AND WILLIAM C. SKAMAROCK

National Center for Atmospheric Research,[a] *Boulder, Colorado*

(Manuscript received 12 January 2016, in final form 19 July 2016)

ABSTRACT

An evaluation of medium-range forecasts of tropical cyclones (TCs) is performed, covering the eastern North Pacific basin during the period 1 August–3 November 2014. Real-time forecasts from the Model for Prediction Across Scales (MPAS) and operational forecasts from the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) are evaluated. A new TC-verification method is introduced that treats TC tracks as objects. The method identifies matching pairs of forecast and observed tracks, missed and false alarm tracks, and derives statistics using a multicategory contingency table methodology. The formalism includes track, intensity, and genesis.

Two configurations of MPAS, a uniform 15-km mesh and a variable-resolution mesh transitioning from 60 km globally to 15 km over the eastern Pacific, are compared with each other and with the operational GFS. The two configurations of MPAS reveal highly similar forecast skill and biases through at least day 7. This result supports the effectiveness of TC prediction using variable resolution.

Both MPAS and the GFS suffer from biases in predictions of genesis at longer time ranges; MPAS produces too many storms whereas the GFS produces too few. MPAS better discriminates hurricanes than does the GFS, but the false alarms in MPAS lower overall forecast skill in the medium range relative to GFS. The biases in MPAS forecasts are traced to errors in the parameterization of shallow convection south of the equator and the resulting erroneous invigoration of the ITCZ over the eastern North Pacific.

1. Introduction

Tropical cyclone (TC) prediction is a societally important problem for which there is increasing emphasis on medium-range forecasts (Gall et al. 2013; Yamaguchi et al. 2015). The challenge is to resolve the finescale aspects of the tropical cyclone while properly simulating the environment within which it occurs. Global models with variable horizontal resolution have recently been explored as a way of providing regional high resolution without lateral boundaries that can compromise the quality and interpretation of medium-range prediction skill (Park et al. 2013). Global, variable-resolution TC prediction was explored by Zarzycki and Jablonowski (2015) using a hydrostatic model. Ultimately, nonhydrostatic models are needed to resolve the inner-core aspects of tropical cyclones.

The Model for Prediction Across Scales (MPAS) is a nonhydrostatic global model designed for mesoscale and convective-scale weather and climate prediction research (Skamarock et al. 2012, 2014; Klemp et al. 2015). One of the main features of MPAS is its use of an unstructured Voronoi mesh, on which the grid spacing can vary smoothly in space. This avoids some of the problems with traditional nesting approaches (Park et al. 2014; Hashimoto et al. 2016). The existence of two meshes, one uniform, the other variable, within the same global model, allows us to consider questions related to prediction skill in the variable-resolution configuration relative to uniform resolution. It is clear that, over time, the effects of the presumably less accurate prediction on the coarse part of the variable mesh will influence the fine-mesh region. An important question is how quickly this will occur, especially in the tropics where influences from the coarse-mesh region will affect TC forecasts.

The present paper addresses two related issues in medium-range prediction of TCs and their environment.

[a] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Christopher A. Davis, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307.
E-mail: cdavis@ucar.edu

DOI: 10.1175/MWR-D-16-0021.1

The first issue concerns how variable- and uniform-resolution forecasts compare to each other out to a lead time of 10 days. The hypothesis is that variable resolution provides a computationally economical strategy for examining tropical cyclones in a particular basin. The viability of variable resolution requires that the relatively coarse resolution over the remainder of the globe does not produce differences that compromise the quality of forecasts within the high-resolution region.

The second issue is the current absence of a verification strategy appropriate for medium-range forecasts of TCs. Most storms present at days 7–10 are not present in the initial condition. Existing mainstream verification approaches[1] do not account for errors due to the erroneous formation of storms during the forecast, or missed events, and do not consider the consequences of errors in the timing of TC-track initiation and dissipation. The present paper, therefore, devotes substantial material to the development of a verification methodology appropriate for extended-range TC forecasts. This verification methodology is then applied to the different MPAS configurations, and to the operational (hydrostatic) Global Forecast System (GFS), to discern important facets of medium-range TC forecast skill and the biases that reduce skill.

The essential consideration for verification is viewing TC tracks as objects whose properties are evaluated. While the matching of TC tracks is fairly trivial for storms already present at model initialization time, the case is far more complicated for tracks that begin during the forecast. Halperin et al. (2013) provide a recent study of genesis prediction skill for the Atlantic, whereas Chan and Kwok (1999) and Elsberry et al. (2009) examine genesis prediction over the northwest Pacific. The present study is made novel by focusing on the eastern North Pacific and, more importantly, the inclusion of TC formation as part of a more comprehensive evaluation framework.

The finest MPAS mesh spacing considered herein is 15 km, which is comparable to current operational global model grid spacing, although less than the operational GFS in 2014. Accurate prediction of TC intensity and structure is potentially compromised at this grid spacing, especially in cases with a small radius of maximum wind. The eastern Pacific basin, the focus of this study, harbors some of the smallest tropical cyclones observed (Knaff et al. 2014).

To provide a context for the verification results, MPAS forecasts are compared with the TC forecasts from the operational GFS model from the National Centers for Environmental Prediction (NCEP).[2] Because both MPAS and the GFS have evolved substantially since 2014, the results herein should not be viewed as a comparison of current model capabilities. The two models have different dynamics, different physical parameterizations, different abilities to resolve TCs, and different biases in the overall tropical atmosphere that lead to some important differences at longer lead times. A significant part of the paper is devoted to understanding how model bias affects the quality of TC forecasts, and to diagnosing the causes of such biases.

Because the verification methodology requires a significant amount of text to explain, we include the details of our track computation, track matching, and skill-score computation in the appendix. Section 3 presents results that are at least partly grounded in operational standards, but which have new components that arise from our verification approach. The conclusions, in section 4, reinforce the need for more comprehensive verification strategies, and support the efficacy of variable-resolution modeling of tropical cyclones in global models.

2. Methodology

a. *Model configuration*

The geographical area of interest is the eastern Pacific basin for the 95-day period of 1 August–3 November 2014. There were 14 tropical cyclones in the eastern Pacific basin during this period, 10 of which had a life cycle fully contained within the evaluation period with an average duration of approximately 9 days. There were six major hurricanes during the period. Hurricane Ana, which formed in the central Pacific, was also included in the sample of storms. Between 1 August and 7 October, there were only 4 days without an active tropical cyclone in the eastern Pacific basin.

For the simulations considered herein, MPAS was configured with a nearly uniform mesh of 15-km centroid spacing, and a variable mesh (Fig. 1) with cell-center spacing that smoothly stretched from 15 to 60 km, with the higher resolution concentrated over the eastern and central Pacific. The variable-resolution configuration is denoted MPAS-EP. In the variable-resolution mesh, roughly 53% of the cells have a mean diameter

[1] For instance, the position and intensity verification statistics computed by the National Hurricane Center of the National Oceanic and Atmospheric Administration consider only storms that exist at model initialization time.

[2] In 2015 the NCEP GFS model was upgraded to a grid spacing of roughly 13 km whereas the 2014 version of the model, examined herein, was roughly twice as coarse.

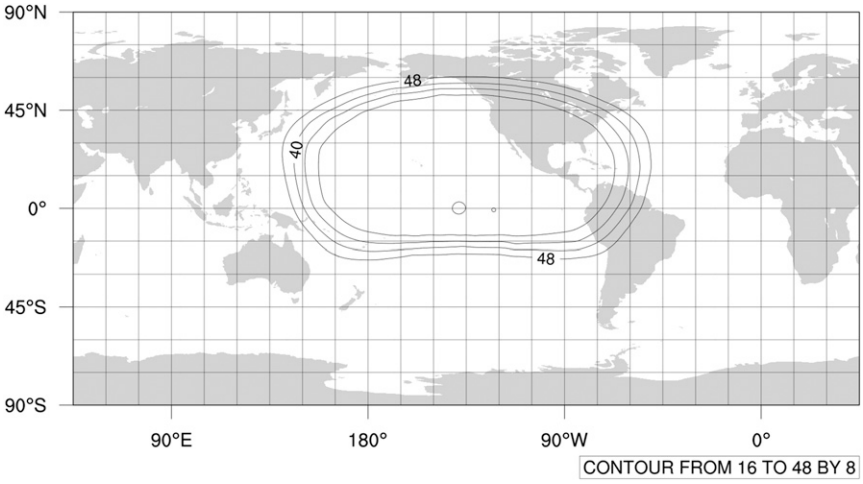


FIG. 1. Centroid spacing for variable-resolution MPAS mesh. Contour interval is 8 km.

less than 16 km. Simulations using the variable mesh require roughly 20% of the computational cost of simulations using the uniform mesh. There is a modest overhead in running the variable-resolution configuration because the time step is constrained by 15-km portion of the mesh.

A total of 95 daily forecasts were integrated out to 10 days during the continuous period 1 August–3 November 2014. MPAS was initialized at 0000 UTC by straight interpolation of the NCEP GFS initial condition, obtained on the native GFS vertical coordinate, to the MPAS grid. The model top was placed at 30 km and a total of 55 levels were positioned as close to the levels of the operational GFS as possible. The operational GFS[3] uses 64 levels, but 9 are above 30 km.

The MPAS physical parameterizations include a version of the Tiedtke cumulus parameterization, the Yonsei University (YSU) planetary boundary layer (PBL) scheme (Noh et al. 2003; Coniglio et al. 2013), the RRTMG radiation scheme, the WSM6 cloud physics scheme, and a simple ocean mixed layer scheme (Pollard et al. 1973; Davis et al. 2008). The YSU, RRTMG, and WSM6 schemes are the same as used in the Weather Research and Forecasting (WRF) Model, version 3.7. The Charnock formulation is used for surface drag, and the Coupled Ocean–Atmosphere Response Experiment (COARE) formulation is used for heat and moisture fluxes over the ocean (Fairall et al. 2003). The Tiedtke scheme used here is based on the modified version of the scheme from Zhang et al. (2011). There are a few changes made for this work (R. Torn 2014, personal communication). The closure for shallow convection is changed to depend on subcloud layer moist static energy following ECMWF Cy37r2, rather than moisture convergence. The trigger function takes into account advection, PBL, and radiation tendencies, which leads to more realistic convection initiation over land. The entrainment rate for shallow convection is reduced from 0.0012 to 0.0006, and this makes shallow convection a bit more active. The convective adjustment time scale is reduced from 1 h to 40 min. It is worth noting that these parameters were based on forecast results for the tropical Atlantic.

b. *Verification*

Traditionally, only storms that are tracked from the initialization time, as a pair of corresponding forecast and observed features, are evaluated. However, for the eastern North Pacific forecasts considered herein, only in 12% of all 10-day forecast periods is a storm present in the initial condition still tracked at day 10. Without considering storms forming during the forecast period, a major component of forecast model evaluation is neglected. As will be shown, systematic errors in the prediction of TC occurrence point the way to further diagnosis to understand model biases that can become large in medium-range forecasts.

Using 6-hourly output from both versions of MPAS and from the GFS, we tracked all tropical storms and hurricanes, even those that developed during the forecast, using essentially the same tracker that is used in operations [the so-called Geophysical Fluid Dynamics Laboratory (GFDL) tracker; see the appendix for details]. We then assessed which tracks in MPAS and the GFS corresponded to tracks of observed tropical

[3] Details about the physical parameterizations in the operational GFS may be found online at http://www.emc.ncep.noaa.gov/GFS/doc.php.

cyclones. The tracks of observed tropical cyclones came from the Tropical Cyclone Guidance Project (Vigh 2015), which maintains an archive of operational "working best tracks." These are not the official post-season best tracks but they have the benefit of including the invest stage before TC genesis. By focusing on tracks, we considered the full storm history in deciding whether a forecast storm corresponded to an observed storm. In addition to considering tracks that began during the forecast period, we considered forecast and observed tracks with differing end points in time, as well as unmatched forecast and observed tracks. This allowed us to assess a full set of statistics of missed events and false alarms. Details of the track matching strategy and verification methodology appear in the appendix.

The outcome of the verification methodology is a Heidke skill score S (Doswell et al. 1990) computed from a 3×3 contingency table. The three event categories—(i) no storm, (ii) weak storm, and (iii) strong storm—allow us to evaluate not just the dichotomous case involving the existence of a storm, but also allow us to evaluate forecast quality relative to a threshold of maximum wind speed V for matching forecast and observed tracks.

While the tracks of storms were obtained from 6-hourly model output fields on a 0.5° grid in latitude and longitude, MPAS intensity was defined as the maximum wind at any point (within 200 km of the cyclone center) on the original unstructured mesh. For the GFS, intensity was taken from the tracker. We compared the maximum GFS wind from the official a-deck files with the maximum wind obtained from the tracker and found only minor intensity differences, as expected. Given the 0.5° grid, we expect that some GFS maximum winds are biased low relative to the winds on the native 27-km grid.

Position and intensity errors were evaluated for all matched (forecast, observed) pairs, including pairs along tracks that began after the model initialization time. We also computed position and intensity error statistics by running the tracker in so-called "deterministic mode." In this mode, only storms that existed at model initialization time were evaluated. This resulted in far fewer samples at lead times beyond 4–5 days than were obtained by including tracks that began during the forecast period. However, this approach is consistent with current operational forecast evaluation procedures.

The software used to compute TC statistics for matched pairs is the Model Evaluation Tools for Tropical Cyclones (MET-TC). This code accepts forecast–observation paired data and computes a variety of standard statistics (mean error, median error, mean absolute error, root-mean-squared error) and produces box-and-whisker graphical representations of error distributions.

3. Results

a. Position and intensity verification

The present section provides some statistics of TC forecasts where there is a corresponding pair of forecast and observed storms along a matched pair of tracks. In deterministic mode, tracks all begin at model initialization time ($t = 0$), whereas in genesis mode, tracks may begin throughout the forecast period. A total of 221 forecast tracks matched the track of an observed storm from the 95 MPAS forecasts. Roughly half of these began after $t = 0$.

For matched tracks, the distribution of position error was computed for all common times along the forecast and observed tracks. The results are presented in Fig. 2 using box-and-whisker plots. There is no statistically significant difference between the position errors of the GFS and MPAS, or between either configuration of MPAS. Careful comparison of Figs. 2a and 2b indicates that the GFS results at day 8 are somewhat better for storms tracked from $t = 0$ (Fig. 2a), but that the small sample size compromises the statistical significance of the result. When genesis cases are included (Fig. 2b), the errors in the three sets of forecasts are indistinguishable.

While we expect differences between the MPAS and GFS forecasts to be small in the short range (because MPAS is initialized with the GFS), the fact that differences are small through day 8 reflects the importance of larger-scale (synoptic-scale) motions on track. In addition, the difference between uniform and variable-resolution MPAS performance is nearly imperceptible through day 8. This demonstrates that variable resolution allows comparable forecast performance for a given basin for a small fraction of the cost of globally uniform resolution.

To evaluate intensity errors, we focus on intensity biases and intensity distributions rather than on deterministic intensity forecast skill because predictability of intensity is rather limited, especially at the long lead times considered herein. Biases provide information about systematic errors in physical processes and about the limitations of resolution and the effects of dissipation. We characterize the distributions by their median values (Fig. 3), rather than the arithmetic means. Over the eastern North Pacific, both configurations of MPAS exhibit similarly small median intensity biases, whereas the GFS median bias is about 10 kt ($1\ \text{kt} = 0.5144\ \text{m s}^{-1}$) weaker than observed. This result is quantitatively similar whether the evaluation is confined to storms that

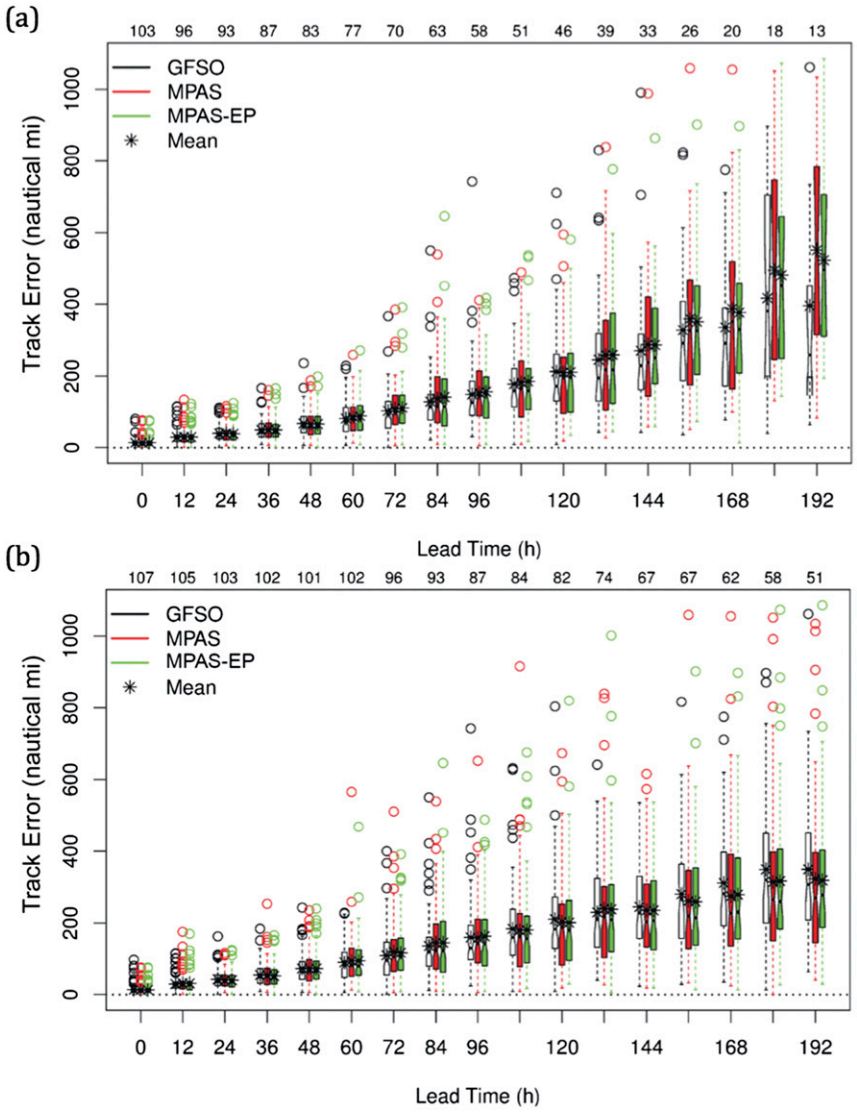


FIG. 2. (a),(b) Distributions of forecast TC position errors for each lead time out to 192 h, for a homogeneous sample of forecasts between 1 Aug and 3 Nov 2014, for the eastern North Pacific basin. Units are nautical miles (n mi; 1 n mi = 1.852 km). Asterisks denote the sample mean, colored bars indicate the inner quartile range with median at the thinnest point, dashed lines span the 5th to 95th percentiles, and circles indicate outliers. MPAS denotes uniform resolution, and MPAS-EP denotes variable resolution. The number of storms at each forecast time appears above the plot. In (a), storms are required to exist at the initial time and in (b) storms are not required to exist at the initial time.

exist at $t = 0$, or whether genesis during the forecast is included. In other basins, where the variable-resolution MPAS has a cell-center spacing of 60 km, the biases of the MPAS-EP forecasts are very similar to those of the GFS, about 10 kt (not shown). The uniform-resolution MPAS maintains smaller median-intensity biases in other basins (not shown).

The comparison of intensity between the different sets of forecasts is not surprising owing to the strong signature of the underlying grid spacing on the maximum winds that can be produced, and because of the smoothing inherent in GFS winds using the tracker. If instead of the winds from the MPAS native grid, we use the maximum wind derived from the tracker, both configurations of MPAS have a similar intensity bias as in the GFS (not shown).

The above does not imply that MPAS intensity forecasts are without bias. This point is brought out by

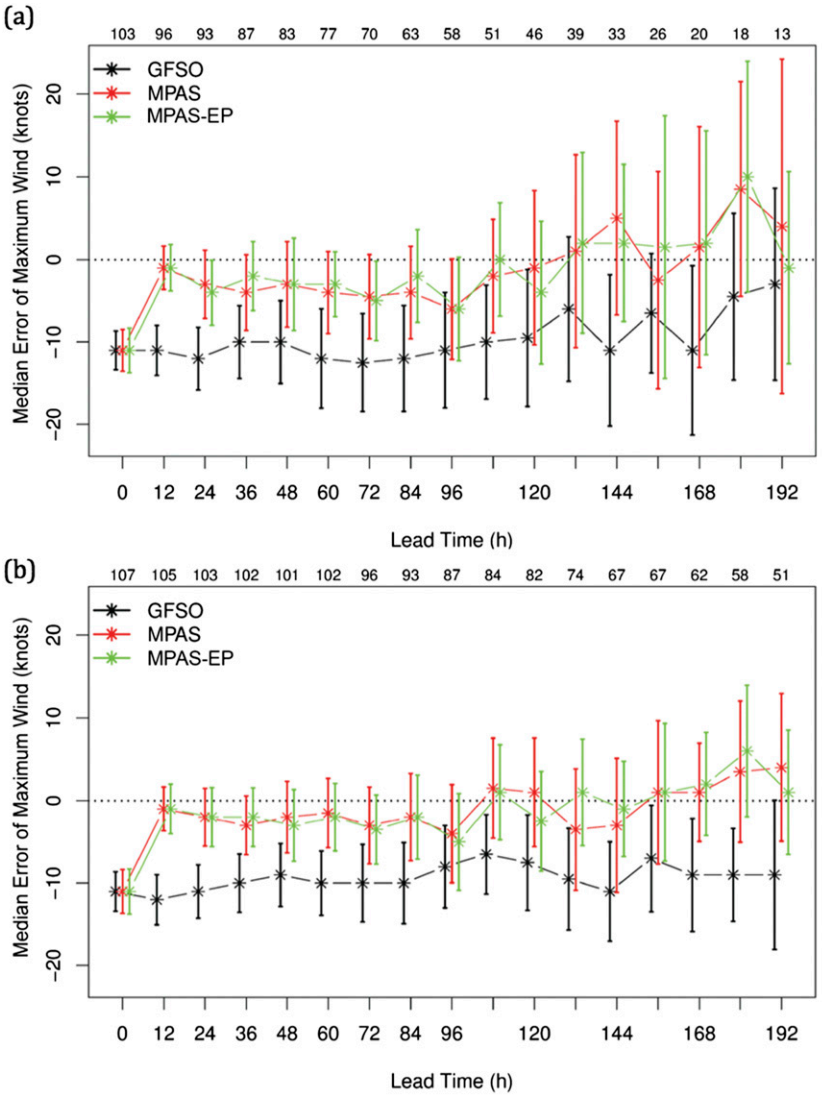


FIG. 3. (a),(b) Median intensity errors (forecast − observation), with bars indicating inner quartile ranges, for GFS (black), uniform-resolution MPAS (red), and variable-resolution MPAS (green) for the east Pacific. In (a), storms are required to exist at the initial time and in (b) storms are not required to exist at the initial time.

looking at the distribution of intensity values for the forecast storms versus the observed intensity values. This metric is not strictly related to bias, but it does examine discrimination, that is, the ability of the model to distinguish storms of different intensities (Murphy 1993). In this case, discrimination is partly influenced by bias and partly influenced by the rarity of intense hurricanes. From Fig. 4, it is clear that MPAS fails to predict intensities greater than about 110 kt over the east Pacific, whereas the observed intensities reach 140 kt. The discrimination of both MPAS and the GFS is relatively good through the range of 40–80-kt maximum intensity, but weak at higher intensities. MPAS better predicts storms of hurricane intensity, but it also overestimates the intensity of the weakest disturbances. MPAS produces relatively unbiased intensity forecasts across the most commonly observed intensities (30–60 kt), and this contributes to the relatively small MPAS intensity bias overall. The negative GFS intensity bias is evident for all observed intensities greater than 50 kt.

b. *Categorical verification*

The preceding section discussed only the forecast TCs corresponding to observed TCs. This section addresses forecasts, especially in the medium range (4 days or more), when there is not necessarily a correspondence

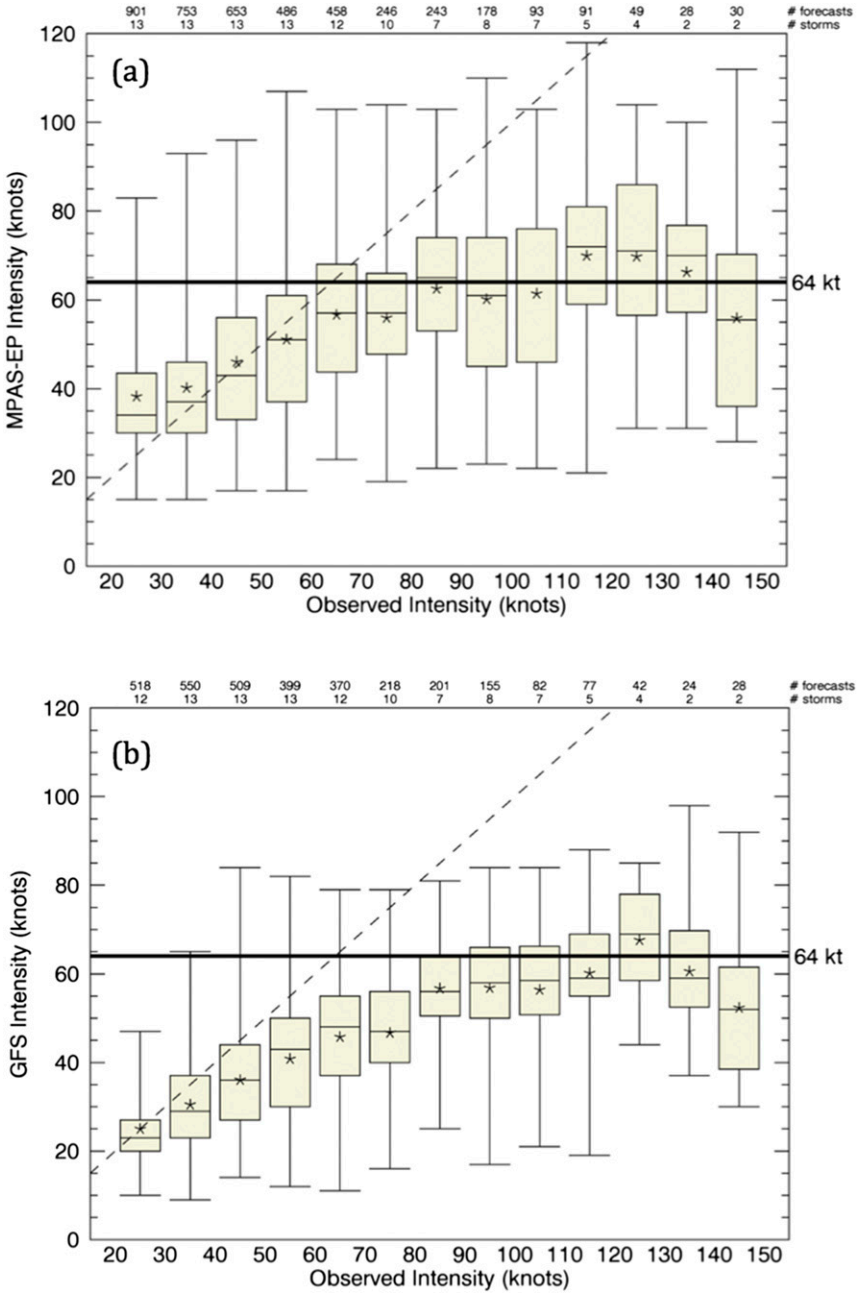


FIG. 4. Intensity distributions of matched storm pairs conditioned on observed intensity binned in 10-kt increments beginning with [15, 25) knots for the observed bin labeled "20 kt." (a) MPAS-EP and (b) GFS. The number of pairs in each bin appears at the top of each plot, below which is the number of distinct storms represented in each bin. Shaded boxes denote the median and interquartile range, asterisks denote the mean, and whiskers denote the minimum and maximum.

between forecast and observed storms. The categorical verification methodology described in section 2 and the appendix is applied here.

For the full 3×3 contingency table and $V = 64$ kt, the Heidke skill curves reveal that the two configurations of MPAS perform comparably at all lead times (Fig. 5a). Furthermore, the GFS attains higher skill scores by an appreciable margin after day 3. The decay rate of S in Fig. 5a is strongly suggestive of synoptic-scale predictability. The fact that some skill is evident even at day

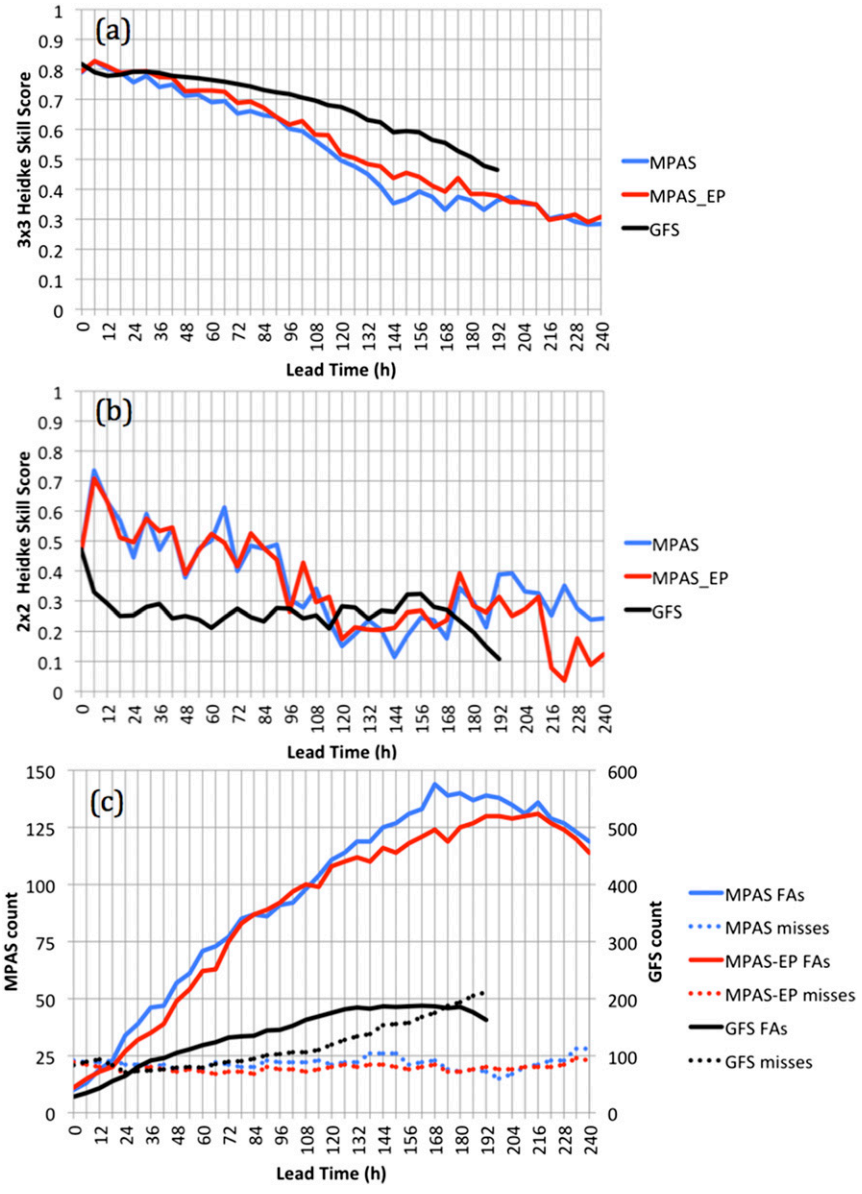


FIG. 5. Heidke skill scores for (a) the full 3×3 contingency table given the intensity threshold $V = 64$ kt, (b) the 2×2 contingency table for matched tracks, and (c) the sum of false alarms ($YN + MN$ in Fig. A1c, solid) and misses ($NY + NM$ in Fig. A1c, dotted).

10 implies that there is value in longer-range TC forecasts. While a threshold intensity is needed to define S, the relatively slow decay of skill implies that the multicategory skill score mainly contains information related to storm tracks.

However, if we restrict the evaluation to matched tracks (represented by the cells enclosed by the red box in Fig. A1c), and compute the Heidke skill score for the dichotomous forecasts of hurricane intensity (following Table A1), it is apparent that the MPAS scores are higher for the first 4 days or so (Fig. 5b). In this 2×2 case, cell MM from the 3×3 table (correct forecasts of storms below intensity V) becomes the sum of correct negative forecasts, and cells MY and YM become misses and false alarms, respectively. For $V = 64$ kt, the number of correct negative forecasts along matched tracks is relatively large; hence, the Heidke score for the 2×2 case reduces approximately to (A3). A summary of both Figs. 5a and 5b is that the GFS forecasts are better overall, but the MPAS forecasts are better able to discern hurricane intensity along matching tracks, at least at short lead times. The latter result is somewhat

expected given the GFS winds come from a relatively coarse 0.5° grid.

The difference in overall skill (Fig. 5a) results mainly from a greater number of false alarms in MPAS than the number of misses in the GFS. The time series of false alarms ($YN + YM$ from Fig. A1c) and missed events ($NY + NM$ from Fig. A1c) indicate that the relatively large number of false alarms in MPAS coincides with the time when the Heidke score from the 3×3 contingency table is higher for the GFS (Fig. 5c). Note that the Heidke score treats all false alarms the same, regardless of intensity, and the same for missed events. It turns out that the sums YN and NY, which represent false alarms and misses of hurricane intensity, respectively, are small compared to terms MN and NM. Thus, most of the false alarms and misses are relatively weak storms. The larger number of false alarms decreases the skill score by inflating the total counts in the denominator of (A1) while also increasing the hits due to random guessing. The number of false alarms would be reduced by roughly a factor of 3 if the MPAS maximum winds were derived from the tracker instead of the native grid. This would increase S computed from the 3×3 contingency table, but it would decrease the Heidke skill of intensity forecasts.

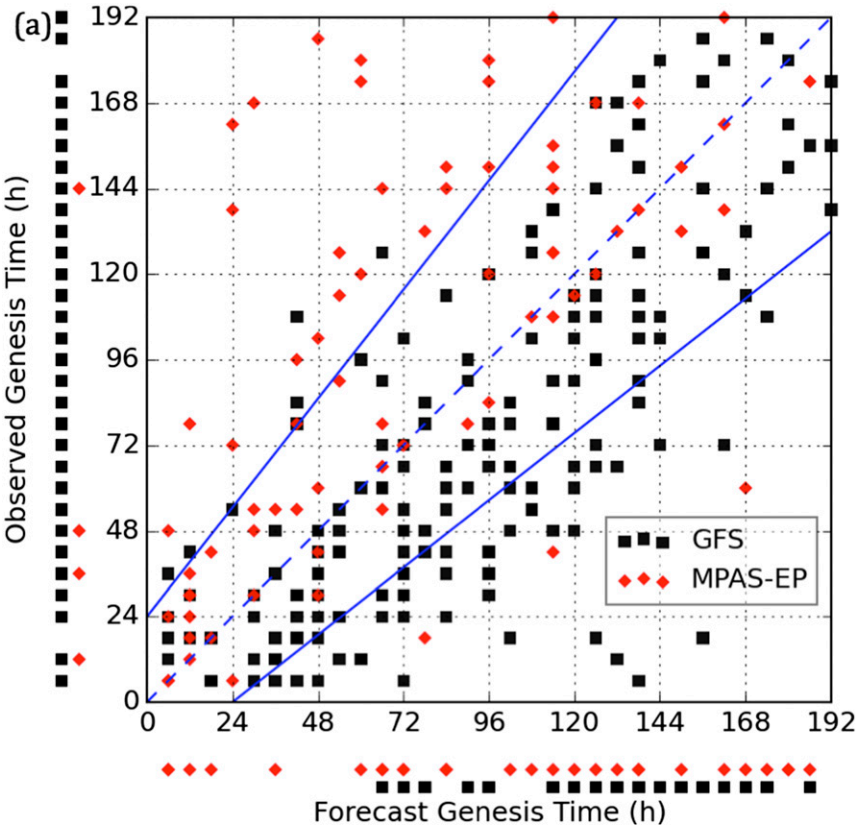
The relatively rapid loss of skill exhibited in Fig. 5b clearly relates to the difficulty of intensity prediction. It also appears that there may be a significant adjustment problem in the GFS at early lead times given the loss of skill in the first 12 h. The nonzero values of S in Fig. 5b at long lead times are not necessarily indicative of skill. For the 2×2 case, Doswell et al. (1990) showed that S is approximately twice the critical success index (CSI) in the limit of rare events, and CSI > 0 as long as $YY > 0$ (see Table A1).

A straightforward extension of our verification method provides information specifically about TC genesis prediction skill. Having a pair of matched tracks, we compute the time of genesis separately for the forecast and observed storms. Genesis is the first occurrence of a maximum wind speed of at least 34 kt. The ordered pairs (forecast genesis, observed genesis) are shown in Fig. 6a. Cases in which an observed storm or a forecast storm has a maximum wind of 34 kt or greater at model initialization time are not counted in genesis statistics. Forecasts of genesis without corresponding observed genesis along a matched track appear below the abscissa; observed genesis with no corresponding forecast genesis appears to the left of the ordinate. Correct forecasts of no genesis are not shown in the figure. Results for all four GFS daily cycles are included.

It is clear that MPAS-EP (and MPAS, not shown) produce genesis too early, and the GFS tends to produce genesis too late. We can quantify genesis skill by computing a Heidke score using (A5), assuming that genesis is a rare event. Pairs that occur between the two solid blue lines in Fig. 6a are identified as correct genesis forecasts (cell YY of Fig. A1c). These lines bound an initial timing error of 24 h. The tolerance grows linearly to 72 h by day 8. For reference, Halperin et al. (2013) used a constant error tolerance of 24 h for forecasts extending to 96 h. Pairs above the higher line represent early genesis (cell YM in Fig. A1c). False alarms (YN) represent forecasts of genesis along a matched track that does not occur during the forecast period. Pairs below the lower solid line represent late genesis (MY). Missed events (NY) denote observed genesis with no genesis predicted along the matched track during the forecast. Correct forecasts of no genesis along a matched track appear in cell MM. These six cells correspond to the same six defined by Halperin et al. (2013, see their Fig. 5b). (Elements NM and MN from Fig. A1c remain unpopulated.) In addition to skill scores for matched tracks, we also compute the statistics for false alarms and missed tracks; each false alarm track will increment the count in cell YN, while each missed genesis event during a forecast period will increment the count in cell NY.

The 3×3 contingency tables corresponding to the scatterplots are presented in Fig. 6b. The Heidke skill scores for genesis along matched tracks are 0.78 for the GFS, 0.68 for MPAS-EP, and 0.71 for MPAS. While the GFS and MPAS exhibited the opposite timing bias (GFS late, MPAS early), the MPAS errors were somewhat larger and, combined with the relatively large number of false alarms, produced a somewhat lower skill score. Adding unmatched tracks to the genesis statistics reduces all Heidke skill scores: 0.66 for GFS, 0.44 for MPAS-EP, and 0.44 for MPAS.

To gain some understanding of what may contribute to the excessive number of false alarms in MPAS, most of which are minimal tropical storms, we examined biases in numerous fields (relative to the GFS analyses used for initialization). Some of the clearest signatures of bias are found in the near-surface variables (Fig. 7). Global plots of the moist static energy (MSE) bias at 120 h, in uniform MPAS, reveal that values over the tropical oceans near and poleward of the equator are too large (Fig. 7a), primarily owing to excessive water vapor mixing ratio. The positive wind speed bias at 10 m (Fig. 7b) bears some resemblance to the bias in MSE over oceans. Over the tropical eastern North Pacific, the wind speed bias is particularly large where the surface winds blow predominantly across the equator in response to the northward-directed gradient of sea surface temperature and associated pressure gradient force. A similar bias occurs over the tropical Atlantic Ocean.



(b)

52	22	38
4	14	-
1	-	MPAS

S=0.71

50	23	40
3	11	-
5	-	MPAS -EP

S=0.68

137	6	34
38	109	-
78	-	GFS

S=0.78

FIG. 6. (a) Scatterplot of (forecast − observed) genesis times for matched tracks. Points plotted below abscissa indicate the time of predicted genesis in cases where observed genesis either did not occur, or occurred after the end of the forecast period. Points plotted to the left of the ordinate indicate observed genesis times in cases where the forecast did not produce genesis by 192 h. GFS (black squares) and MPAS-EP (red diamonds) forecasts are shown only through day 8. Dashed blue line is the 1:1 line and solid blue lines bound the region of correct genesis forecasts. (b) Counts in the 3 × 3 contingency table (see Fig. A1c) for genesis forecasts along matched tracks in MPAS, MPAS-EP, and the GFS.

Given that the analyses over tropical oceans have uncertainties as well, we also compared 10-m winds to wind estimates from the Advanced Scatterometer (ASCAT). ASCAT wind speeds[4] were paired with the closest MPAS centroid (to which both wind components were mapped), allowing at most a 1.5-h time displacement. The MPAS–ASCAT wind speed pairs were grouped into 1° bins and the speed difference was averaged to obtain the wind speed bias. An analogous procedure was followed for the GFS winds on the 0.5° grid. While the MPAS winds have a high bias (Fig. 7c), the GFS forecast surface winds appear to be too weak over the eastern Pacific (and elsewhere) (Fig. 7d). This

[4] In accord with the ASCAT user manual guidance, winds were not used if the monitoring flag, the KNMI quality control flag or the variational quality control flag were set.

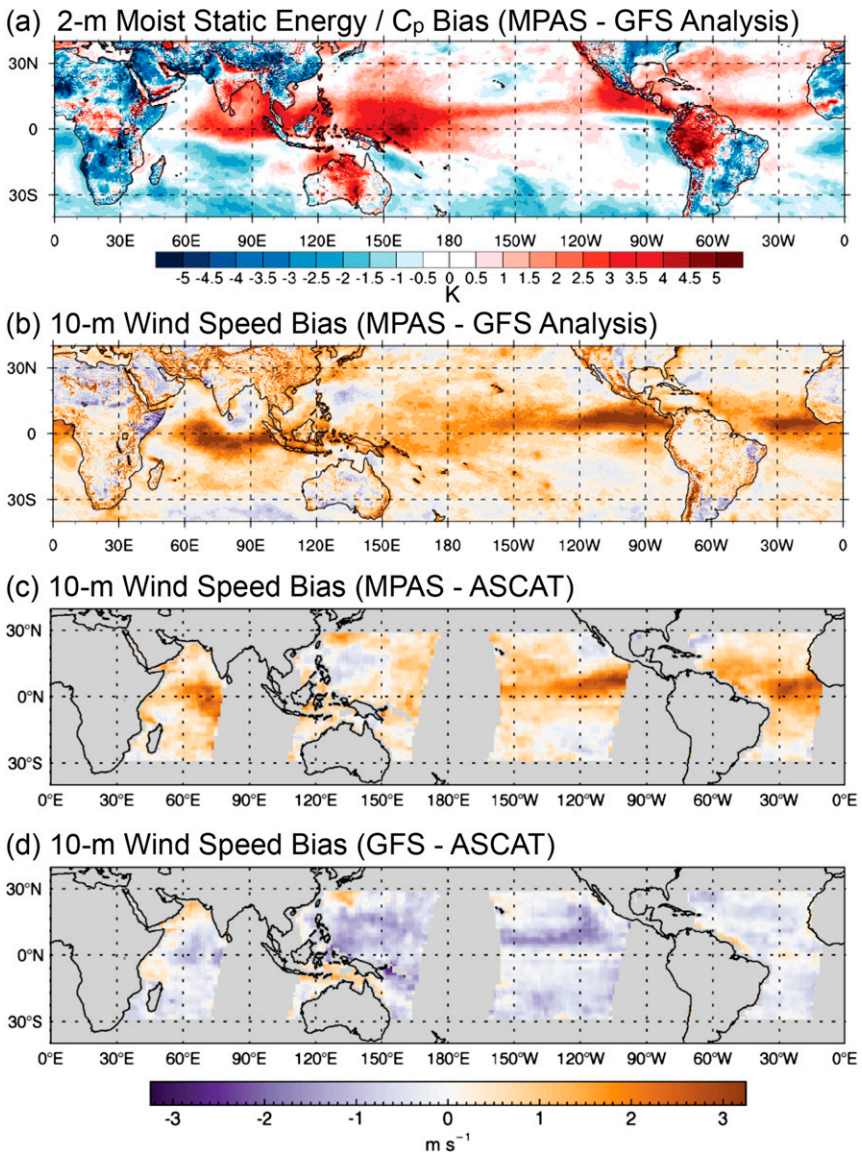


FIG. 7. Uniform-resolution MPAS bias spatial distributions at 120 h, averaged over all forecasts, for (a) 2-m moist static energy, normalized by heat capacity C_p (K); (b) 10-m wind speed ($m\,s^{-1}$) compared to GFS analysis; (c) 10-m MPAS wind speed bias compared to ASCAT winds ($m\,s^{-1}$); and (d) 10-m GFS wind speed bias compared to ASCAT winds ($m\,s^{-1}$).

GFS bias and its attendant low bias in surface fluxes (not shown) may contribute to the reduced number of TCs in the GFS compared with observations. While Chou et al. (2013) have suggested a correction to the ASCAT data for underestimates of wind at wind speeds above roughly 12 $m\,s^{-1}$, that correction was not applied here.

Over cooler water near and to the south of the equator, stratocumulus is common beneath a warm and dry midtroposphere. Air flowing across the equator is warmed and moistened by surface fluxes, reducing convective inhibition as air approaches the intertropical convergence zone (ITCZ) (Raymond et al. 2006). Zonal averages of meridional wind and sea level pressure (Fig. 8) reveal the excessive southerly flow across the equator that is consistent with an anomalous pressure gradient. Enhanced destabilization results in more widespread rainfall. Rainfall in the forecast ITCZ averages more than 25 mm day^{-1}, which is nearly twice the typical rate indicated by Raymond et al. (2006). Enhanced boundary layer convergence contributes to relative vorticity, producing a maximum of about $2 \times 10^{-5}\,s^{-1}$, which is twice the composite relative vorticity at 0 h,

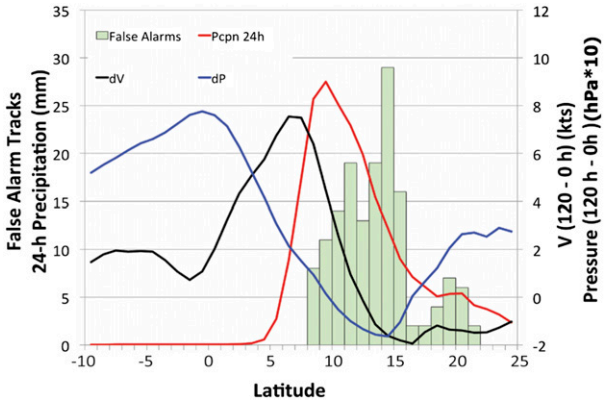


FIG. 8. Averages over 95 forecasts and 90°–120°W longitude of precipitation (red) from 120 to 144 h, 10-m meridional wind difference between 120 and 0 h (*dV*, black), and sea level pressure difference (*dP*, 120 minus 0 h, blue) from MPAS-EP. Also graphed are the track-initiation latitudes of false alarms in MPAS-EP (green bars).

and roughly 2/3 of the Coriolis parameter at 12°N. The origins of false alarm tracks (Fig. 8) are in close proximity to the anomalously strong convergence, rainfall and vorticity, but the false alarms are skewed poleward toward larger values of the Coriolis parameter.

The pressure anomaly of about 0.8 hPa near the equator is consistent with a cold bias in the 850-hPa temperature (Fig. 9a) over the tropical South Pacific basin, as well as over the tropical Atlantic. A representative sounding from the equatorial eastern Pacific shows that the cold bias arises from erroneous lifting of the inversion atop the stratocumulus layer (Fig. 9b). It appears that shallow convection in the Tiedtke scheme is too vigorous in this region. A temperature reduction averaging 3°C over the layer between 880 and 820 hPa is consistent with a decrease in layer thickness of about 6 m. If we assume that this reduction in layer thickness is manifested in the sea level pressure, it accounts for most of the 0.8 hPa excess near the equator. Therefore, we hypothesize that the error in shallow convection to the south of the equator enhances the pressure gradient force in the boundary layer, which in turn, erroneously accelerates the cross-equatorial flow and invigorates the ITCZ farther north. Other effects arising from erroneous temperature gradients produced by errors in shallow convection were also noted by Torn and Davis (2012).

Although the entrainment parameter we use is smaller than that used by Zhang et al. (2011), sensitivity tests indicate that the bias in shallow convection is not altered by simply increasing its entrainment rate. An entirely new version of the Tiedtke scheme (C. Zhang 2015, personal communication) has been tested in MPAS and appears to have no cold bias at 850 hPa.

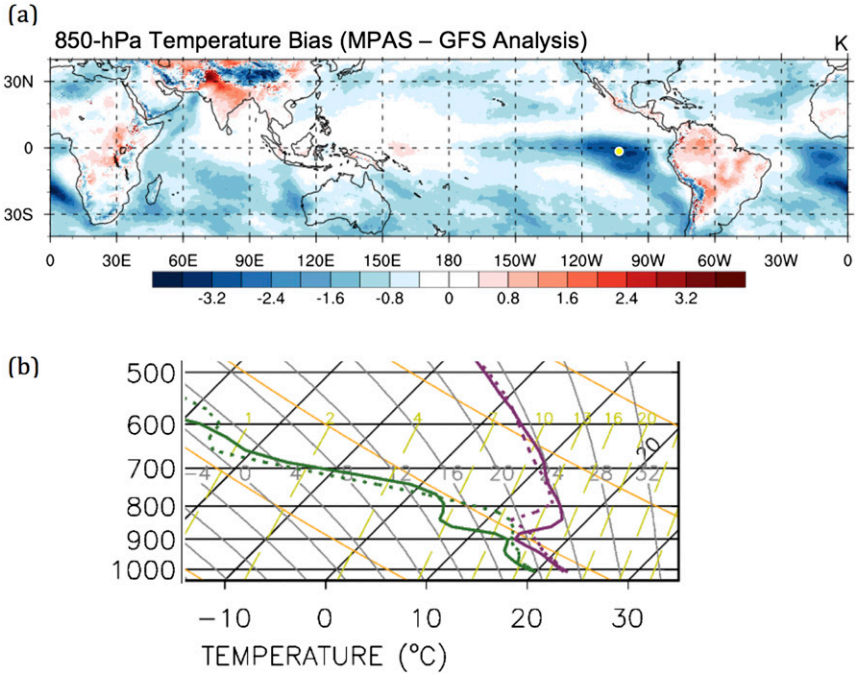


FIG. 9. (a) Temperature bias at 850 hPa averaged over all 120-h uniform-MPAS forecasts. (b) Representative skew *T*–log*p* showing the temperature (purple) and dewpoint temperature (green) profiles at the point indicated by the yellow dot in (a), which is valid at 0000 UTC 29 Aug 2014. The dashed line is the 96-h forecast and the solid lines show the corresponding analysis.

Future work will be devoted to understanding why this is so, and investigating the consequences for the model ITCZ representation and TC prediction.

c. *Prediction with variable resolution*

While it is apparent that TC prediction using variable resolution produces similar results to prediction with uniform resolution out to 10 days, at least in the basin where the resolution is the same, this section examines more general forecast aspects, as well as how the forecasts using the two MPAS configurations begin to diverge. For these purposes, it is useful to partition the eastern North Pacific into two regions: tropical (0°–23.5°N) and extratropical (23.5°–45°N). In what follows, we normalize root-mean-square differences between the two MPAS configurations by the spatial standard deviation computed from the analysis over the corresponding latitude belt and over all longitudes. For each variable there is one normalizing factor for the tropics and a different factor for the extratropics. The normalization allows a comparison of different variables and regions with differing intrinsic variability.

Wind differences at 500 hPa grow faster in the tropics initially (Fig. 10). The curves cross around day 4, with faster growth in the midlatitudes thereafter. Differences grow rather slowly in the tropics between days 2 and 7. The behavior in the tropics may be related to initially rapidly growing differences due to the intrinsically unpredictable nature of deep convection. However, slower difference growth at longer time scales may result from the quasi-linear nature of tropical waves compared with quasi-exponential difference growth in midlatitude baroclinic waves.

Some insight into the behavior of forecast differences comes from examination of time–longitude (Hovmöller) diagrams of the evolution of 500-hPa wind differences (Fig. 11). Other variables (e.g., 200-hPa geopotential height and 500-hPa temperature) exhibit similar behavior. Differences in the extratropics grow relatively uniformly around the hemisphere, with some modest enhancement in the Pacific and Atlantic storm tracks. In the tropics, the growth of normalized differences varies strongly with longitude. Differences grow rapidly over North Africa, punctuated by the diurnal cycle, and relatively rapidly over Southeast Asia and the Indian Ocean. Differences grow most slowly over the longitude band 80°–50°W, and also 30°–60°E, which are both regions of reduced rainfall (not shown). Strong tropical cyclones that differ in the two models are evident as quasi-vertical streaks in the tropics.

The behavior in Fig. 11 appears driven by differences in treating deep convection in the uniform and variable MPAS configurations. If convection is weak, or if the resolution in the two configurations is similar, the solutions diverge at a later time. Over Africa and Southeast Asia, differences grow rapidly over the first diurnal cycle, presumably because of the differing treatment of deep convection on the 60- and 15-km meshes even though the cumulus scheme is the same. The slower growth of differences over the tropical eastern Pacific contrasts with more rapid difference growth elsewhere that arises because the parameterization scheme is not "scale aware."

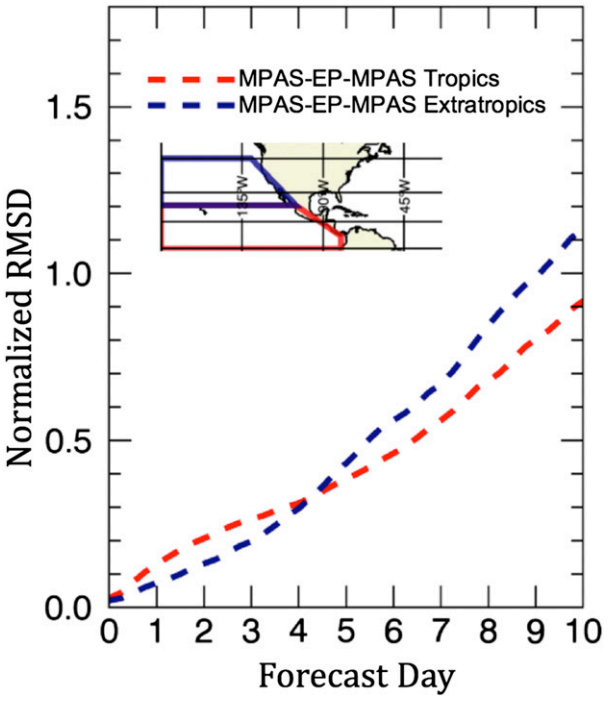


FIG. 10. RMS differences in 500-hPa vector wind between MPAS and MPAS-EP for the tropics (red) and extratropics (blue). RMS differences are normalized by the RMS spatial variance within the respective regions. Inset map shows the geographical extent of the two regions.

Over the longitude band 130°E–180°, the apparent eastward-propagating signal arises because of the gradient in cell-center spacing (Fig. 1) and the increasing similarity of convection as the cell spacing in the two configurations converges farther east. Westward movement of differences from Africa is evident, and may have a physical basis because it does not occur in the gradient of mesh spacing, and because westward movement is expected in the tropical easterlies. However, the westward migration of differences appears to affect only the eastern Atlantic during the forecast period.

Over the eastern Pacific, we can see how tropical cyclones tend to exacerbate the differences between the two model configurations (Fig. 11). Differences tend to grow faster over the longitude range 120°–90°W, where most of the TCs develop. Differences near day 10 in this longitude band approach values in other longitude bands where errors initially grow much more rapidly.

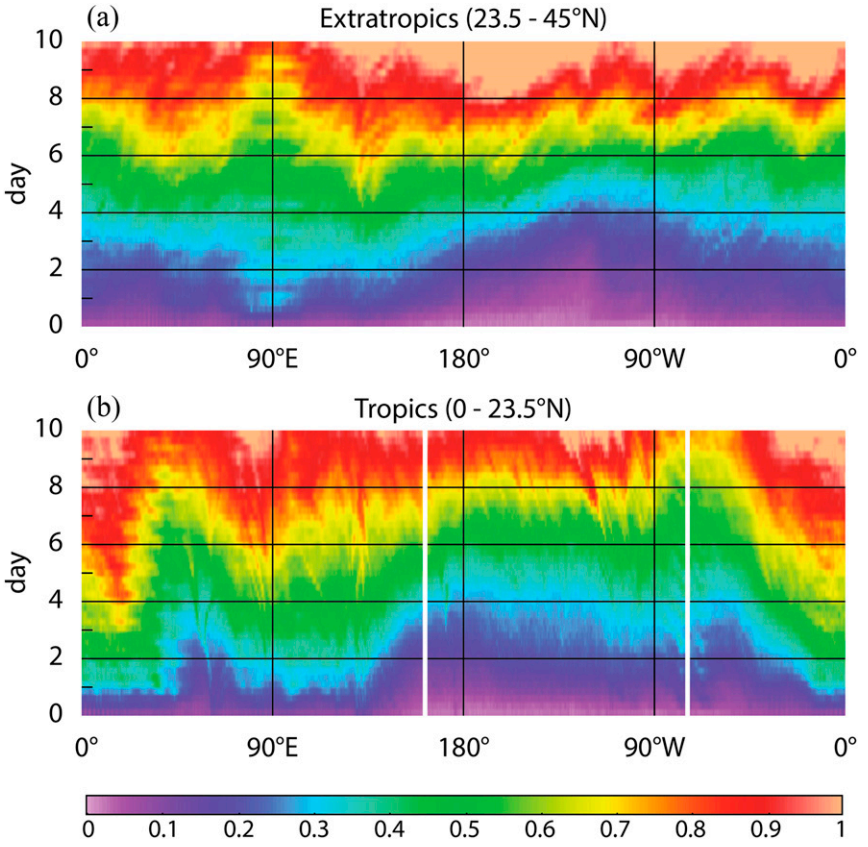


FIG. 11. (a),(b) Time–longitude diagrams of the latitude and forecast-averaged standard deviation of vector wind difference (MPAS minus MPAS-EP) at 500 hPa. The RMS differences are normalized by the respective spatial standard deviations for each latitude band. The latitude bands are illustrated above each panel. Vertical solid white lines in (b) bound the approximate extent of 15-km grid spacing in the variable-resolution configuration.

Overall, it appears that differences between the two configurations do not exhibit significant zonal propagation; hence, the region of finer variable resolution is not rapidly influenced by differences elsewhere. It is possible that differences in the midlatitudes eventually influence the tropical eastern Pacific, perhaps by influencing tropical cyclone behavior. Overall, the local nature of differences between the two configurations is consistent with the similarity of TC statistics over the eastern Pacific. This suggests that a variable-resolution approach to forecasting tropical cyclones appears practical well into the medium range. Further examination of variable-resolution forecasts in other basins will be needed to demonstrate the generality of this finding.

4. Conclusions

We have examined the performance of the Model for Prediction Across Scales (MPAS) for eastern Pacific tropical cyclones in 2014, and compared forecasts integrated on a nearly uniform 15-km mesh with those integrated on a variable-resolution mesh ranging from a 15-km cell spacing over the eastern Pacific to 60 km across the remainder of the globe (denoted MPAS-EP). Both forecasts were also compared with the operational GFS forecasts.

The time period was 1 August–3 November 2014 (95 forecasts). Because the models were integrated to 10 days, a new verification method was developed to account for tropical cyclones that developed during the forecast period. This method, based on defining storm tracks as objects, allowed us to assess false alarms (forecast tracks that were not observed), missed events (observed tracks with no counterpart in a particular forecast), and correct negative forecasts. The method also allowed us to evaluate track and intensity errors for storms not present in the model initial condition.

Overall, the evaluation of forecast skill showed that the two configurations of MPAS performed nearly identically to each other over the eastern Pacific and

performed comparably to the GFS. For a homogeneous sample of matched tracks, the position errors were similar. MPAS had a smaller intensity bias than GFS and more ability to discriminate hurricane intensity than GFS, consistent with its somewhat finer resolution. The GFS was more skillful for TC genesis. Both configurations of MPAS produced an excessive number of weak storms late in the forecast period, whereas the GFS missed a large number of events. The false alarms in MPAS reduced the Heidke skill score and the skill of genesis forecasts.

It was shown that the excessive number of weak storms in MPAS stemmed from an overall bias of boundary layer wind speed in the cross-equatorial flow over the eastern Pacific. Related to this was a positive bias in moist static energy, average rain rate, and boundary layer convergence in the intertropical convergence zone. The wind biases were traced to pressure-gradient errors that resulted from excessive vertical mixing of the Tiedtke shallow convection parameterizations in the tropical eastern South Pacific. A future publication will quantify the performance of TC forecasts using a different version of the Tiedtke scheme that does not have this bias.

The variable and uniform configurations of MPAS produced very similar forecasts of TCs, with quantitative similarity in most cases through day 7. Furthermore, the difference between solutions over the tropical eastern Pacific exhibited slow growth between days 2 and 7. Where the MPAS configurations possessed widely different horizontal resolution, and where convection was frequent, the growth of differences was much more rapid. However, differences in such regions were fairly localized. This suggests that variable resolution is a cost-effective approach to medium-range prediction of TCs.

Future investigation will expand the use of variable-resolution MPAS to TC prediction in other basins, utilizing improvements in convective parameterization and the multicategorical verification methodology. While the forecasts analyzed herein were performed at hydrostatic scales, future work will report on the extension to nonhydrostatic scales where the TC inner core should be well resolved.

Acknowledgments. The authors wish to thank Dr. Barbara Brown of NCAR for helpful comments about the verification methodology and Dr. Jonathan Vigh of NCAR for answering questions about the best track datasets. Also, the authors thank Dr. Mike Fiorino of NOAA for numerous helpful comments and assistance with the tracking software, and the authors thank an anonymous reviewer for helpful comments. Funding for this research was provided by the National Center for Atmospheric Research through support from the National Science Foundation under Cooperative Support Agreement AGS-0856145. The content of this publication is solely the responsibility of the author(s) and does not necessarily represent the official views of the National Science Foundation.

APPENDIX

Tracking, Track Matching, and Verification Methodology

Output from MPAS was saved every 6 h. To process MPAS output, evaluate forecasts, and provide input for the tracking algorithm [the "GFDL tracker," hereafter simply "the tracker;" Knutson et al. (2007), their appendix B],[A1] an interpolation routine mapped each MPAS forecast to the same 0.5° latitude–longitude grid using linear interpolation from the MPAS cells to the latitude–longitude gridcell centers. Prior to interpolation, smoothing was performed iteratively on the native MPAS mesh. During each pass the cell values were averaged with their neighbors. Each cell was assigned a number of smoothing passes that varied inversely with its size. Following the smoothing, Delaunay triangulation was used to find the closest MPAS cell centers and obtain weights for the final interpolated values. These weights varied inversely with the local separation of centroids, yielding a relatively uniformly smoothed mesh even with variable resolution. For vorticity, which is defined at cell vertices, the interpolation was preceded by a step where the values of vorticity at cell vertices were averaged to obtain centroid values. The smoothed output was helpful for identifying TC tracks, and was essential in preventing the vortex tracker from identifying a plethora of short tracks.[A2]

Output on the 0.5° grid was used to identify and track tropical cyclones. Cyclone centers were based on a consensus of 850- and 700-hPa vorticity, 850- and 700-hPa wind speed, 850- and 700-hPa geopotential height, MSLP, 10-m wind speed, and 10-m vorticity. For thresholds, we used the same namelist settings as in the tracker test cases: *trkrinfo%mslpthresh=0.0015, trkrinfo%v850thresh=1.5, and trkrinfo%contint=100*. These parameters are slightly different than those described in Knutson et al. (2007). Identification of a warm-core vortex utilized two diagnostics, cyclone phase space (CPS; Hart 2003) and an

[A1] The tracker code was obtained online at http://www.dtcenter.org/HurrWRF/users/downloads/index.tracker.php, version v3.5b.

[A2] To allow all tracks to be discovered, the parameter *maxstorm_mg* had to be set to 5000.

TABLE A1. The 2 × 2 contingency table for dichotomous (forecast − observed) pairs, where Σ denotes the sum of either a row or column.

	Observed		
Forecast	*YY* (hit)	*YN* (false alarm)	ΣF_Y
	NY (miss)	*NN* (correct negative)	ΣF_N
	ΣO_Y	ΣO_N	*T*

area-mean temperature anomaly criterion (*phaseflag='n', phasescheme='both', and wcore_depth=1.0*), motivated by Vitart et al. (1997). The CPS calculation determined if the storm structure resided in the symmetric warm core category.

It is common to apply a skill score to assess model performance. In this case, the Heidke skill score was applied to multicategory TC forecasts that are able to account for more characteristics of TCs than simply their existence. The Heidke score for dichotomous forecasts was summarized in detail by Doswell et al. (1990). Aberson (2008) discussed a multicategory form of the Heidke skill score for TC intensity evaluation. The score takes the following form:

$$S = \frac{C - E}{T - E}, \tag{A1}$$

where *C* is the number of correct forecasts, *T* is the total number of forecasts, and *E* is an estimate of the number of correct forecasts that could be obtained by random guessing. Essentially, *S* measures the ability of the forecasts to predict the correct category, measured across all possible categories, relative to the ability of a random chance forecast.

Dichotomous forecasts lend themselves to an evaluation using a 2 × 2 contingency table (Table A1). In this case, the elements of (A1) are as follows:

$$C = YY + NN; \quad E = \frac{\Sigma F_Y \Sigma O_Y + \Sigma F_N \Sigma O_N}{T};$$
$$T = YY + NY + YN + NN, \tag{A2}$$

where *YY* and *NN* represent the number of hits and correct negative forecasts, respectively. Here, the double-letter notation refers to elements of Table A1. In addition, $\sum F_Y$ is the sum of forecast events, $\sum F_N$ is the sum of forecast nonevents, and $\sum O_Y$ and $\sum O_N$ are the analogous sums for observations. Furthermore, *NY* represents the number of events observed but not predicted (misses), and *YN* refers to the number events forecast but not observed (false alarms). In the limit that *NN* is large, which is true for rare events, (A1) can be approximated as

$$S = \frac{YY}{YY + \dfrac{YN + NY}{2}}. \tag{A3}$$

This expression is related to the critical success index (CSI) as noted by Doswell et al. (1990).

Dichotomous forecasts are not sufficient to capture the variety of possibilities concerning TC forecasts. It is straightforward to extend the Heidke skill score to arbitrarily many categories (Aberson 2008). In our methodology, we will consider three possible situations: (i) no tracked disturbance, (ii) a tracked disturbance that has maximum wind speed (intensity) less than a threshold *V* at some instant, and (iii) a tracked disturbance with intensity ≥*V*.[A3] For tropical storms, *V* = 34 kt in the observations and in each model; for hurricanes, *V* = 64 kt. We assert that three categories are the minimum needed to describe forecast quality.

Using three categories (Fig. A1), the totals in (A1) are defined as follows:

$$C = YY + MM + NN;$$
$$E = \frac{\Sigma F_Y \Sigma O_Y + \Sigma F_M \Sigma O_M + \Sigma F_N \Sigma O_N}{T}, \tag{A4}$$

where *T* is the sum of all elements in the 3 × 3 table; and the subscripts *Y*, *M*, and *N* refer to a tracked disturbance that is instantaneously at or above the threshold *V* (*Y* for "yes"), below *V* (*M* for "minimal intensity"), or no tracked disturbance at all (*N* for "no"). If *NN*, the correct forecast of no tracked disturbance, is large compared to all other elements of the 3 × 3 table, then the Heidke skill score reduces to

$$S = \frac{YY + MM + \dfrac{MY + YM}{2}}{YY + MM + MY + YM + \dfrac{NY + NM + YN + MN}{2}}. \tag{A5}$$

From (A5), it is clear that the multicategory score offers what amounts to "partial credit" through the appearance of off-diagonal elements *MY* and *YM* in the numerator. These cells account for points along a matched track where the intensity is incorrectly predicted relative to the threshold *V*; cell *MY* includes forecasts below

[A3] Because the observations only include tracks of disturbances reaching tropical storm strength, we do not include simulated disturbances that never achieve at least minimal tropical storm strength.

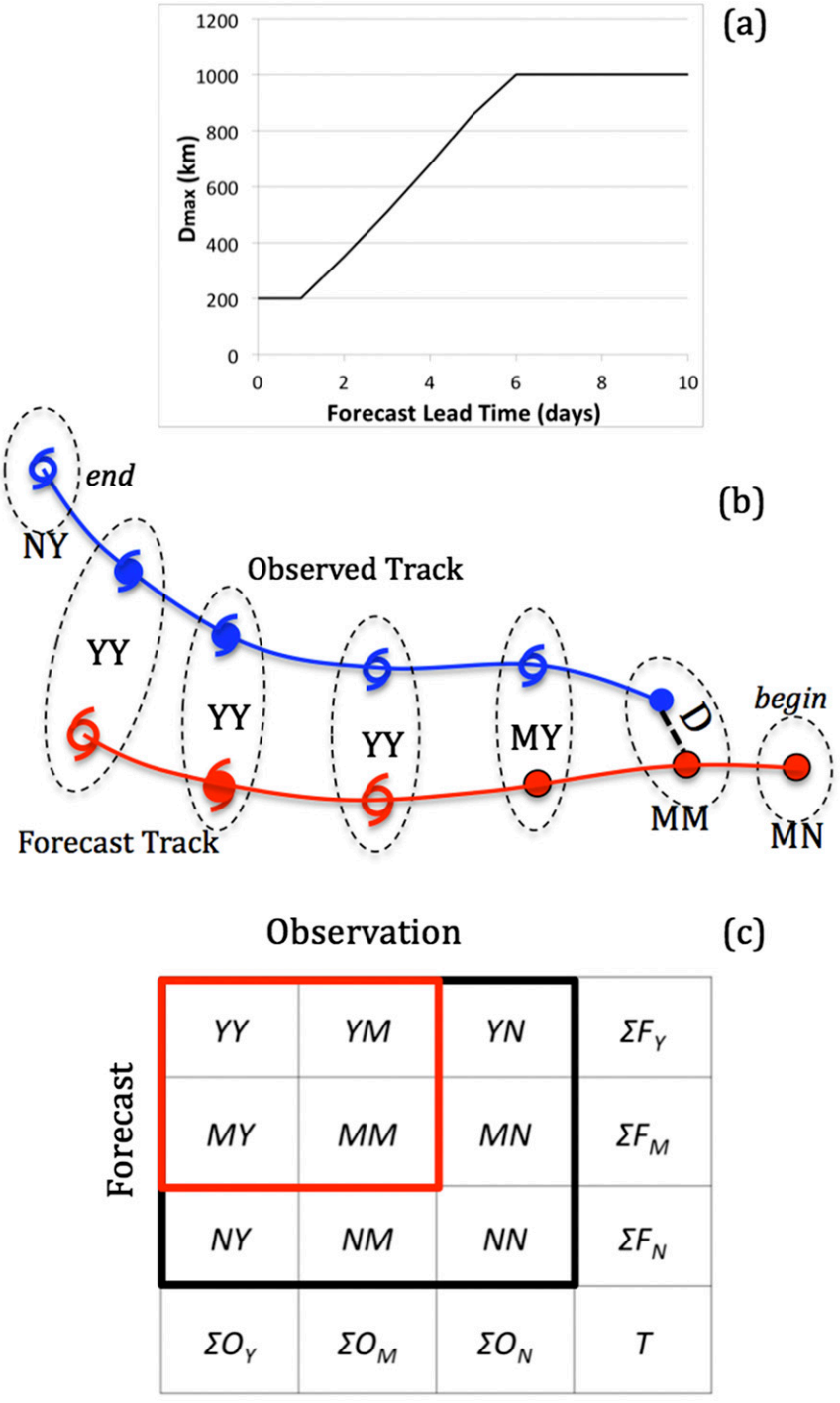


FIG. A1. (a)–(c) Summary of track-matching and evaluation method. (a) The tolerance for the track separation at the initial common time D in order to define matched tracks. (b) Schematic of a pair of tracks where storm location is indicated by symbols (closed circle for less than tropical storm intensity). Dashed ellipses encircle common times. Black pairs of capital letters refer to specific elements of the 3×3 contingency table in (c) that are incremented according to the example in (b) for $V = 34$ kt (see text for details).

intensity V paired with observations above V, and vice versa for cell YM.

The track-matching and contingency-table population methodology is illustrated in Fig. A1. The tracks of a forecast and an observed storm, moving from right to left, are shown in Fig. A1b, where the red curve and symbols indicate the forecast track. For each forecast and observed track pair that overlaps in time, at least partially, we evaluate a match based on the position error at the earliest common time. The tracks match if the separation D is less than the threshold $D_{\max}(t)$ (Fig. A1a). This threshold is approximately the 95th percentile of NHC forecast errors for a given lead time out to day 5 (Fig. 2).[A4] The maximum separation allowed is 1000 km, and this value is used beyond day 5. While this separation threshold is large, it is still small enough to ensure that the forecast and observed storms are part of the same tropical wave trough.

One might question why $D_{\max}$ should depend on lead time at all. The intent is to identify a tropical cyclone in a forecast that is the counterpart of an observed tropical cyclone. Because tropical cyclone formation is slaved to synoptic-scale tropical waves, this amounts to equating matching tolerance with the growth of synoptic-scale errors during a forecast. The lead-time dependence of $D_{\max}(t)$ is an attempt to account for that error growth. The NHC track forecast error dependence on lead time is a convenient quantitative measure for the growth of errors, and hence, for the variation of spatial displacement tolerance with time.

Along the portion of matched forecast and observed tracks coincident in time, YY, MY, YM, and MM are incremented according to whether the forecast and observed storms simultaneously achieve intensity V. For the schematic in Fig. A1, assuming $V = 34$ kt, YY will increase by 3, whereas MM and MY will each increment by 1. The first forecast point has no observed counterpart, and because the intensity is less than 34 kt, MN is incremented by 1; this is a false alarm. The final observed point has no corresponding forecast, and since the observed intensity exceeds 34 kt, NY is incremented by 1; this is a missed event. For $V = 64$ kt, YY would be incremented by 1; MM by 3; and MY, NM, and MN by 1 each. From (A5), one can deduce that the Heidke score does not depend on V in this example. More generally there is a dependence on V, but the dependence is fairly weak. This also means that the evaluation results using the full 3×3 contingency table should not be sensitive to the low-intensity bias that characterizes coarse-resolution models.

In the event that there is no matching track for the forecast in Fig. A1b, YN and MN are incremented at each 6-h interval along the track according to whether or not the forecast storm attains intensity V. For $V = 34$ kt, YN and MN would each be incremented by 3. In the event that the observed track in Fig. A1b has no counterpart, NY and NM would be incremented by 5 and 1, respectively. The partitioning between NY and NM, or YN and MN depends on the threshold V, but the overall skill score is not affected. However, because the maximum intensity achieved by storms is weakly correlated with their longevity,[A5] unmatched tracks that achieve a high intensity will tend to result in a greater penalty than unmatched tracks of weak events. To qualify as a false alarm, an unmatched simulated storm must be over water at some point, last at least 24 h, and reach tropical storm intensity between 0° and 30°N at some point. Matched forecast tracks are not required to reach tropical storm intensity. This is done to account for initialization limitations, especially for decaying storms, or timing errors of genesis late in the forecast.

Completion of the full contingency table for detecting tropical cyclones requires an estimate of correct negative forecasts, in other words, correct forecasts of the nonoccurrence of a TC. We assume that "skillful" forecasts of the nonoccurrence of a TC are confined within an area, A_T, that covers the region where tropical cyclones typically occur within a given basin. If a tropical cyclone track begins in a forecast at some time t, then our matching distance threshold (Fig. A1) precludes a correct negative forecast within an area $A_S(t) = \pi[D_{\max}(t)]^2$. If an observed storm is present, by analogy, we must exclude the same area $A_S(t)$ from A_T in our estimation of the number of correct negative forecasts. We assert that the sum of all area at lead-time t that is not within the area A_S surrounding a forecast or observed storm is equal to the total area of correct negative forecasts. In the absence of any forecast or observed storm, the area occupied by a correct negative forecast is simply A_T. Dividing the sum of all area of correct negative forecasts by $A_S(t)$ (with units of area per event) yields the number of correct negative forecasts. The number of correct negative forecasts can therefore be expressed as

[A4] The growth of the 95th percentile of NHC errors is assumed to approximate the growth of synoptic-scale errors. The 95th percentile is chosen to allow large errors that still permit a plausible connection of forecast and observed storms.

[A5] Analysis of all eastern Pacific storms from 2001 to 2013 reveals $R^2 = 0.25$ for the correlation between maximum intensity and storm longevity, according to the Extended Best Track dataset (Demuth et al. 2006).

$$NN = \frac{nA_T - A_S(\sum F_Y + \sum F_M + \sum O_Y + \sum O_M - YY - MY - YM - MM)}{A_S}$$
$$= \frac{nA_T}{A_S} - (\sum F_Y + \sum F_M + \sum O_Y + \sum O_M - YY - MY - YM - MM), \quad \text{(A6)}$$

where sums apply to n forecasts valid at lead time t. The tallies YY, MY, YM, and MM are subtracted in (A6) because the separation of points along matched tracks is usually much less than $D_{\max}(t)$ (by construction); hence, the forecast and observed storms correspond to approximately the same area. For a basin such as the eastern North Pacific, we assume that A_T is represented by a region defined as 10°–30°N, 100°–160°W. The number of correct negative forecasts is dominated by the first right-hand-side term in (A6) at early forecast lead times due to the smallness of $D_{\max}(t)$ (Fig. A1a). At long lead times, $D_{\max}(t)$ increases to the point where the bracketed term in (A6) is not negligible with respect to the first term. At these longer lead times, the number of correct negative forecasts becomes comparable to other elements of the 3×3 table, especially if the number of storms produced by the model increases.

REFERENCES

Aberson, S. D., 2008: An alternative tropical cyclone intensity forecast verification technique. *Wea. Forecasting*, **23**, 1304–1310, doi:10.1175/2008WAF2222123.1.

Chan, J. C. L., and R. H. F. Kwok, 1999: Tropical cyclone genesis in a global numerical weather prediction model. *Mon. Wea. Rev.*, **127**, 611–624, doi:10.1175/1520-0493(1999)127<0611:TCGIAG>2.0.CO;2.

Chou, K.-H., C.-C. Wu, and S.-Z. Lin, 2013: Assessment of the ASCAT wind error characteristics by global dropwindsonde observations. *J. Geophys. Res. Atmos.*, **118**, 9011–9021, doi:10.1002/jgrd.50724.

Coniglio, M. C., J. Correia Jr., P. T. Marsh, and F. Kong, 2013: Verification of convection-allowing WRF model forecasts of the planetary boundary layer using sounding observations. *Wea. Forecasting*, **28**, 842–862, doi:10.1175/WAF-D-12-00103.1.

Davis, C., and Coauthors, 2008: Prediction of landfalling hurricanes with the Advanced Hurricane WRF model. *Mon. Wea. Rev.*, **136**, 1990–2005, doi:10.1175/2007MWR2085.1.

Demuth, J., M. DeMaria, and J. A. Knaff, 2006: Improvement of advanced microwave sounder unit tropical cyclone intensity and size estimation algorithms. *J. Appl. Meteor. Climatol.*, **45**, 1573–1581, doi:10.1175/JAM2429.1.

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585, doi:10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Elsberry, R. L., W. M. Clune, and P. A. Harr, 2009: Evaluation of global model early track and formation prediction during the combined TCS08 and T-PARC field experiment. *Asia-Pac. J. Atmos. Sci.*, **45**, 357–374.

Fairall, C. W., E. F. Bradley, J. E. Hare, A. A. Grachev, and J. B. Edson, 2003: Bulk parameterization of air–sea fluxes: Updates and verification for the COARE algorithm. *J. Climate*, **16**, 571–591, doi:10.1175/1520-0442(2003)016<0571:BPOASF>2.0.CO;2.

Gall, R., J. Franklin, F. Marks, E. N. Rappaport, and F. Toepfer, 2013: The Hurricane Forecast Improvement Project. *Bull. Amer. Meteor. Soc.*, **94**, 329–343, doi:10.1175/BAMS-D-12-00071.1.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of tropical cyclone genesis forecasts from global numerical models. *Wea. Forecasting*, **28**, 1423–1445, doi:10.1175/WAF-D-13-00008.1.

Hart, R. E., 2003: A cyclone phase space derived from thermal wind and thermal asymmetry. *Mon. Wea. Rev.*, **131**, 585–616, doi:10.1175/1520-0493(2003)131<0585:ACPSDF>2.0.CO;2.

Hashimoto, A., J. M. Done, L. D. Fowler, and C. L. Bruyère, 2016: Tropical cyclone activity in nested regional and global grid-refined simulations. *Climate Dyn.*, **47**, 497–508, doi:10.1007/s00382-015-2852-2.

Klemp, J. B., W. C. Skamarock, and S.-H. Park, 2015: Idealized global nonhydrostatic atmospheric test cases on a reduced-radius sphere. *J. Adv. Model. Earth Syst.*, **7**, 1155–1177, doi:10.1002/2015MS000435.

Knaff, J. A., 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.

Knutson, T. R., J. J. Sirutis, S. T. Garner, I. M. Held, and R. E. Tuleya, 2007: Simulation of the recent multidecadal increase of Atlantic hurricane activity using an 18-km-grid regional model. *Bull. Amer. Meteor. Soc.*, **88**, 1549–1565, doi:10.1175/BAMS-88-10-1549.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Noh, Y., W. G. Cheon, S.-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer Meteor.*, **107**, 401–427, doi:10.1023/A:1022146015946.

Park, S.-H., W. C. Skamarock, J. B. Klemp, L. D. Fowler, and M. G. Duda, 2013: Evaluation of global atmospheric solvers using extensions of the Jablonowski and Williamson baroclinic wave test case. *Mon. Wea. Rev.*, **141**, 3116–3129, doi:10.1175/MWR-D-12-00096.1.

——, J. B. Klemp, and W. C. Skamarock, 2014: A comparison of mesh refinement in the global MPAS-A and WRF models using an idealized normal-mode baroclinic wave simulation. *Mon. Wea. Rev.*, **142**, 3614–3634, doi:10.1175/MWR-D-14-00004.1.

Pollard, R. T., P. B. Rhines, and R. O. R. Y. Thompson, 1973: The deepening of the wind-mixed layer. *Geophys. Fluid Dyn.*, **3**, 381–404.

Raymond, D. J., C. S. Bretherton, and J. Molinari, 2006: Dynamics of the intertropical convergence zone of the east Pacific. *J. Atmos. Sci.*, **63**, 582–597, doi:10.1175/JAS3642.1.

Skamarock, W. C., J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, and T. D. Ringler, 2012: A multiscale nonhydrostatic atmospheric

model using centroidal Voronoi tesselations and C-grid staggering. *Mon. Wea. Rev.*, **140**, 3090–3105, doi:10.1175/MWR-D-11-00215.1.

——, S.-H. Park, J. B. Klemp, and C. Snyder, 2014: Atmospheric kinetic energy spectra from global high-resolution nonhydrostatic simulations. *J. Atmos. Sci.*, **71**, 4369–4381, doi:10.1175/JAS-D-14-0114.1.

Torn, R. D., and C. A. Davis, 2012: The influence of shallow convection on tropical cyclone track forecasts. *Mon. Wea. Rev.*, **140**, 2188–2197, doi:10.1175/MWR-D-11-00246.1.

Vigh, J., 2015: Tropical cyclone guidance project. Accessed May 2015. [Available online at http://www.ral.ucar.edu/guidance/.]

Vitart, F., J. L. Anderson, and W. F. Stern, 1997: Simulation of the interannual variability of tropical storm frequency in an ensemble of GCM integrations. *J. Climate*, **10**, 745–760, doi:10.1175/1520-0442(1997)010<0745:SOIVOT>2.0.CO;2.

Yamaguchi, M., F. Vitart, S. T. K. Lang, L. Magnusson, R. L. Elsberry, G. Elliott, M. Kyouda, and T. Nakazawa, 2015: Global distribution of the skill of tropical cyclone activity forecasts on short- to medium-range time scales. *Wea. Forecasting*, **30**, 1695–1709, doi:10.1175/WAF-D-14-00136.1.

Zarzycki, C. M., and C. Jablonowski, 2015: Experimental tropical cyclone forecasts using a variable-resolution global model. *Mon. Wea. Rev.*, **143**, 4012–4037, doi:10.1175/MWR-D-15-0159.1.

Zhang, C., Y. Wang, and K. Hamilton, 2011: Improved representation of boundary layer clouds over the Southeast Pacific in ARW-WRF using a modified Tiedtke cumulus parameterization scheme. *Mon. Wea. Rev.*, **139**, 3489–3513, doi:10.1175/MWR-D-10-05091.1.