# Diagnosing Conditions Associated with Large Intensity Forecast Errors in the Hurricane Weather Research and Forecasting (HWRF) Model

DANIEL J. HALPERIN[a] AND RYAN D. TORN

*Department of Atmospheric and Environmental Science, University at Albany, State University of New York, Albany, New York*

(Manuscript received 14 June 2017, in final form 20 December 2017)

ABSTRACT

Understanding and forecasting tropical cyclone (TC) intensity change continues to be a paramount challenge for the research and operational communities, partly because of inherent systematic biases contained in model guidance, which can be difficult to diagnose. The purpose of this paper is to present a method to identify such systematic biases by comparing forecasts characterized by large intensity errors with analog forecasts that exhibit small intensity errors. The methodology is applied to the 2015 version of the Hurricane Weather Research and Forecasting (HWRF) Model retrospective forecasts in the North Atlantic (NATL) and eastern North Pacific (EPAC) basins during 2011–14. Forecasts with large 24-h intensity errors are defined to be in the top 15% of all cases in the distribution that underforecast intensity. These forecasts are compared to analog forecasts taken from the bottom 50% of the error distribution. Analog forecasts are identified by finding the case that has 0–24-h intensity and wind shear magnitude time series that are similar to the large intensity error forecasts. Composite differences of the large and small intensity error forecasts reveal that the EPAC large error forecasts have weaker reflectivity and vertical motion near the TC inner core from 3 h onward. Results over the NATL are less clear, with the significant differences between the large and small error forecasts occurring radially outward from the TC core. Though applied to TCs, this analog methodology could be useful for diagnosing systematic model biases in other applications.

## 1. Introduction

Tropical cyclone (TC) intensity forecasting remains a vexing challenge as a result of the multiscale nature of the processes that affect TC intensity change and an incomplete understanding of how these factors interact. Statistical–dynamical models such as the Statistical Hurricane Intensity Prediction System (DeMaria et al. 2005) highlight the importance of large-scale environmental factors on TC intensity change, such as vertical wind shear, midlevel moisture, and upper-level divergence. Moreover, Hendricks et al. (2010) suggested that intensification rates may be controlled by smaller-scale processes, assuming that favorable large-scale environmental conditions exist. Indeed, several studies have discussed the importance of symmetric inner-core deep convection in intensifying TCs. Jiang (2012) found a statistically significant relationship between the intensity of inner-core convection as measured by 11 years' worth of TRMM data and TC intensification rates. Rogers et al. (2013) used airborne Doppler radar data to examine the structural differences between intensifying and steady-state TCs. They determined that intensifying (steady state) TCs are characterized by convective bursts that are preferentially located within (outside) the radius of maximum wind (RMW). Furthermore, intensifying TCs were found to exhibit deeper, stronger inflow layers and stronger upward motion in the eyewall. More recently, Rogers et al. (2016) conducted a case study of Hurricane Edouard (2014) and compared a time period of intensification to the period of peak intensity. The period of intensification was marked by deep convection at radii closer to the TC center, stronger low-level convergence, and greater CAPE compared to the peak intensity period.

Despite these gains in knowledge, operational TC intensity forecast skill has shown small incremental improvement over the past two decades, particularly for <48 h, despite notable improvements in TC track

[a] Current affiliation: Department of Applied Aviation Sciences, Embry-Riddle Aeronautical University, Daytona Beach, Florida.

*Corresponding author*: Daniel J. Halperin, daniel.halperin@erau.edu

DOI: 10.1175/WAF-D-17-0077.1

forecasting (NHC 2016). Improvements in TC forecasting are generally attributed to better operational statistical and/or dynamical forecast guidance products being available to forecasters. (Rappaport et al. 2012). The Hurricane Weather Research and Forecasting (HWRF) Model (Tallapragada et al. 2016) has become one of the primary TC intensity forecast guidance sources for the operational community. Like all guidance products, it can exhibit large intensity forecast errors. Numerous studies have been conducted using HWRF (e.g., Gopalakrishnan et al. 2011, 2012; Bao et al. 2012; Gopalakrishnan et al. 2013; Tallapragada et al. 2014; Chen and Gopalakrishnan 2015; Zhang and Marks 2015) to examine the impact of various parameterizations on intensity forecast errors and predictability. While numerous, these studies typically involve either idealized studies or real-data studies applied to a limited number of cases; therefore, the results may not be generally applicable to larger sets of cases.

The goal of this study is to introduce a methodology for diagnosing the potential mechanisms associated with large intensity forecast errors. Given its prominence in the literature, the emphasis of this paper is on investigating whether differences in the intensity of convection in HWRF are related to large forecast intensity errors. Systematic biases are obtained by comparing composites of forecasts with large errors with a set of analog forecasts that exhibit a similar intensity and shear magnitude evolution, but which are associated with small intensity forecast errors. These comparisons are conducted within a shear-relative reference framework to preserve the structural asymmetries that typically are observed for TCs in shear. For example, increased (decreased) convection typically is observed in the downshear (upshear) quadrants of TCs (e.g., Corbosiero and Molinari 2002, 2003; Reasor et al. 2013; DeHart et al. 2014; Hazelton et al. 2015; Rios-Berrios et al. 2016a,b).

This approach will yield more meaningful results when a large database of forecasts with the same model configuration is available; here, the 2015 configuration of HWRF is used. In many ways, this approach is similar to other studies that have demonstrated the use of reforecast datasets in bias correction and model diagnostics (e.g., Hamill et al. 2004, 2006, 2013; Hamill and Whitaker 2006; Galarneau and Hamill 2015). While applying these results to a real-time intensity guidance product may be of interest and utility, it is beyond the scope of this paper. Here, the main goal is to present the analog framework. Ideally, model developers will find utility in applying this methodology to identifying model biases.

The remainder of the paper is structured as follows. Section 2 describes the composite analog comparison methodology. Section 3 discusses the application of the methodology to reforecasts of the 2015 HWRF over the North Atlantic (NATL) and eastern North Pacific (EPAC) basins. A summary and conclusions are provided in section 4.

## 2. Methodology

Forecast performance is evaluated for the 2015 configuration of HWRF (H215; Tallapragada et al. 2016). Prior to the start of the 2015 hurricane season, reforecasts of all 2011–14 TCs were generated from the 2015 model configuration, which provided a set of >900 initialization times for each of the NATL and EPAC basins. The forecasts were separated by basin because of the differences in climatological shear, TC size, and HWRF configuration. The focus of this paper is on 24-h intensity forecast errors because that is a time period where intensity forecast errors have exhibited minimal change over time (NHC 2016) and where it is likely that improvements to the model, including the initial conditions, can result in more accurate intensity forecasts before large-scale model biases may become dominant. It is worth pointing out that this approach could theoretically be applied to intensity forecasts at any lead time.

All intensity forecasts were verified against National Hurricane Center best-track data (Jarvinen et al. 1984; McAdie et al. 2009). The cumulative distribution of all 24-h intensity forecast errors for each basin is presented in Figs. 1a and 1b. Over the EPAC, 66% of H215 forecasts underintensified the TC compared to the best track (i.e., negative intensity forecast errors), where the errors range from −64 kt (i.e., the model forecast is 64 kt weaker than observed; 1 kt = 0.51 m s$^{-1}$) to +37 kt, which suggests that the larger-magnitude errors are overwhelming negative errors. Moreover, 0.5% of the forecasts overintensify the TC by at least 25 kt, while 9% of the forecasts underintensify the TC by more than 25 kt. This result may be due to the inability of HWRF to capture the large number of rapid intensification (RI) events within this basin during the 2011–14 reforecast period. For the NATL, 56% of the forecasts exhibit underintensification. The larger magnitude errors also are negative errors over the NATL, ranging from −41 to +30 kt, while 0.9% (3.5%) of forecasts overintensify (underintensify) the TC by at least 25 kt.

To narrow the focus of the present study and to test the usefulness of the analog methodology, a subset of forecasts were selected. First, all forecasts within 50 km of land during the 0–24-h forecast period were removed to ensure that intensity errors were not due to land interaction. The forecasts in the top 15% of the 24-h intensity (VMAX) mean absolute error distribution

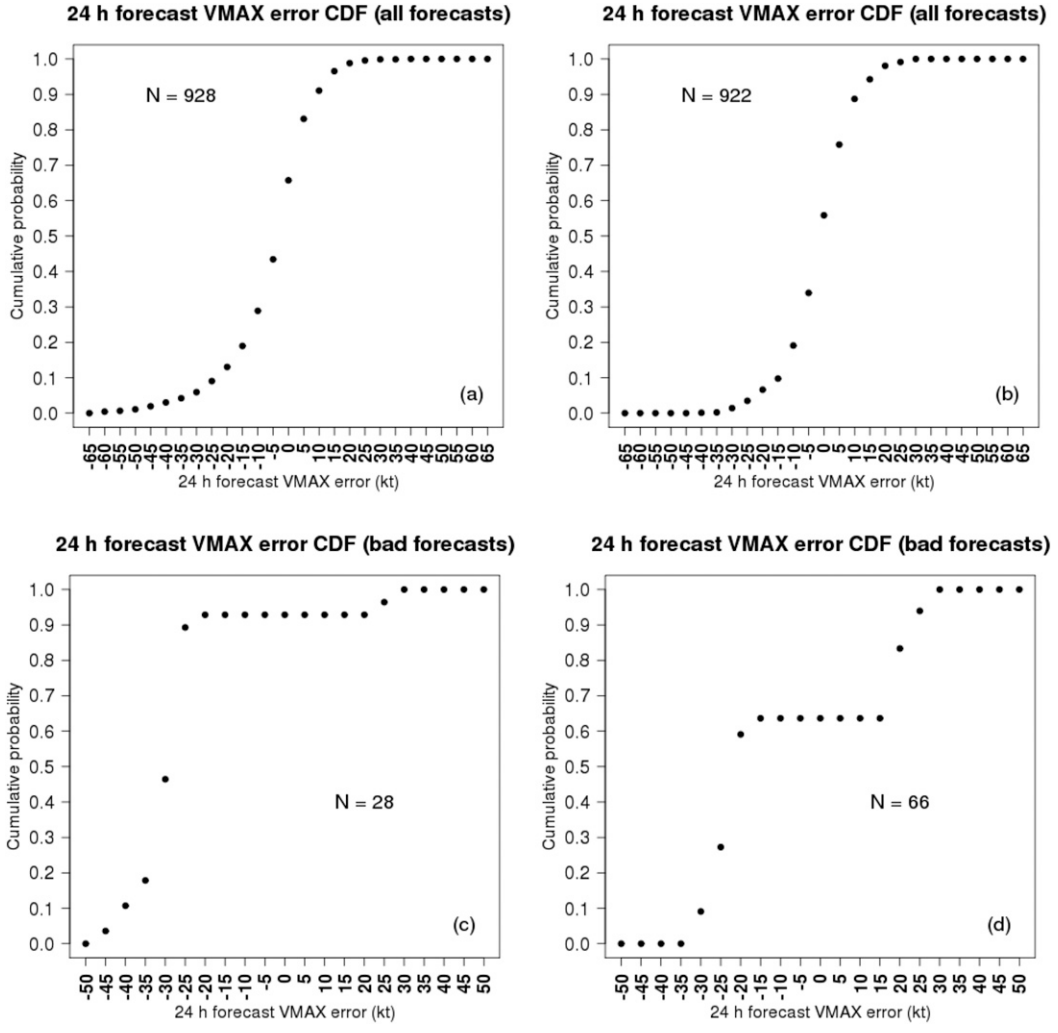



FIG. 1. CDF of the 24-h VMAX error for (a) all H215 forecasts over the EPAC, (b) all H215 forecasts over the NATL, (c) the final set of bad forecasts over the EPAC, and (d) the final set of bad forecasts over the NATL.

were considered the ''bad'' forecasts ($N = 98$ over the EPAC; $N = 92$ over the NATL). For each bad forecast, a potential analog forecast is selected from a pool of ''good'' forecasts in the bottom 50% of the 24-h VMAX mean absolute error distribution ($N = 437$ over the EPAC; $N = 396$ over the NATL). These percentiles are somewhat arbitrary, but were necessary to produce a sufficient sample size to identify a suitable analog.

Additional filtering criteria are also employed to remove situations that are associated with large errors, but are outside the scope of this study. All forecasts that occur during rapid intensity changes [$\pm 30\,\text{kt}\,(24\,\text{h})^{-1}$] in the best track were removed. This reduced the sample size to 42 (68) bad and 389 (388) good forecasts over the EPAC (NATL). The large reduction in cases, especially over the EPAC, indicates that many of the bad forecasts were cases where H215 failed to capture rapid intensity changes. While these cases are important to understand, it is well known that models have difficulty predicting RI (e.g., Gopalakrishnan et al. 2011; Chen and Gopalakrishnan 2015). Indeed, the focus here is on non-RI cases, which might be expected to exhibit smaller errors. Finally, TCs that move over SSTs $< 25.5$°C during the 0–24-h forecast period also are removed. This criterion removed cases characterized by track errors that yielded intensity differences when the TC moves over the strong EPAC SST gradient, yielding a sample size of 28 (66) bad and 236 (276) good forecasts for the EPAC (NATL). Moreover, this criterion also provides an objective method of removing cases where TCs may be undergoing extratropical transition in the NATL—a process that contains timing uncertainties and subjectivity in identification (e.g., Zarzycki et al. 2016). Cumulative distribution functions

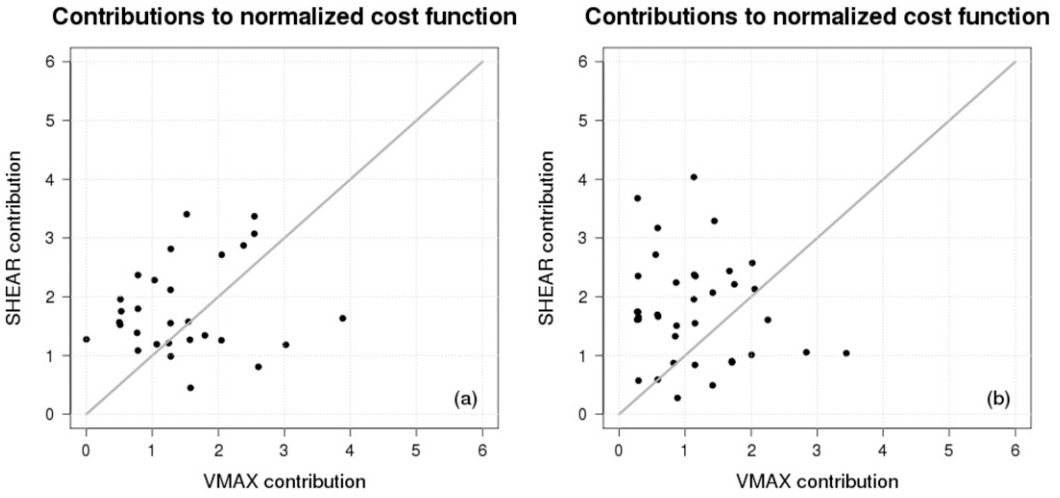



FIG. 2. The VMAX vs shear difference contributions to the analog cost function expressed in normalized units for (a) EPAC and (b) NATL forecasts.

(CDFs) of the 24-h intensity error indicate that nearly 90% (65%) of the remaining EPAC (NATL) bad forecasts are cases where HWRF underintensifies the TC compared to the best track (Figs. 1c,d). As a consequence, the focus of this study is on these forecasts.

It is also desirable to compare each bad forecast to a good forecast with some similar characteristics to help identify any systematic differences between the two. This analog good forecast should have a best-track intensity time series that is similar to those of the bad forecasts; otherwise, it is possible that the composite differences will mainly reflect intensity differences. In addition, the analog forecast should have a similar shear evolution since shear magnitude is a strong intensity predictor (e.g., DeMaria et al. 2005; DeMaria 2009). As a consequence, a cost function was developed to identify the best analog for each bad forecast, which is defined as

$$J_{i,k} = \sum_{t=0}^{24} \left\| \frac{\mathrm{VMAX}_{\mathrm{bad}_{i,t}} - \mathrm{VMAX}_{\mathrm{good}_{k,t}}}{\sigma_{\mathrm{VMAX}_{\mathrm{good}_t}}} \right\| + \sum_{t=0}^{24} \left\| \frac{\mathrm{SHEAR}_{\mathrm{bad}_{i,t}} - \mathrm{SHEAR}_{\mathrm{good}_{k,t}}}{\sigma_{\mathrm{SHEAR}_{\mathrm{good}_t}}} \right\|, \quad (1)$$

where VMAX refers to the best-track maximum wind speed; $i$ and $k$ denote the index for bad and good cases, respectively; $t$ refers to lead time; SHEAR denotes the magnitude of vertical wind shear; $\sigma$ is the standard deviation of that quantity over all cases, which is used to create an equivalent weight between these two metrics; the subscript bad refers to a particular bad forecast; and the subscript good refers to the set of HWRF forecasts in the bottom 50% of the 24-h VMAX error distribution (i.e., the pool of potential analog forecasts). For each bad forecast, its analog good forecast is the case that minimizes $J$. To ensure similarity between bad and analog good cases, the bad forecast's analog had to have $J < 6$ to be included in the composite analysis. This seemingly arbitrary threshold was selected to remove bad forecasts that did not have any good analogs while ensuring a sufficient sample size is available to conduct this analysis. Here, $J = 6$ corresponds to an average difference in VMAX and SHEAR of 6.5 kt at each 6-h time step. Figure 2 shows the contributions of each component to the total cost function. In both basins, the shear differences between the bad and analog forecasts generally contribute more to the cost function than do VMAX differences, implying that it is more difficult to match the time evolution of shear magnitude compared to maximum wind speed.

The final sample size of forecasts for the EPAC (NATL) basin is 26 (30). Figure 3 shows the best-track VMAX and SHIPS shear magnitude time series for each forecast included in the composite analysis, with one random bad–analog pair set in bold. The TCs span a wide range of intensities over both basins, from minimal TCs to category 4 hurricanes. The shear magnitude values are mainly in the 0–15-kt (5–20 kt) range for the EPAC (NATL), which is low-to-moderate shear (e.g., Rios-Berrios and Torn 2017).

In addition to VMAX, the RMW can be another important metric for determining the structure of a TC. To determine the sensitivity of the analog methodology and the composite results to consideration of TC size, the above process was repeated using a cost function similar to Eq. (1) that includes the RMW:

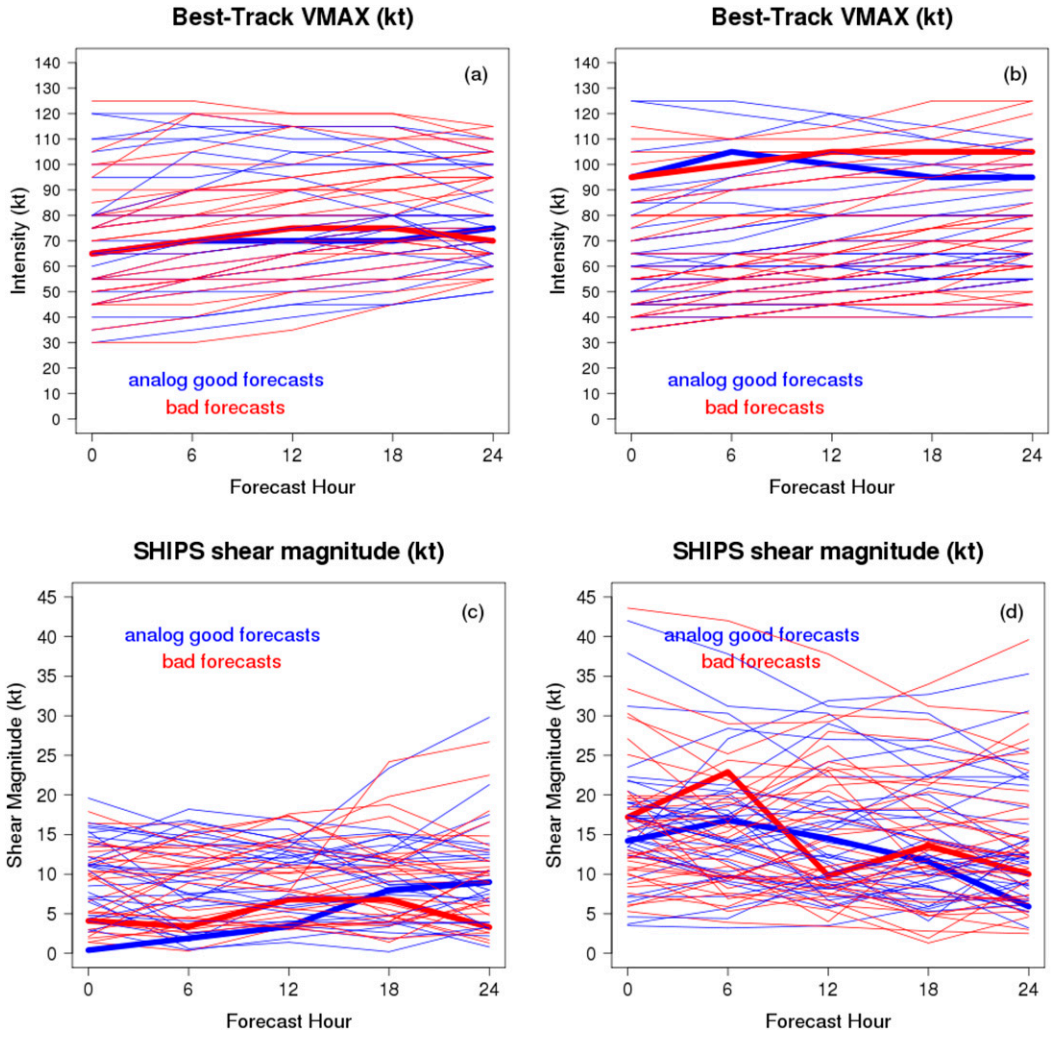



FIG. 3. Best-track intensity time series of the analog good and bad forecasts for the (a) EPAC and (b) NATL basins. SHIPS shear magnitude time series of the analog good and bad forecasts for the (c) EPAC and (d) NATL basins. Bold lines denote a randomly selected bad and good analog pair.

$$J_{i,k} = \sum_{t=0}^{24} \left\| \frac{\mathrm{VMAX}_{\mathrm{bad}_{i,t}} - \mathrm{VMAX}_{\mathrm{good}_{k,t}}}{\sigma_{\mathrm{VMAX}_{\mathrm{good}_t}}} \right\| + \sum_{t=0}^{24} \left\| \frac{\mathrm{SHEAR}_{\mathrm{bad}_{i,t}} - \mathrm{SHEAR}_{\mathrm{good}_{k,t}}}{\sigma_{\mathrm{SHEAR}_{\mathrm{good}_t}}} \right\| + \sum_{t=0}^{24} \left\| \frac{\mathrm{RMW}_{\mathrm{bad}_{i,t}} - \mathrm{RMW}_{\mathrm{good}_{k,t}}}{\sigma_{\mathrm{RMW}_{\mathrm{good}_t}}} \right\|. \quad (2)$$

This cost function yields a final sample size of 26 (29) forecasts for the EPAC (NATL) basin. It is worth noting that 65% (50%) of the bad and analog cases identified in the EPAC (NATL) basin using Eq. (2) (i.e., with RMW included the cost function) were also identified using Eq. (1) (i.e., without RMW included in the cost function). Composites of the cases identified in both cost functions are provided in section 3 to illustrate the sensitivity to including RMW in the cost function.

## 3. Results

Recall that the bad forecasts are characterized by the HWRF forecasts being too weak relative to the best track. One hypothesis is that the bad forecasts have weaker inner-core convection compared to the analog forecasts, which in turn yields weaker TCs. Further, it is hypothesized that weaker surface fluxes may lead to the lack of inner-core convection. This hypothesis is tested by computing the normalized difference (e.g., Torn et al. 2015; Lamberson et al. 2016) for convection-related

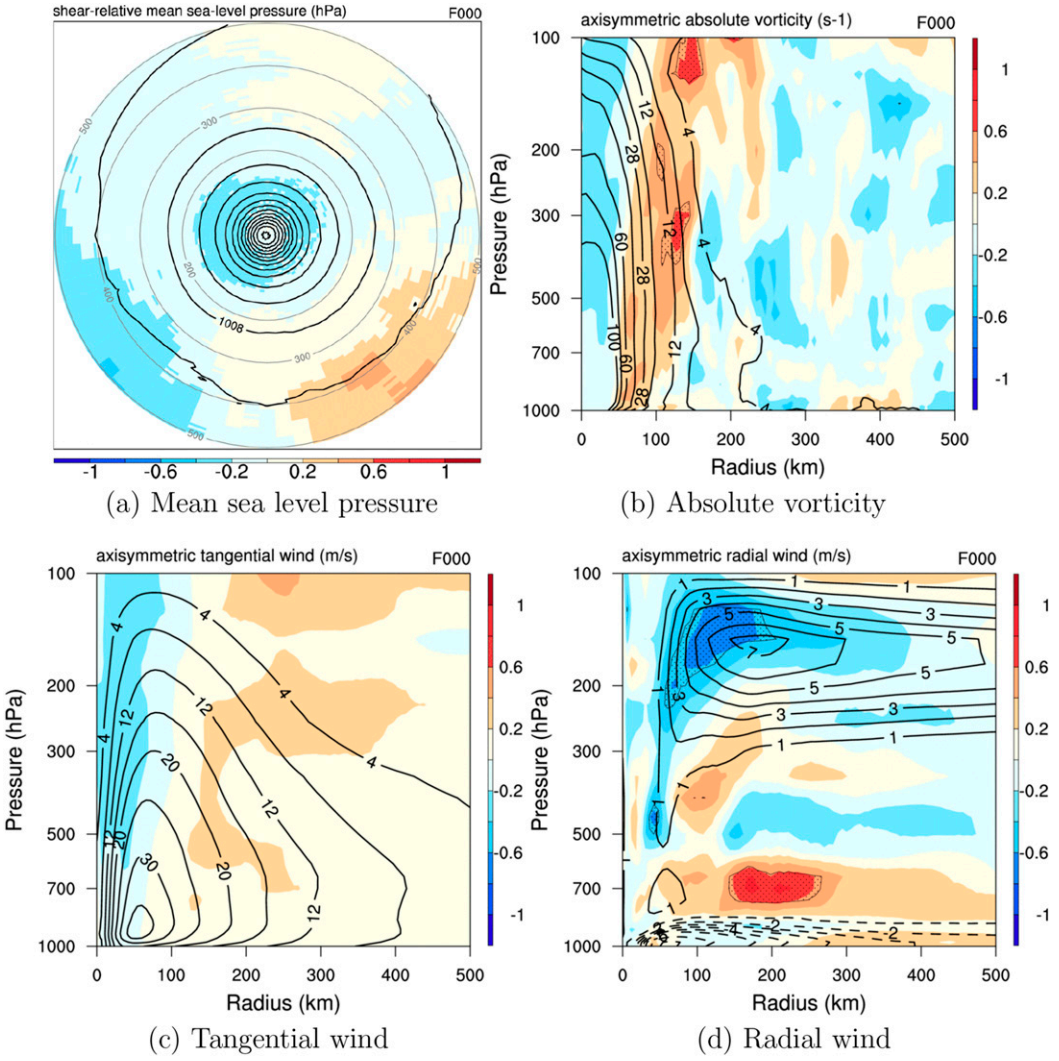



FIG. 4. Normalized composite difference in 0-h (a) MSLP (hPa), (b) axisymmetric absolute vorticity ($10^{-5}\,s^{-1}$), (c) tangential wind ($m\,s^{-1}$), and (d) radial wind ($m\,s^{-1}$) between the bad forecasts and analog good forecasts within the EPAC basin (shading). Cool (warm) colors denote areas where the bad forecasts had smaller (larger) values of the field analyzed. Stippled regions denote where the difference is statistically significant at the 95% confidence interval. Black contours denote the composite bad forecast values. For (a), the differences are computed in a reference frame relative to the 200–850-hPa shear vector, where the shear is oriented toward the top of the page, and the gray lines denote range rings from the TC center (km).

fields between bad and analog forecasts. Lamberson et al. (2016) provide justification for the use of normalized differences, including straightforward comparisons of the differences between the two subsets of forecasts across fields, space, and time. In all cases, these normalized differences are constructed within a 200–850-hPa shear-relative reference frame to preserve structural asymmetries. The radius–height plots (e.g., Fig. 4b) are averaged around a particular azimuth range. The statistical significance of the difference is evaluated using bootstrap resampling (5000 random resamples of all cases at the 95% confidence interval).

### *a. Eastern North Pacific*

The focus of the present study is on bad forecasts where HWRF underintensifies TCs, and the analog forecasts have similar best-track intensity time series to the bad forecasts. Thus, it is expected that the composite differences in fields related to intensity should be small at 0 h, but the bad forecast composite should be weaker than the analog composite by 24 h. At 0 h, there are only small areas of statistically significant differences in MSLP, 850-hPa relative vorticity, and radial and tangential wind speeds (Fig. 4);

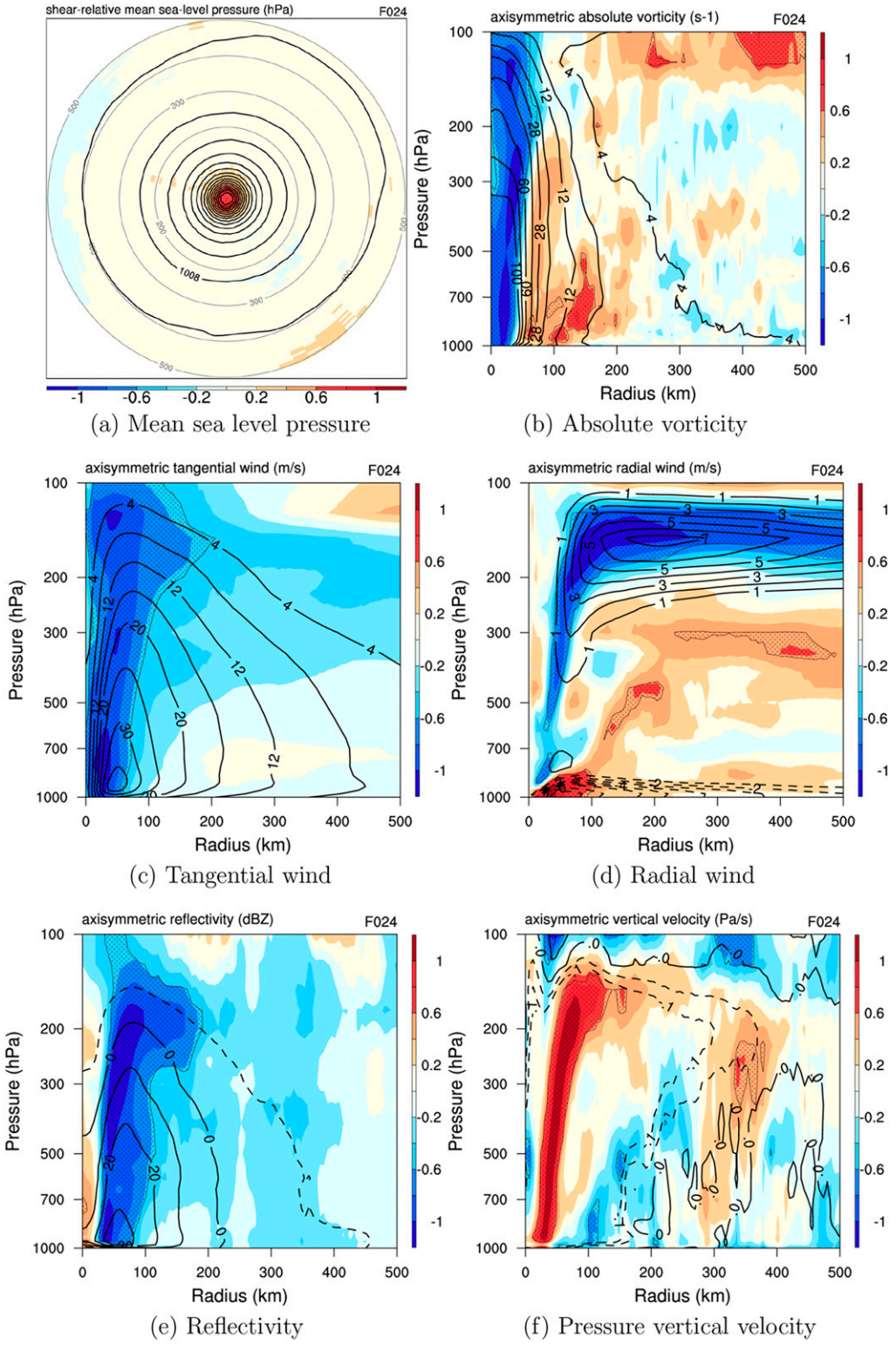



FIG. 5. As in Fig. 4, but for 24 h. Depicted are the (e) axisymmetric reflectivity (dB$Z$) and (f) pressure vertical velocity (Pa s$^{-1}$).

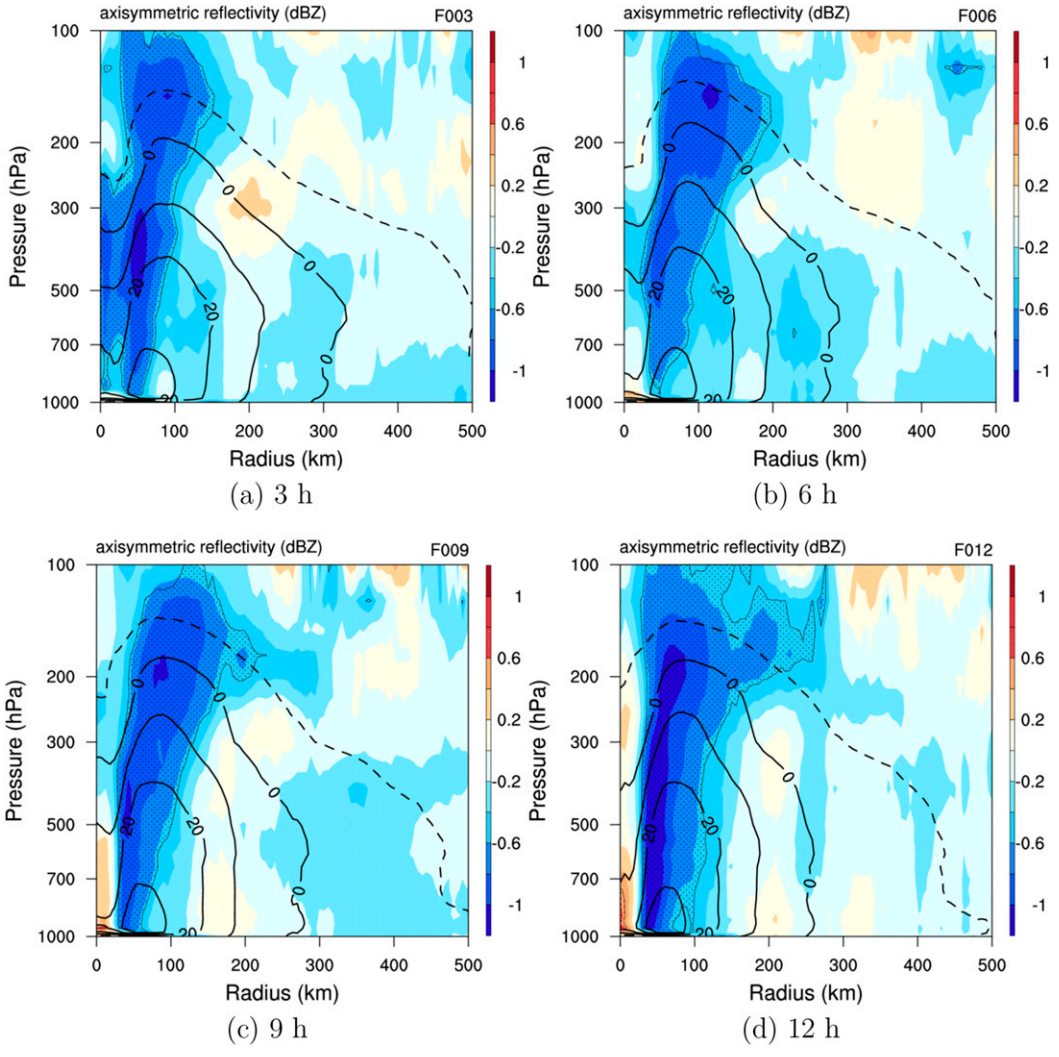



FIG. 6. Normalized composite differences of azimuthally averaged reflectivity at (a) 3, (b) 6, (c) 9, and (d) 12 h (shading) for EPAC forecasts. The contours denote the composite bad forecast values (dBZ). Stippled regions denote where the difference is statistically significant at the 95% confidence interval.

however, more substantive differences, mainly within 100 km of the TC center, are present by 24 h (Fig. 5). The bad forecasts exhibit higher MSLP and smaller 850-hPa absolute vorticity. The tangential wind speeds are also weaker near the TC core for the bad forecasts, and the radial wind profile suggests the bad forecasts have a weaker secondary circulation. In addition, the bad forecasts contain weaker inner-core convection as indicated by the reflectivity and pressure vertical velocity $\omega$ fields. As a consequence, the remainder of this subsection is devoted to examining select physically relevant variables to investigate the hypothesis that the cause of the divergence between bad and analog forecasts during the first 24 h is due to inner-core convection.

A time evolution of reflectivity (Fig. 6) indicates that the weaker inner-core convection in the bad forecasts is present for much of the forecast period throughout the entire depth of the troposphere.[1] The largest axisymmetric differences initially occur at 50-km radius in the 300–500-hPa layer, implying that convection in the bad forecasts does not reach as high in the troposphere near the core. Over time, the differences become more pronounced and extend closer to the surface. In addition, the weaker convection in the bad forecasts extends outward to 200-km radius above 300 hPa by 12 h,

[1] This field is first available at 3 h.

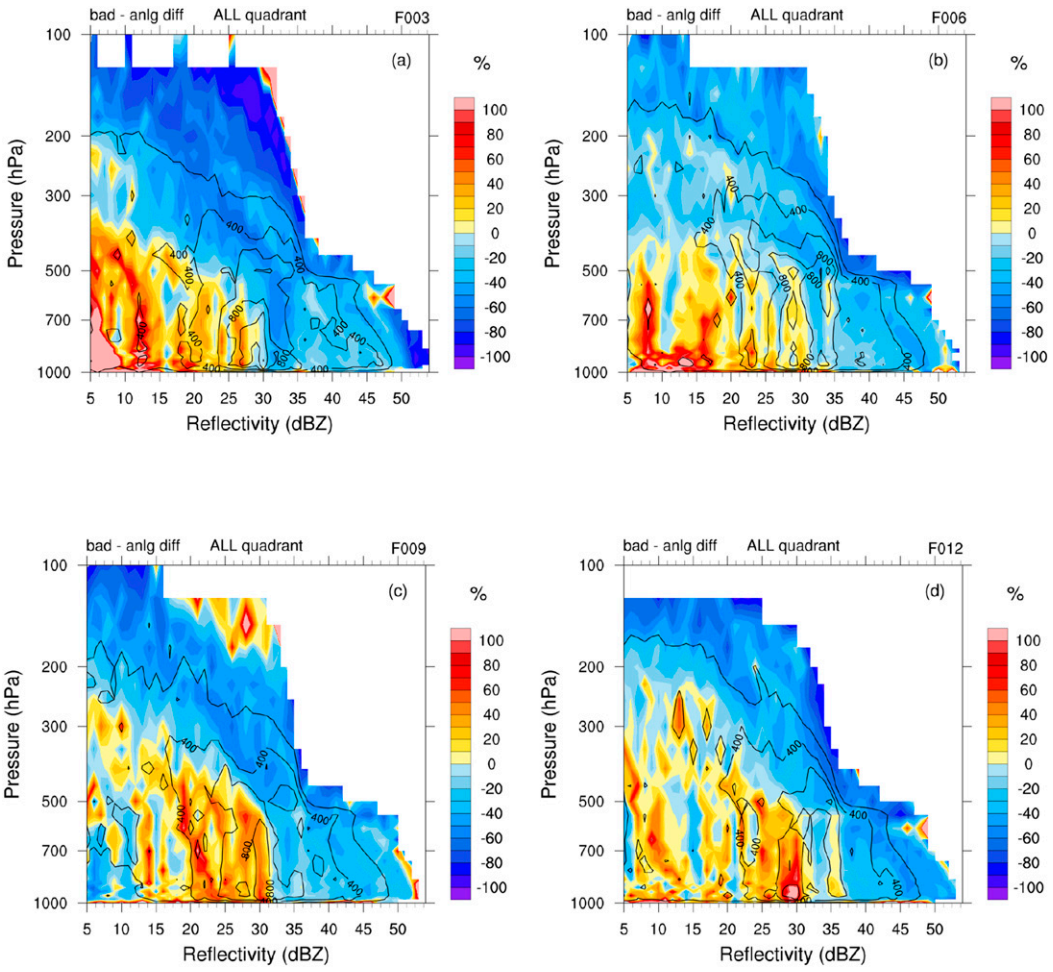



FIG. 7. CFADs of percentage differences in reflectivity (shaded) at (a) 3, (b) 6, (c) 9, and (d) 12 h for EPAC forecasts. Cool (warm) colors denote areas where the bad forecasts had fewer (greater) counts of reflectivity compared to the analog forecasts. Black contours denote the CFAD counts of the composite bad forecast. Only data within 2 × RMW of the center are included.

implying that the convection also is not as deep outside of the TC core.

An examination of the reflectivity profiles in various shear-relative quadrants reveals that at 3 h, significant differences exist in all quadrants above 500 hPa, but mainly extend closer to the surface in the upshear quadrants, especially upshear left (not shown). By 6 h, the largest magnitude and area of significant differences exists in the downshear-right quadrant, suggesting that the bad forecasts exhibit weaker convection in the quadrant that might be expected to have the most active convection (e.g., Corbosiero and Molinari 2002, 2003; Reasor et al. 2013; DeHart et al. 2014; Hazelton et al. 2015; Rios-Berrios et al. 2016a,b). Significant differences in the downshear-left quadrant indicate that the mature convection also is weaker in the bad forecasts (not shown). The differences in reflectivity become more homogeneous across all quadrants by 12 h, similar to Fig. 6d.

Quadrant and axisymmetric averaging can mask differences associated with individual updrafts; therefore, to gain a better understanding of the distribution of the convection near the inner core, reflectivity differences also are shown in the form of contoured frequency by altitude diagrams [CFADs; Yuter and Houze (1995); DeHart et al. (2014)]. The shading indicates the percentage change in a particular reflectivity–height bin for the bad forecasts relative to the analog forecasts. To isolate the inner core, only data at radii within 2 × RMW[2] are included. Figure 7 shows that the bad forecasts generally have greater

[2] The average RMW of TCs at each time period is used.

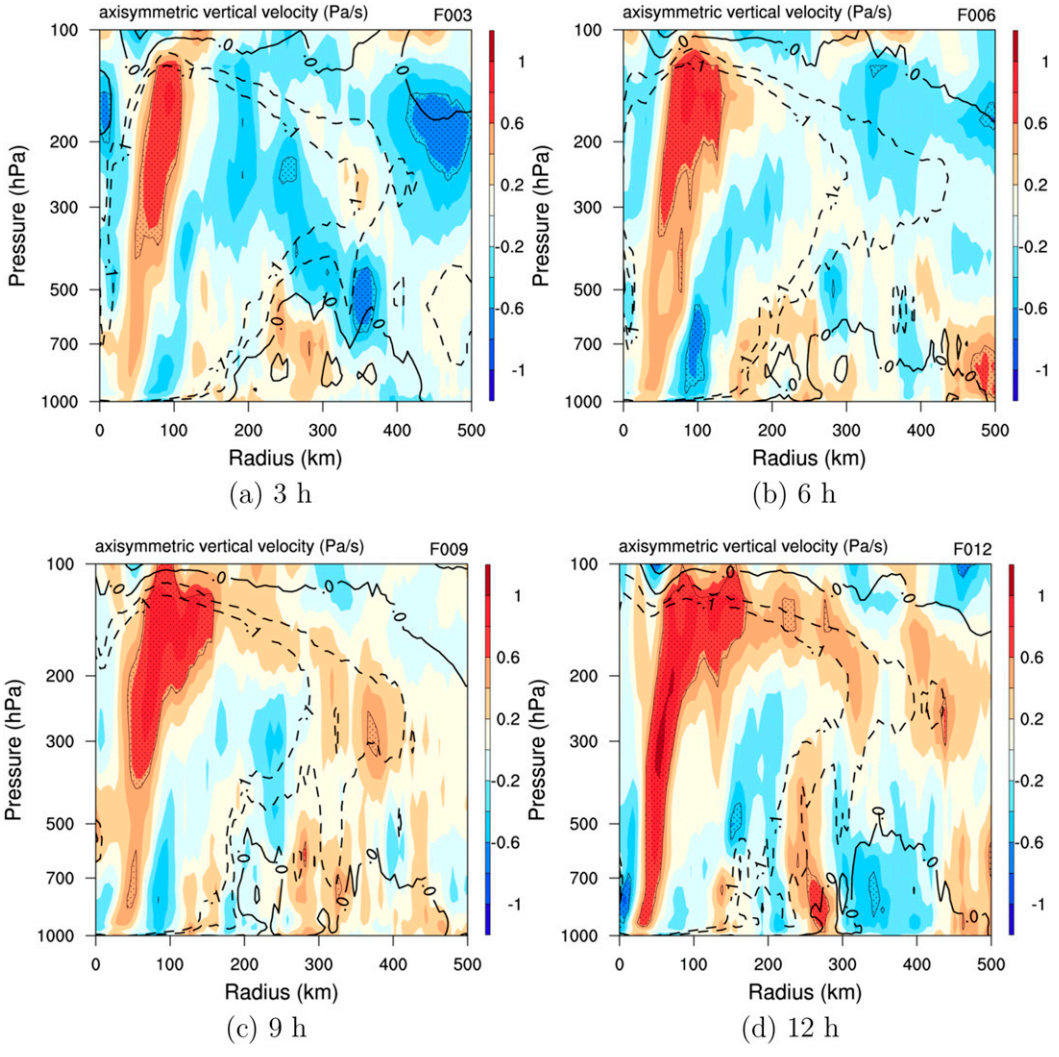



FIG. 8. As in Fig. 6, but for pressure vertical velocity (Pa $s^{-1}$).

counts of reflectivity values less than 30 dB$Z$ at and below 500 hPa, implying that convection in the bad forecasts is weaker and shallower compared to the analog forecasts. Meanwhile, the bad forecasts have fewer counts of reflectivity greater than 35 dB$Z$ at all levels, confirming that the convection is weaker in the bad forecasts. Above 300 hPa, the bad forecasts have fewer counts at nearly all reflectivity bins, suggesting that the convection in the bad forecasts is also shallower compared to the analog forecasts. These results suggest that the bad forecasts have less vigorous convection, which may result in weaker latent heat release and vortex stretching. This result is consistent with the observational study of Reasor et al. (2013), who found that TC intensification typically occurs when deep convection is found near the TC core, within the RMW.

Vertical profiles of $\omega$ present a similar picture (Fig. 8). There are significant differences in inner-core vertical velocity as early as 3 h (first time that this field is available), but they are confined to the upper levels. Bad forecasts exhibit weaker upward motion above 300 hPa between 3 and 9 h, again suggesting that the convection is shallower in the bad forecasts. By 12 h, the weaker upward motion in the bad forecasts extends to the surface, presumably because the bad forecasts become considerably weaker overall than the analog forecasts at this point. CFAD differences also were constructed for $\omega$ (Fig. 9). Again, only data within 2 × RMW are included to highlight inner-core differences. At 3 h, the bad forecasts exhibit fewer counts of stronger updrafts in the mid- to upper levels, consistent with the vertical profiles of mean $\omega$. However, it is interesting to note that the bad forecasts

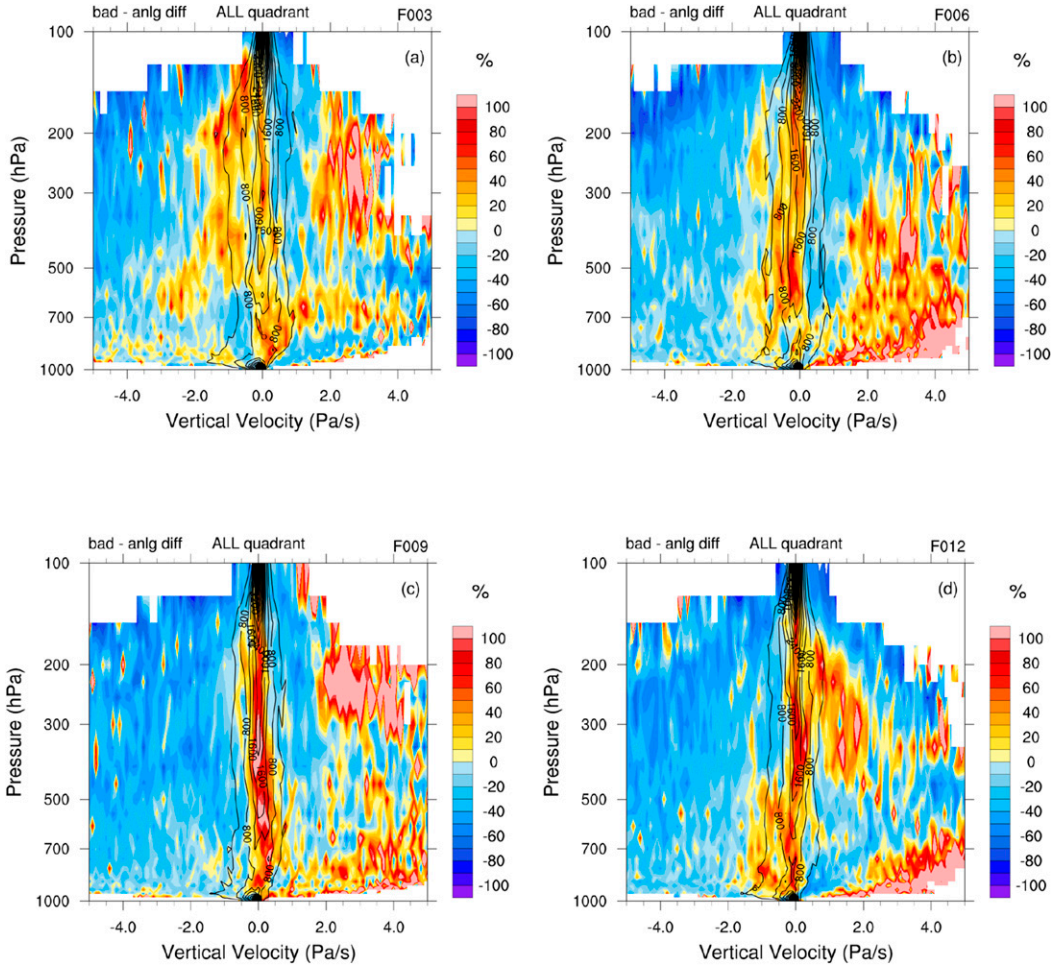



FIG. 9. As in Fig. 7, but for pressure vertical velocity (Pa $s^{-1}$).

have more weak updrafts compared to the analog forecasts. Bad forecasts also have more strong downdrafts, except in the 600–450-hPa layer. Over time, the bad forecasts exhibit fewer updrafts of all magnitudes at most levels compared to the analog forecasts. Meanwhile, the bad forecasts have more strong downdrafts, especially below 700 hPa and in the 200–300-hPa layer. This too suggests that convection is overall weaker and shallower in the bad forecasts, potentially resulting in the observed HWRF underintensification. Since the reflectivity and vertical velocity fields are available beginning at 3 h, it is not known whether these differences exist at the initial time or if they are a result of other structural differences that are present in the model initial conditions. Therefore, it is also unclear whether the weaker upper-level outflow (Fig. 4d) at 0 h in the bad forecasts is a causal or associative factor of the weaker updrafts in the bad forecasts at 3 h.

Another critical factor in the maintenance and intensification of TCs is the surface latent heat flux from the ocean. Weaker surface fluxes yield a near-surface environment less favorable for intensification. Indeed, the bad forecasts exhibit significantly weaker fluxes, especially on the right-of-shear side of the TC at 3 h; however, these differences become largely insignificant over time (Fig. 10). Nevertheless, the weaker surface latent heat fluxes early in the forecast cycle suggest that the near-surface temperature and moisture fields may not be as conducive for extracting energy from the ocean and hence intensification in the bad forecasts. Azimuthally averaged vertical profiles of equivalent potential temperature $\theta_e$ (Fig. 11) therefore may provide insight regarding why the bad forecasts have smaller near-surface wind speeds but statistically equivalent surface latent heat fluxes. Small near-surface $\theta_e$ in the bad forecasts would suggest that the convection is less buoyant, but the thermal disequilibrium is larger.

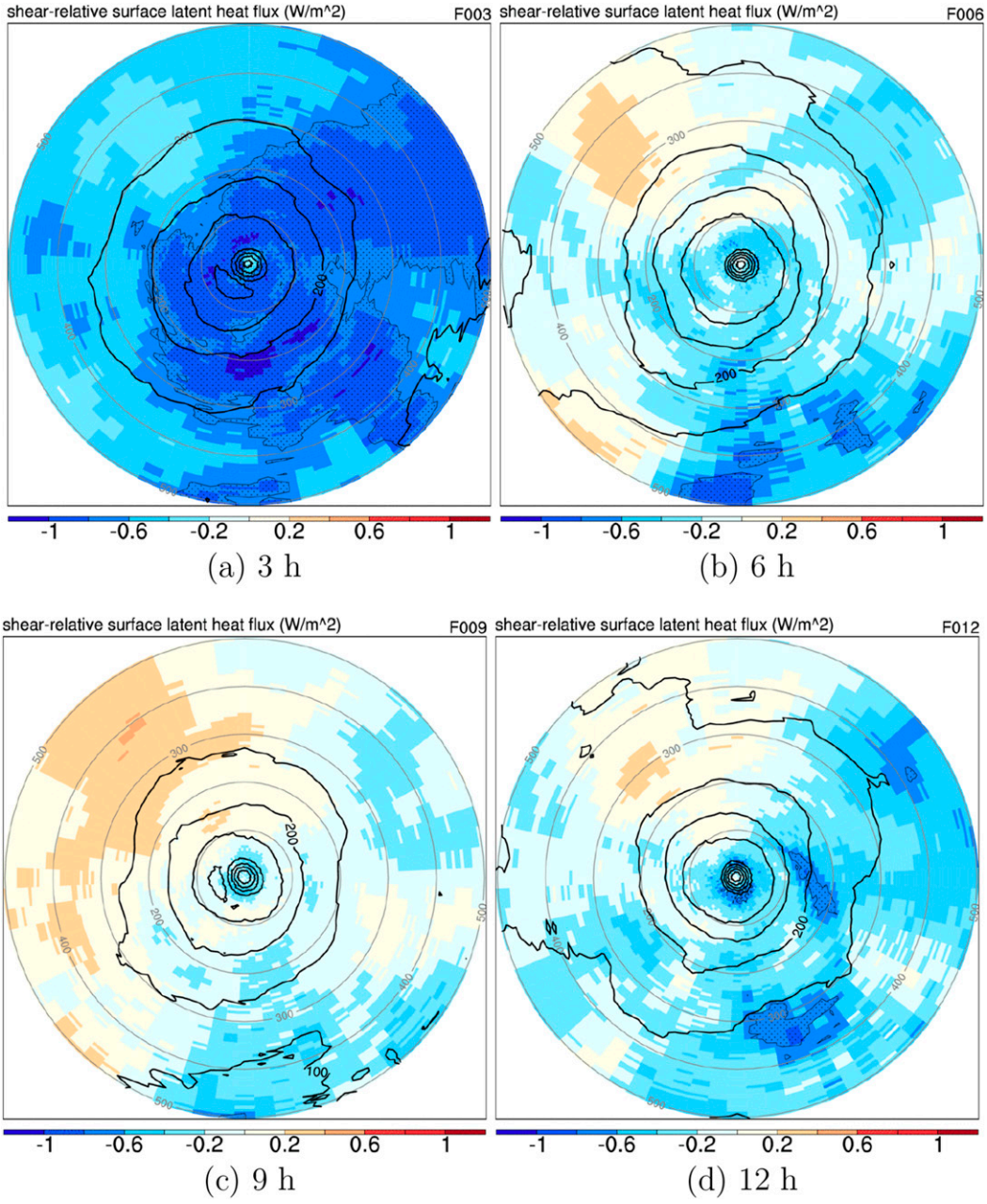



FIG. 10. As in Fig. 4a, but for surface latent heat flux (W $m^{-2}$) at (a) 3, (b) 6, (c) 9, and (d) 12 h.

Indeed, beginning at the initial time, the bad forecasts exhibit significantly smaller values of $\theta_e$ near the surface. This difference persists such that by 24 h, the lower-$\theta_e$ air has been advected by the radial inflow into the TC core. Thus, we hypothesize that the lower-$\theta_e$ air combined with the smaller surface latent heat fluxes in the bad forecasts cause the bad forecasts to exhibit less buoyancy, yielding shallower convection and hence weaker TCs in the forecast. This hypothesis is supported by composite CAPE differences (Fig. 12) that show the bad forecasts exhibiting significantly less CAPE near the TC center at 0 and 6 h. This difference shifts outward to 100–200 km from the TC center at 12 and 24 h. Frequency percentage change plots of CAPE (similar to CFADs but for only one vertical level) also confirm that the bad (analog) forecasts exhibit larger frequencies of CAPE values from 0 to ~1200 J $kg^{-1}$ (from ~1200 to 3600 J $kg^{-1}$) at 0–24 h (Fig. 13). At 6, 12, and 24 h, the bad forecasts exhibit more frequent values of CAPE > 3600 J $kg^{-1}$, but the sample size in these

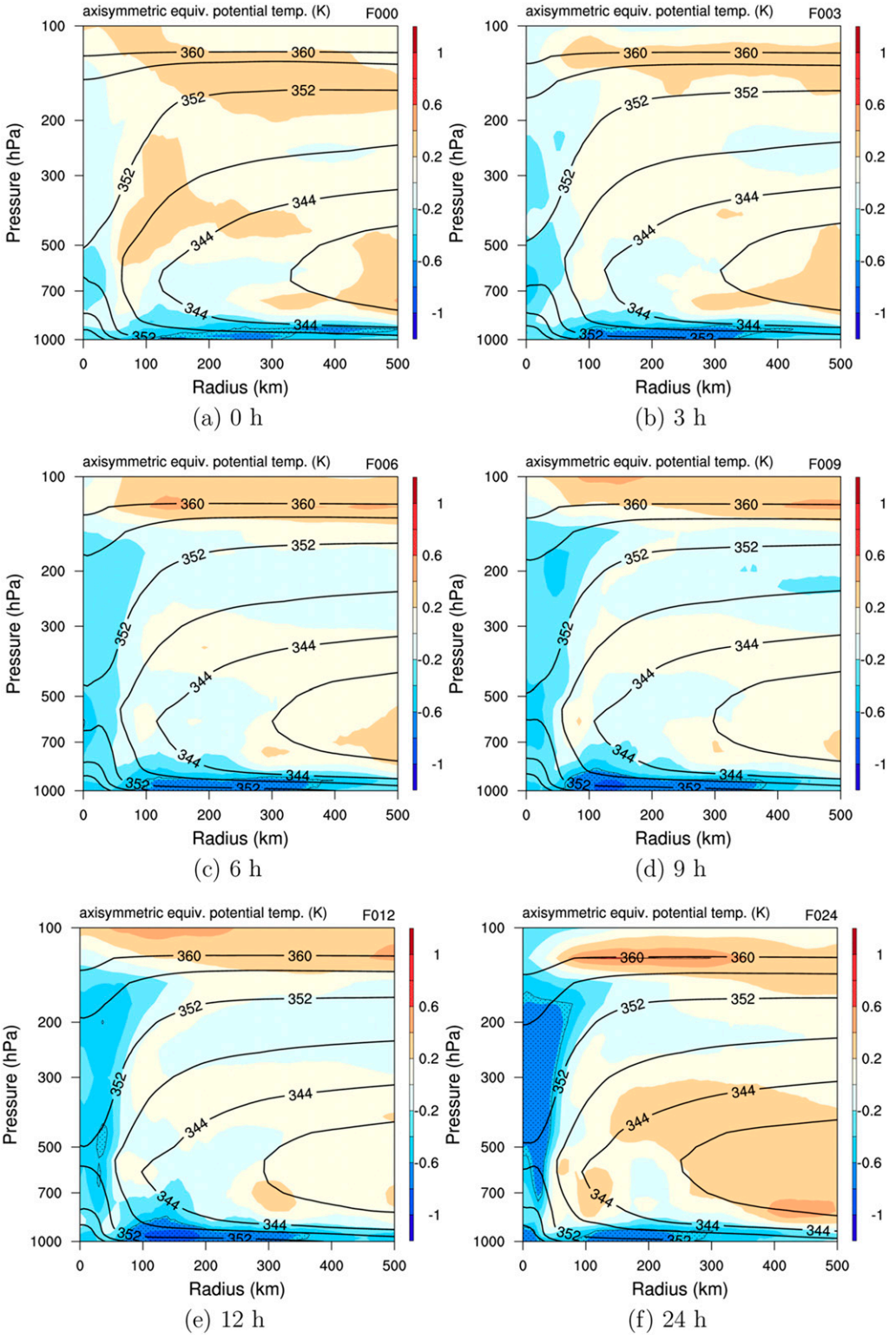



FIG. 11. As in Fig. 6, but for equivalent potential temperature (K) at (a) 0, (b) 3, (c) 6, (d) 9, (e) 12, and (f) 24 h.

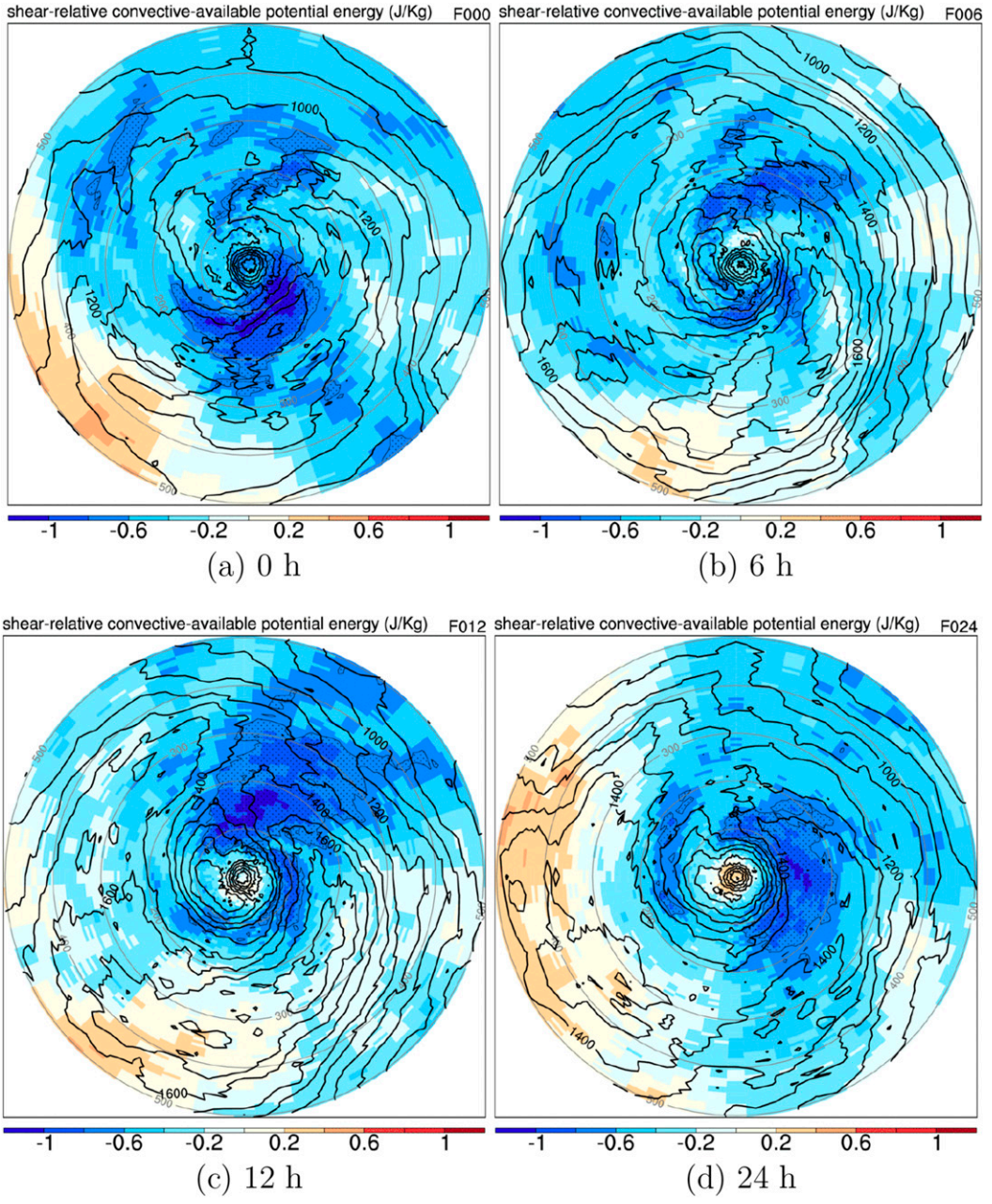



FIG. 12. As in Fig. 4a, but for CAPE (J kg$^{-1}$) at (a) 0, (b) 6, (c) 12, and (d) 24 h.

CAPE bins are orders of magnitude smaller than at CAPE bins $< \sim 3000\,\mathrm{J\,kg^{-1}}$.

As noted in section 2, composites were also created using a cost function that included RMW; overall, the results show little sensitivity to this choice. Figures 14 and 15 are quite similar to Figs. 4 and 5, respectively. Overall, there are few if any significant differences between the bad and analog cases in the TC intensity metrics at 0 h (Fig. 14). However, by 24 h (Fig. 15), the bad forecasts are significantly weaker in terms of both intensity metrics and convection. Furthermore, the surface latent heat flux differences between the two cost functions at 12 and 24 h exhibit few differences (Fig. 16); the bad forecasts have weaker flux values near and upshear of the TC center. The evolutions of the differences in CAPE also are similar for both cost functions

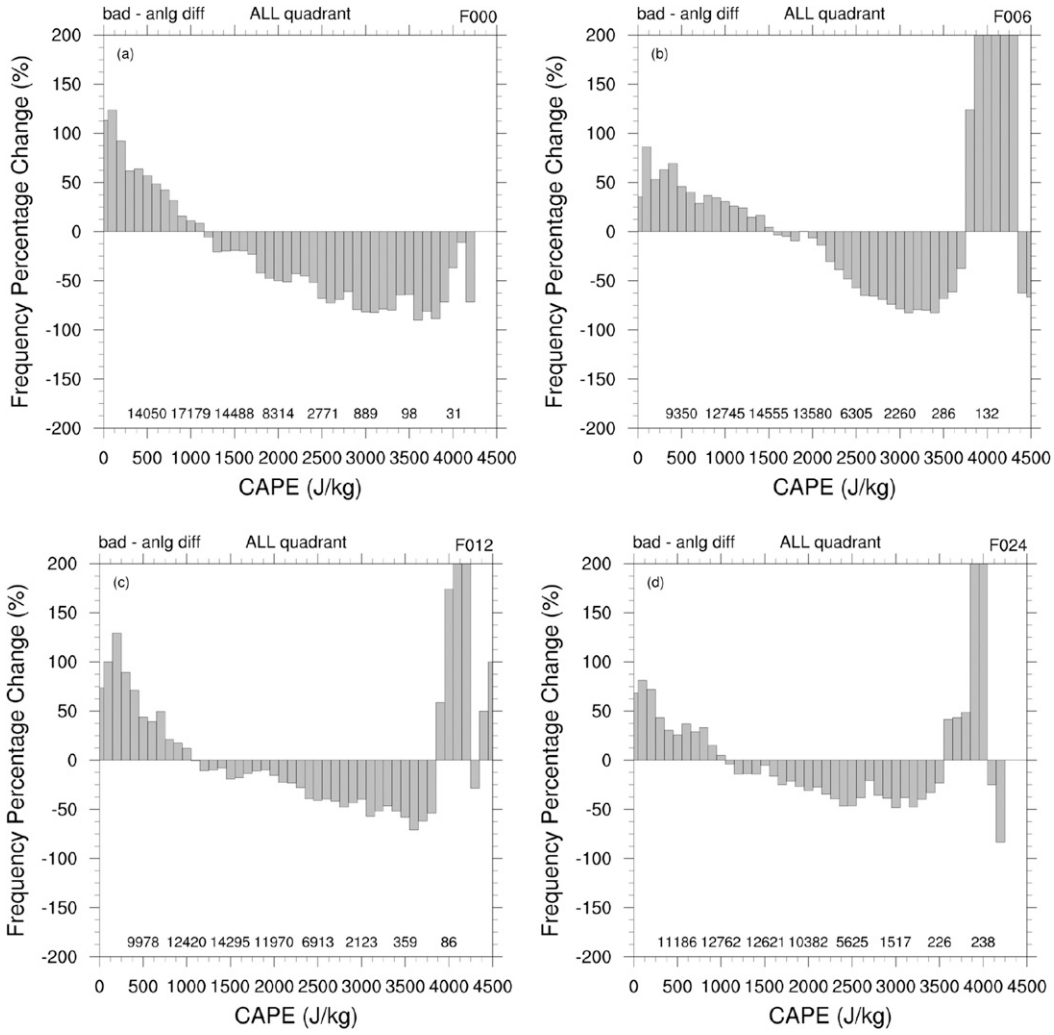



FIG. 13. Frequency percentage difference between the bad and analog good forecast CAPE within 4500-km radius in the EPAC basin at (a) 0, (b) 6, (c) 12, and (d) 24 h. Sample size of all cases in each fifth bin is provided to show the decreased sample at the largest CAPE values.

(Fig. 17), with the bad cases exhibiting smaller CAPE values upshear of the TC center at 0 h, then downshear and right of shear at 12 and 24 h. Given these similarities, it appears that the results are not sensitive to including RMW in the cost function; therefore, the remainder of the paper utilizes the cost function without RMW [i.e., Eq. (1)].

### *b. North Atlantic*

Given the relatively clear differences between the bad and analog good forecasts in the EPAC, it might be expected that bad forecasts over the NATL were also characterized by weaker convection and weaker surface fluxes relative to their analogs. As a consequence, the same atmospheric variables were examined for the NATL forecasts.

As in the EPAC, it is expected that the 0-h fields related to TC intensity (Fig. 18) are similar between the bad and analog forecasts. However, unlike for the EPAC, the NATL bad forecasts exhibit significantly weaker absolute vorticity below 500 hPa between 50- and –150-km radius. In addition, the tangential wind speed also is weaker in the bad forecasts below 600 hPa at radii $> 125$ km, with insignificant differences near the RMW. The secondary circulation also appears weaker for the NATL bad forecasts.

The composite differences over the NATL are quite different compared with those for the EPAC at 24 h (Fig. 19). One striking difference is that, unlike the EPAC, the bulk of the significant differences between NATL bad and analog forecasts occur at larger radii. For example, the bad forecasts exhibit statistically

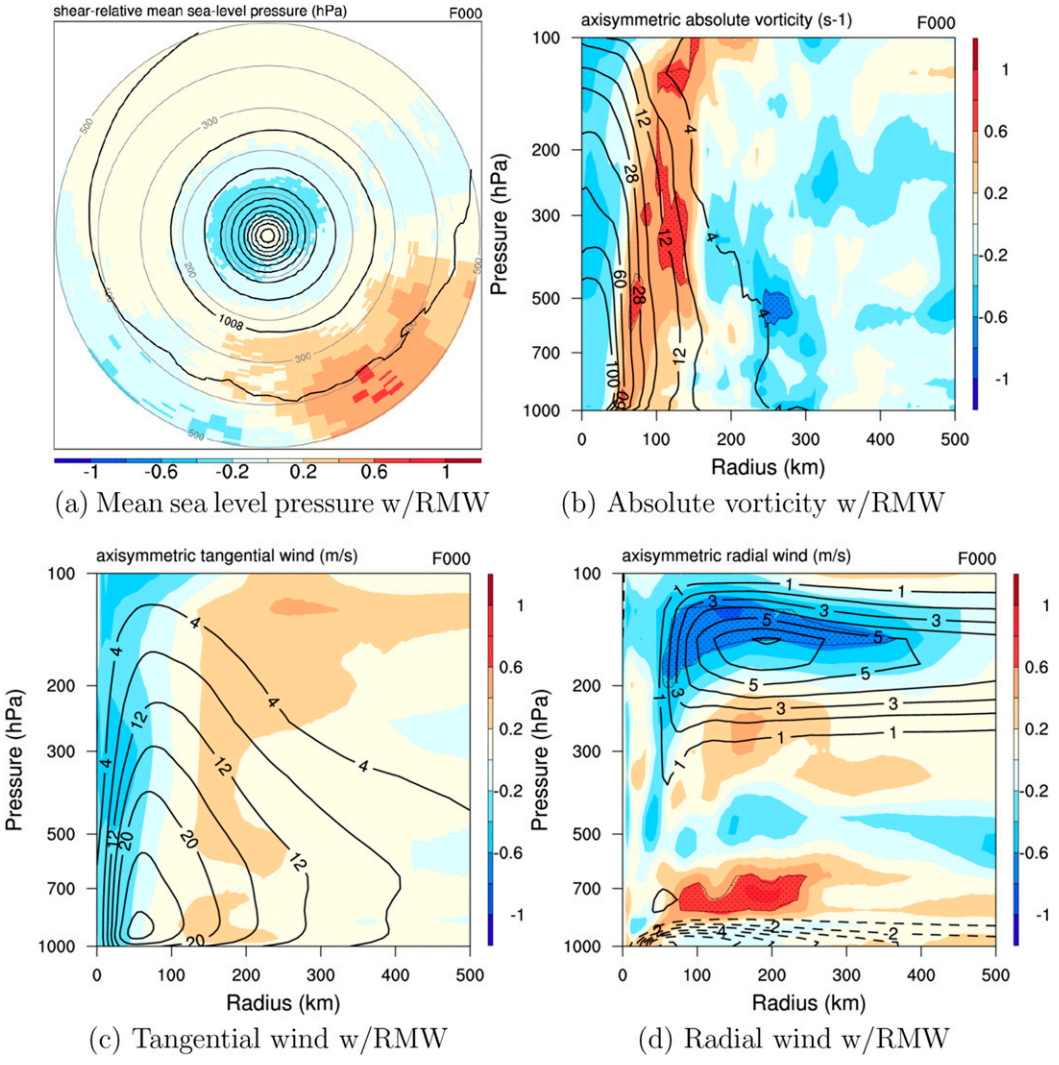
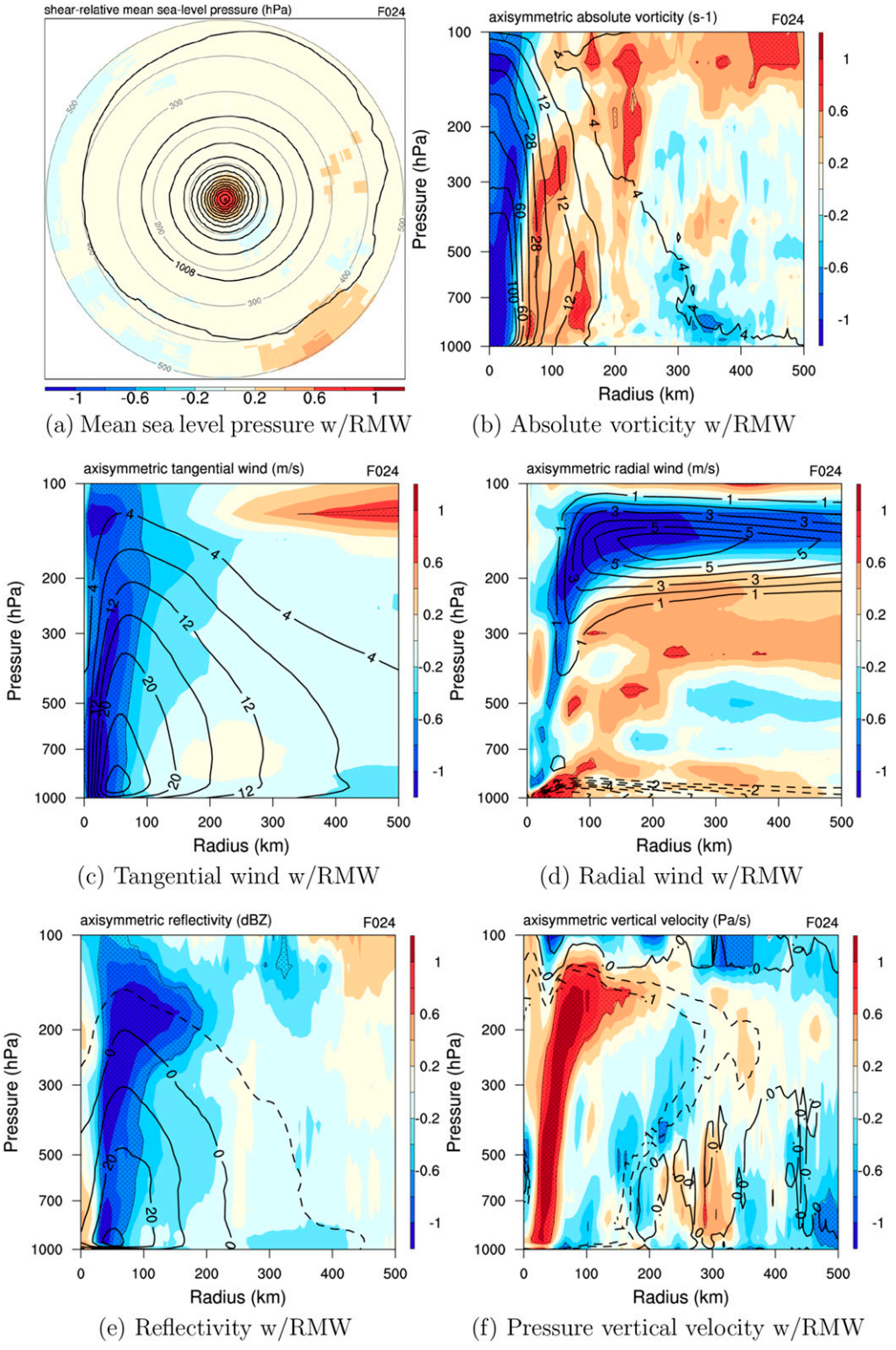



FIG. 14. As in Fig. 4, but for cases with RMW included in the cost function.

higher MSLP values than the analog forecasts approximately 100–300 km from the TC center. The bad forecasts are also weaker in terms of 10-m tangential wind beyond 75-km radii. These results suggest that the bad forecasts are smaller in size than the analog forecasts. Indeed, the mean best-track radius of the outer closed isobar (ROCI) of the bad forecasts is 95 km (22%) smaller than the mean ROCI of the analog forecasts. This difference is statistically significant at the 97% confidence level according to a Wilcoxon rank-sum test (Wilks 2011). In contrast, the mean ROCIs for the bad and analog forecasts over the EPAC were within 4 km (not significantly different). Therefore, some of the differences in structure noted away from the core in the NATL cases may be due primarily to differences in TC size. Despite the differences in TC size as measured by ROCI, the analog cost function that included RMW [i.e., Eq. (2)] depicted similar results to those presented in Figs. 18 and 19. That is, the significant differences between bad and analog forecasts occurred primarily outside of the RMW.

Weaker convection in the bad forecasts may be contributing to the underintensification in this basin as well; however, it appears that the differences are not maximized in the inner core (Fig. 20). In fact, the reflectivity differences are rather inconsistent as a function of lead time. For example, the bad forecasts have significantly smaller reflectivity above 200 hPa between 100 and 300 km at 3 h. These differences extend downward toward the surface at 6 h, then outward to 300–500-km radius by 12 h. It is possible that these inconsistencies result from differing positions of outer rainbands in the various TCs. Regardless, the results are consistent with the bad forecasts being smaller TCs (i.e., weaker winds at larger radii).




(a) Mean sea level pressure w/RMW (b) Absolute vorticity w/RMW

(c) Tangential wind w/RMW (d) Radial wind w/RMW

(e) Reflectivity w/RMW (f) Pressure vertical velocity w/RMW

FIG. 15. As in Fig. 5, but for cases with RMW included in the cost function.

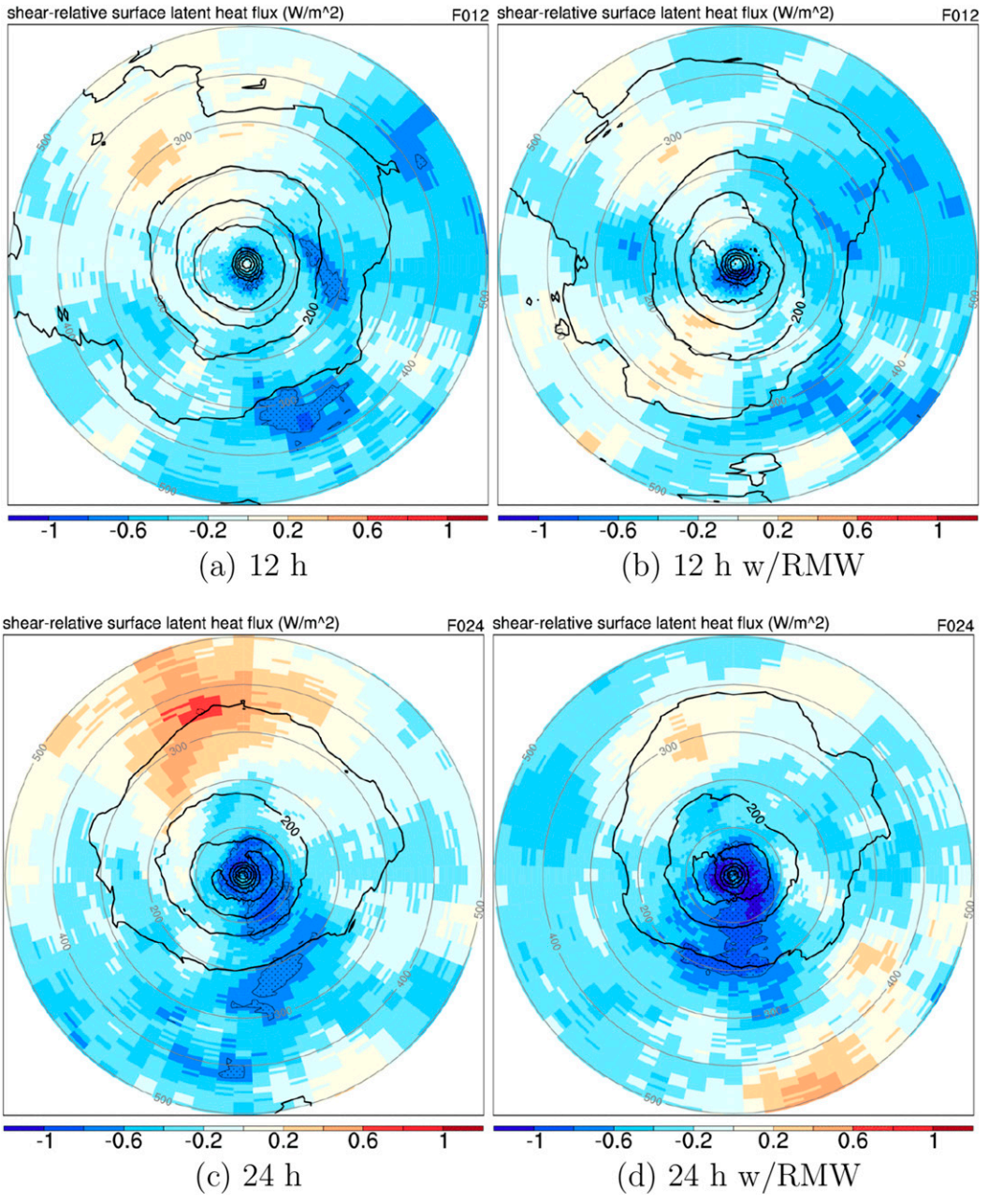



FIG. 16. As in Fig. 10, but comparing cases (right) with and (left) without RMW included in the cost function at 12 and 24 h.

An inspection of the reflectivity profiles by quadrant does not provide much clarification. One consistent difference, however, is significantly weaker reflectivity in the downshear-right quadrant of bad forecasts between 250- and 400-km radius. These differences are primarily confined to above 400 hPa at 3 and 6 h, but extend throughout the depth of the troposphere at 9 and 12 h (not shown). As in the EPAC, this result suggests that convection is shallower in the bad forecasts in the quadrant where convective initiation should be taking place. By contrast, the significant differences in the NATL occur at larger radii than the EPAC. The differences in other quadrants were less consistent (not shown).

Given that the region of significant differences is beyond the RMW, CFAD differences were constructed for the radius range from 2 × RMW to 500 km to highlight the areas of significant differences (Fig. 21). With the

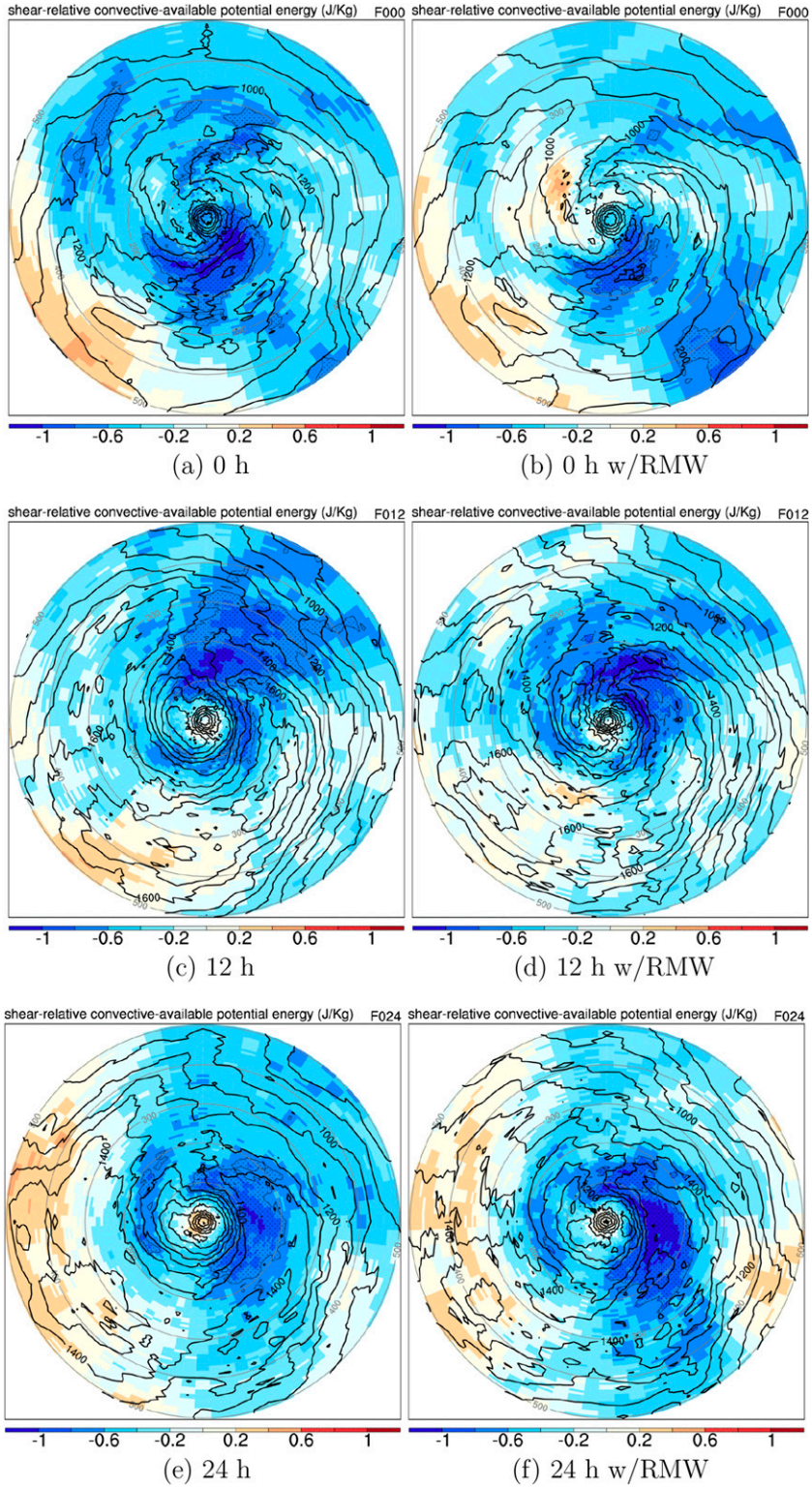



FIG. 17. As in Fig. 12, but comparing cases (right) with and (left) without RMW included in the cost function at 0, 12, and 24 h.

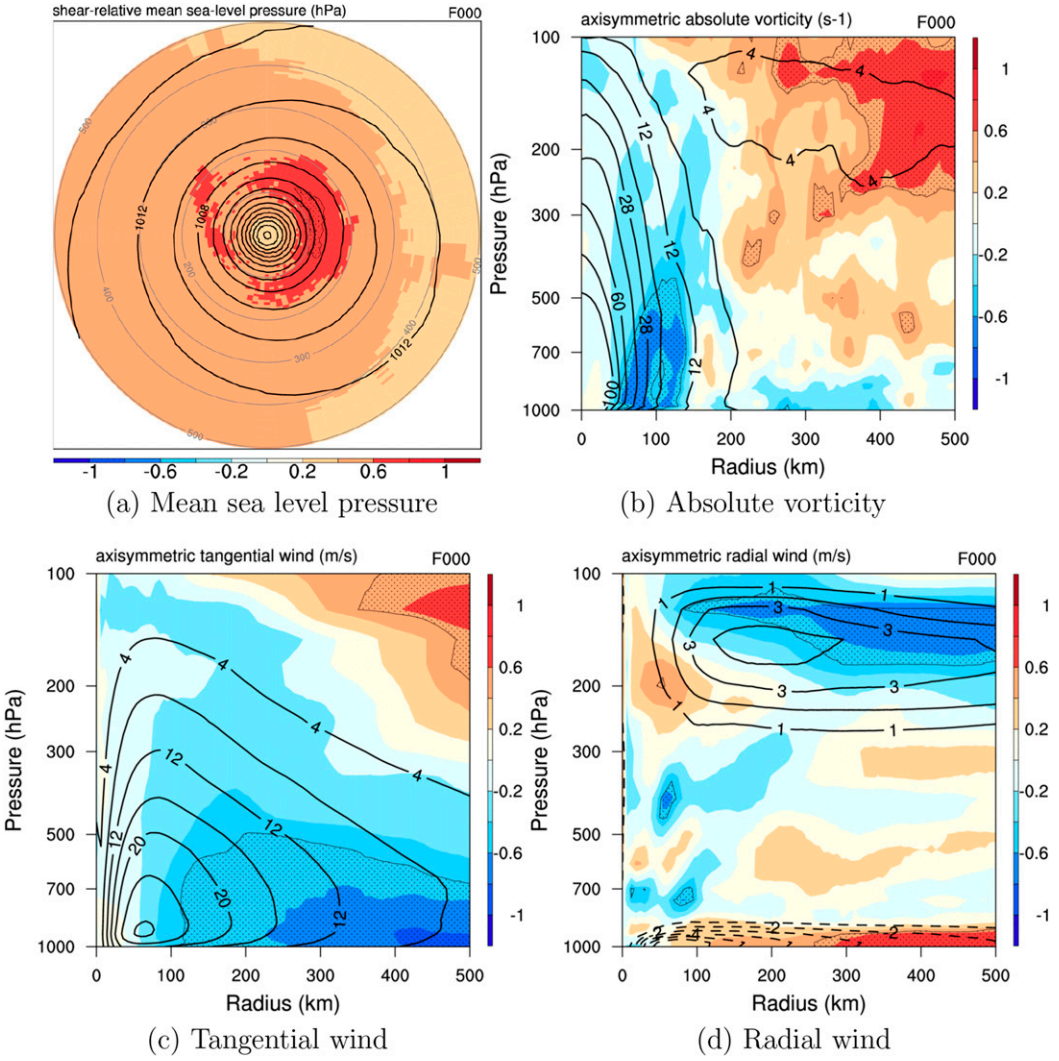



FIG. 18. As in Fig. 4, but for the NATL basin.

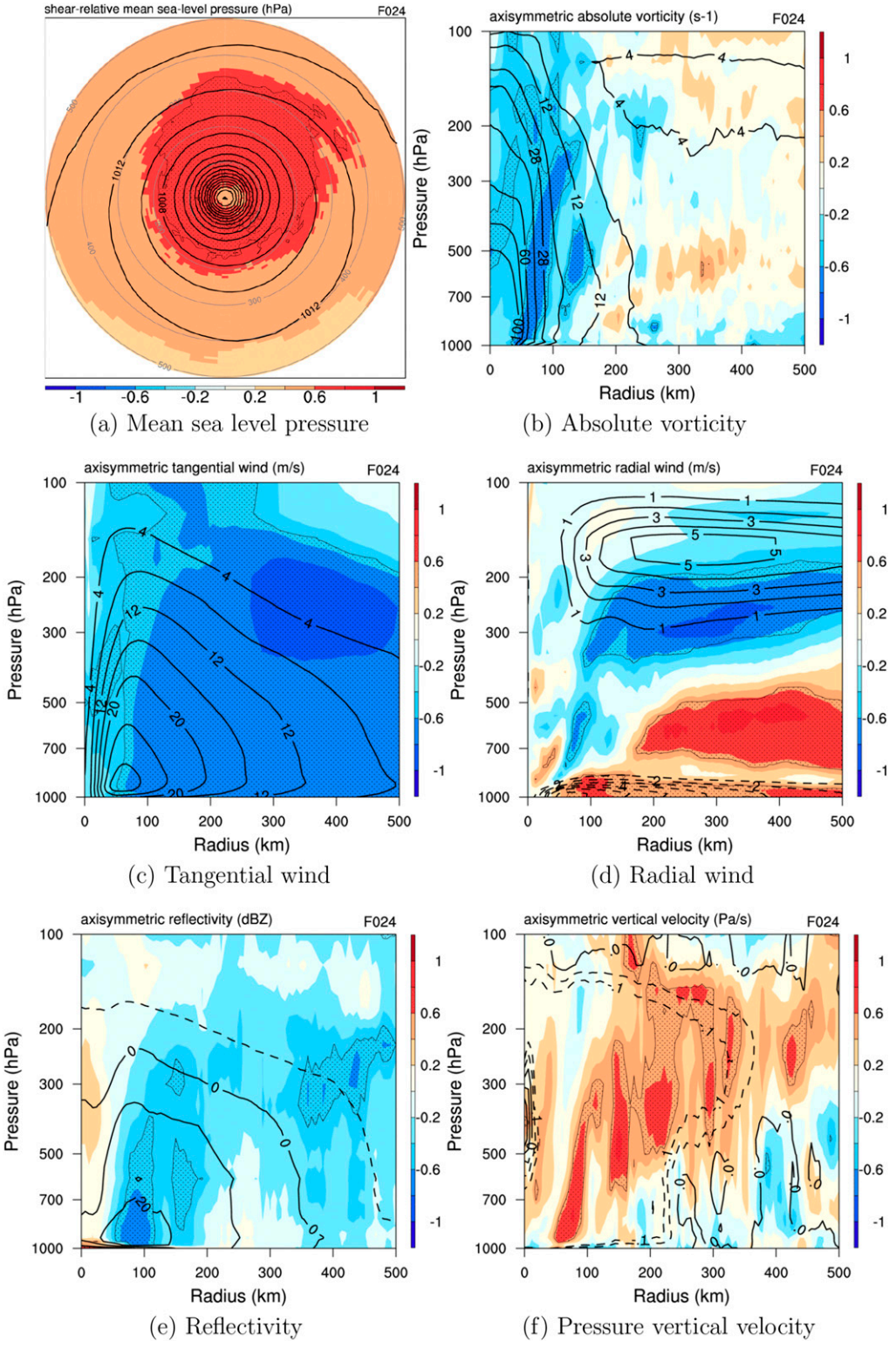
exception of reflectivities weaker than 10 dB$Z$ below 500 hPa, the bad forecasts have fewer reflectivity counts in all reflectivity–height bins (Fig. 21). Differences are greatest above 200 hPa and for reflectivities $> 35$ dB$Z$, implying that the bad forecasts have shallower and weaker convection than their analogs.

The bad forecasts also exhibit smaller surface latent heat flux values compared to the analog forecasts, especially on the right-of-shear side at radii $> 200$ km (Fig. 22). Compared to the EPAC, the differences are not as expansive in area, but they are persistent in time.

Despite the overall contrast in results between the basins, the NATL $\theta_e$ vertical profiles (Fig. 23) show that significantly lower-$\theta_e$ air is present near the surface in the bad forecasts, similar to the EPAC forecasts. Also, unlike the other NATL results, which show significant differences occurring outside of the TC core, the bad forecasts over the NATL have lower-$\theta_e$ air near the surface and in the TC core up to 400 hPa at 0 h. These structural differences persist throughout the 24-h forecast period.

Overall, the NATL results are less clear than those of the EPAC. While some similar characteristics are observed (e.g., weaker reflectivity and smaller near-surface $\theta_e$ values in the bad forecasts), the significant differences in the NATL forecasts generally occur away from the TC core. Such differences likely are present because the bad forecasts are associated with significantly smaller TCs. A more homogeneous set of TC sizes may be required to yield more concrete results; however, this might not be possible without an even larger set of retrospective forecasts.

Finally, the potential impact of forecast track errors on the results is considered. Large errors in forecast




(a) Mean sea level pressure (b) Absolute vorticity

(c) Tangential wind (d) Radial wind

(e) Reflectivity (f) Pressure vertical velocity

FIG. 19. As in Fig. 5, but for the NATL basin.

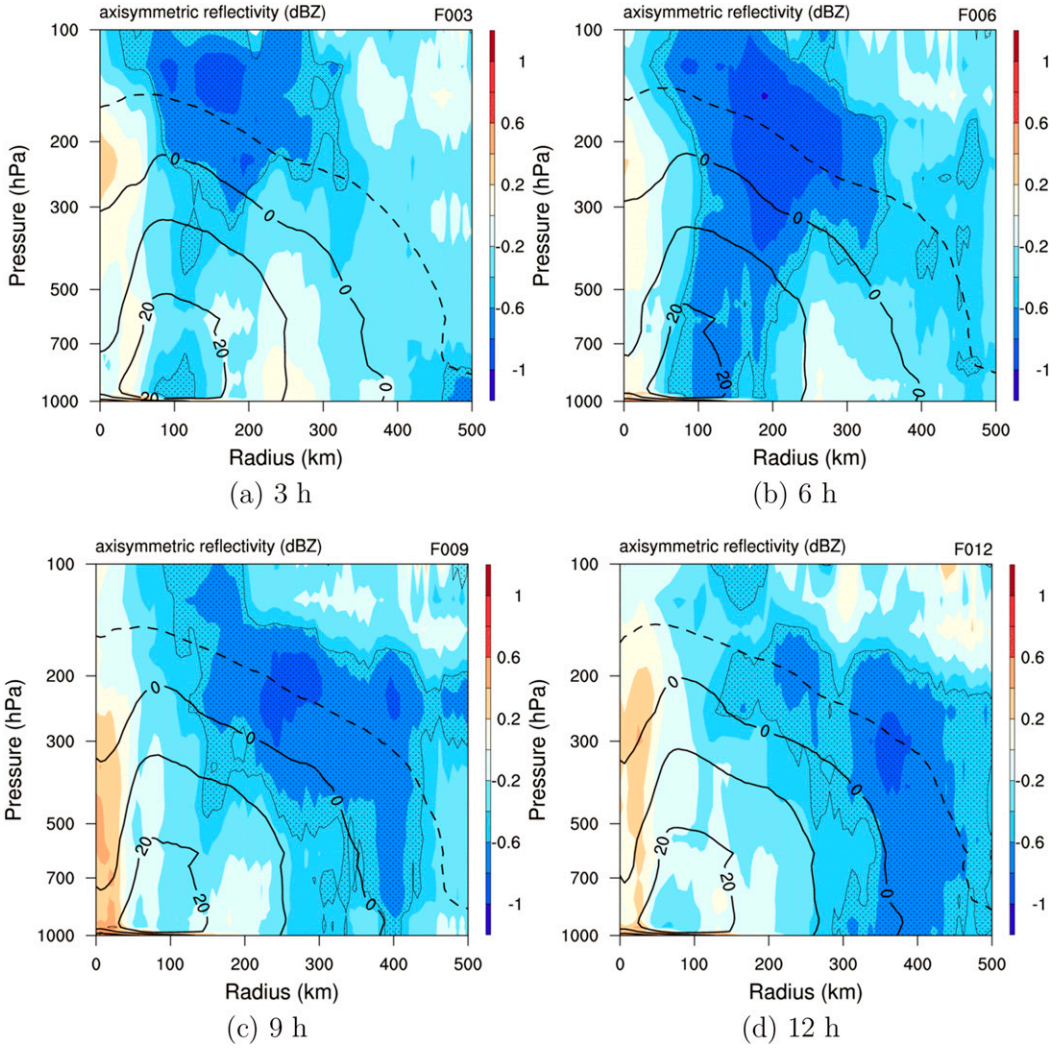



FIG. 20. As in Fig. 6, but for the NATL basin.

track could lead to errors in forecast SST, which could be a causal factor in the underintensification present in the bad forecasts. Thus, the 24-h track and SST forecast errors were computed for all cases. In both basins, the difference in the means of the forecast track and SST errors between the bad and analog forecasts were statistically indistinguishable. In addition, the average paired difference in 24-h SST errors between the bad and analog forecasts was less than 0.05°C in both basins. The paired difference of the absolute value of the 24-h SST errors was 0.76°C (0.51°C) for the EPAC (NATL) basin. This result suggests that these differences are small enough, so that they are not responsible for the large underintensification errors present in the bad forecasts.

### c. *Role of model initialization*

One common result among the basins is the presence of lower-$\theta_e$ air near the surface at 0 h for the bad forecasts. This difference could yield the weaker bad forecasts at 24 h. In an attempt to determine the cause of the $\theta_e$ differences, we investigate whether they are introduced as a result of the H215 vortex initialization/correction procedure (Tallapragada et al. 2016).

The H215 analysis fields begin with the 6-h forecast from the previous model cycle and are adjusted according to the vortex initialization/correction procedure. Therefore, it is possible to isolate the adjustments made by the vortex initialization procedure by examining the difference between the model analysis field and the 6-h forecast from the previous cycle.

We create composite difference plots similar those in the previous subsections, but now examine differences between the analysis and 6-h forecast from the previous cycle for the bad and analog forecasts. Results indicate that the analysis fields exhibit lower $\theta_e$ values near the surface compared to the 6-h forecast from the previous

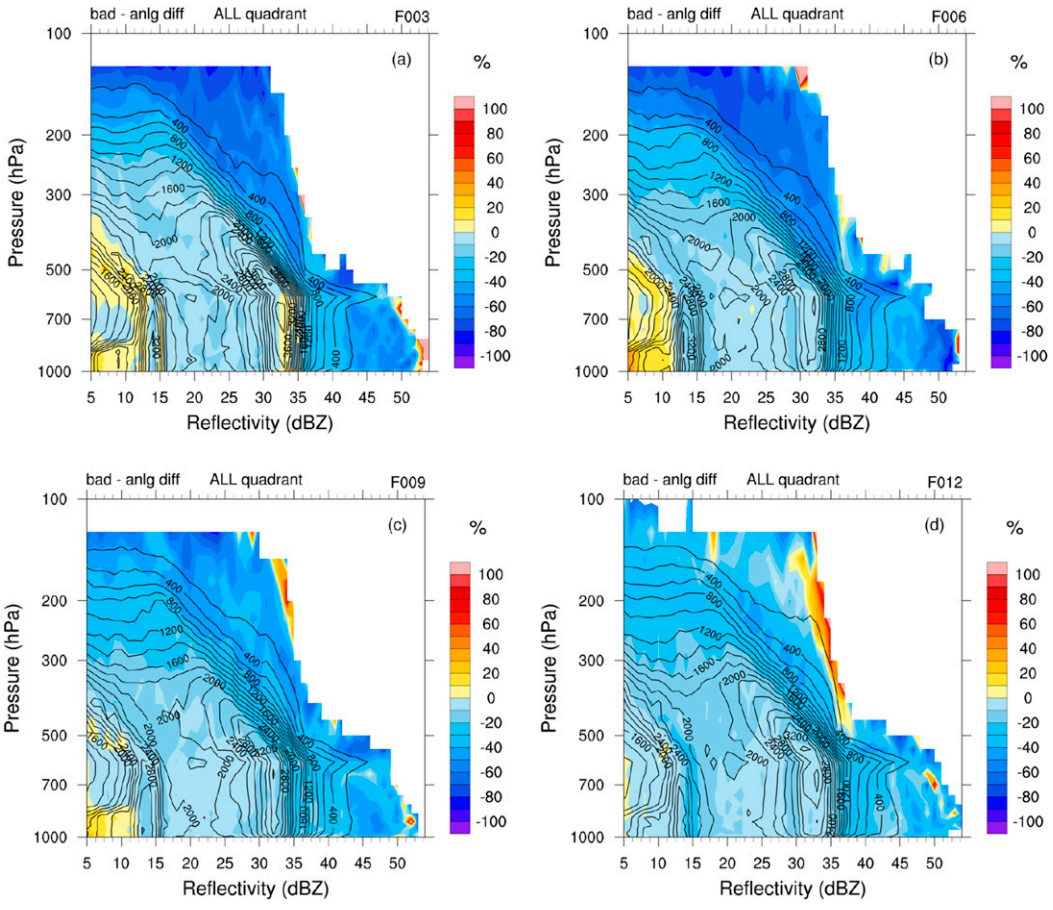



FIG. 21. As in Fig. 7, but for the NATL basin and only data outside 2 × RMW of the center are included.

cycle over both basins (not shown). However, this result is valid for both bad and analog forecasts. Thus, to determine if the bad forecasts have significantly lower-$\theta_e$ air in the analysis compared to the analog forecasts, we subtract the analog differences from the bad differences, to create the difference of the differences between the analysis and 6-h forecast from the previous cycle. Figure 24 shows that there is a small area near the surface where the difference between the analysis and 6-h forecast from the previous cycle is significantly greater for the bad forecasts than the analog forecasts in terms of $\theta_e$ and specific humidity. The analyses have lower $\theta_e$ and specific humidity than the 6-h forecasts from the previous cycle. This result is valid for both basins, but the significant differences are found at large radii (350–500 km) over the EPAC, and closer to the TC core (50–200 km) over the NATL. However, given the relatively small area of significant differences, it is unlikely that the vortex initialization/correction procedure alone is responsible for the smaller $\theta_e$ in the bad forecasts described previously.

## 4. Summary and conclusions

This paper presented a methodology for diagnosing conditions associated with large tropical cyclone forecast intensity errors and applied it to a set of reforecasts using the 2015 HWRF configuration. Large error (i.e., ''bad'') forecasts were identified and compared to smaller error forecasts that exhibit a similar evolution of intensity and shear magnitude in the best-track and SHIPS analyses, respectively (i.e., ''analog'' forecasts). Comparing composites of bad and analog forecasts can assist in identifying the systematic conditions associated with large forecast errors.

Identifying analog forecasts by minimizing a cost function that only included TC intensity and shear magnitude appears to be appropriate for 24-h forecasts; results showed little sensitivity to including the radius of maximum wind in the cost function. However, it is possible that extending this methodology to longer-term forecasts might require incorporating more environment and structure metrics into the cost function to identify suitable analogs. This could be the subject of future work.

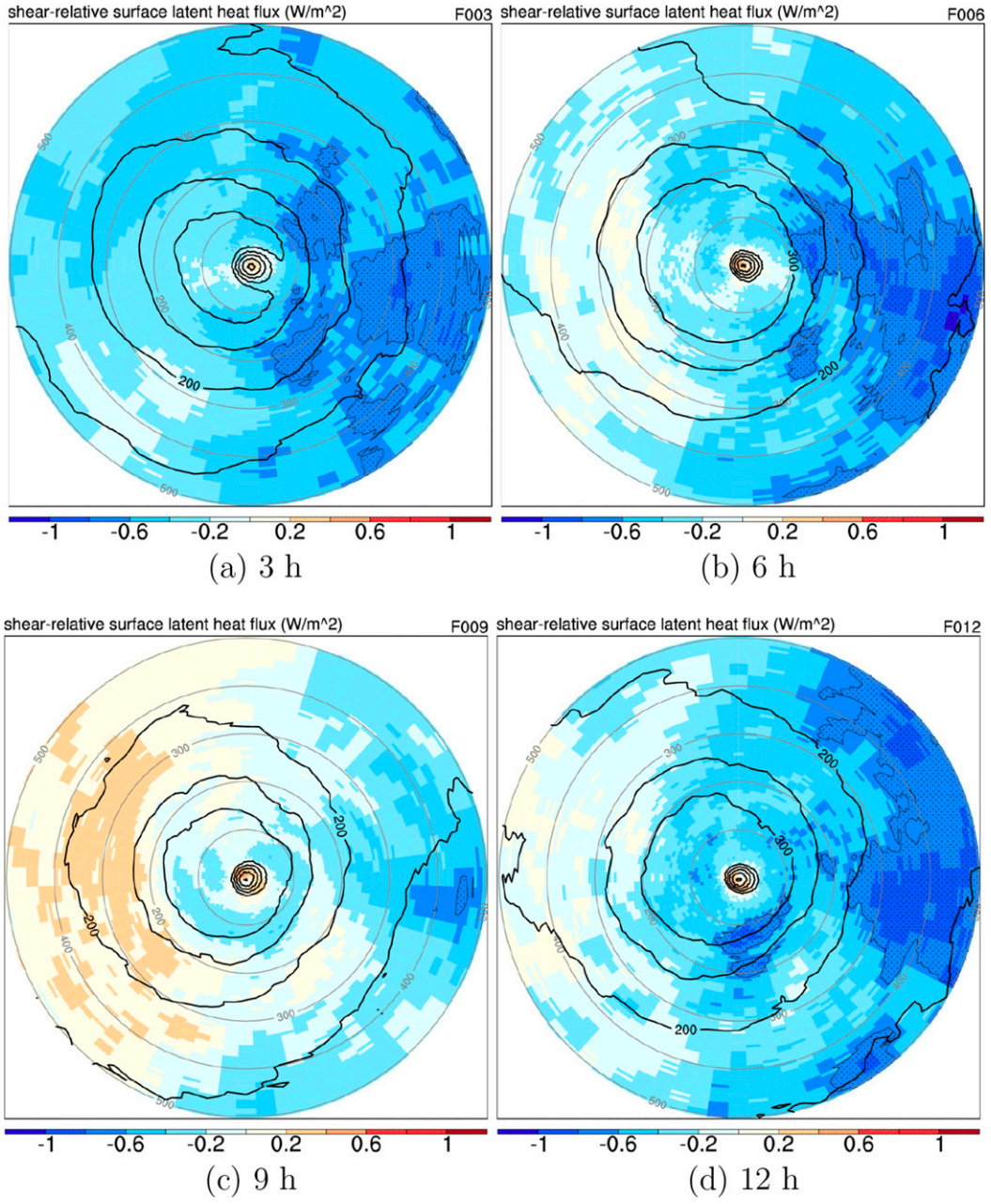



FIG. 22. As in Fig. 10, but for the NATL basin.

A majority of bad HWRF intensity forecasts in the EPAC and NATL basins were the result of HWRF underforecasting the 24-h intensity. Thus, the methodology was applied to this subset of cases. It was hypothesized that weaker convection near the TC core was responsible for the underintensification.

Results for EPAC forecasts showed that nearly all of the significant differences between the bad and analog forecasts occurred near the TC core (i.e., radius $<$ 75–100 km). The strength and extent of convection, as measured by reflectivity and vertical motion $\omega$, was indeed weaker in the bad forecasts compared to the analog forecasts from 3 to 24 h. Differences largely occurred in the upper levels during the earlier time periods, indicating that the convection was initially shallower in the bad forecasts. These differences are associated with smaller surface latent heat fluxes or lower-$\theta_e$ air near the surface in the bad forecasts compared to the analog forecasts.

The EPAC and NATL composites exhibited some similar results. For example, the convection was weaker

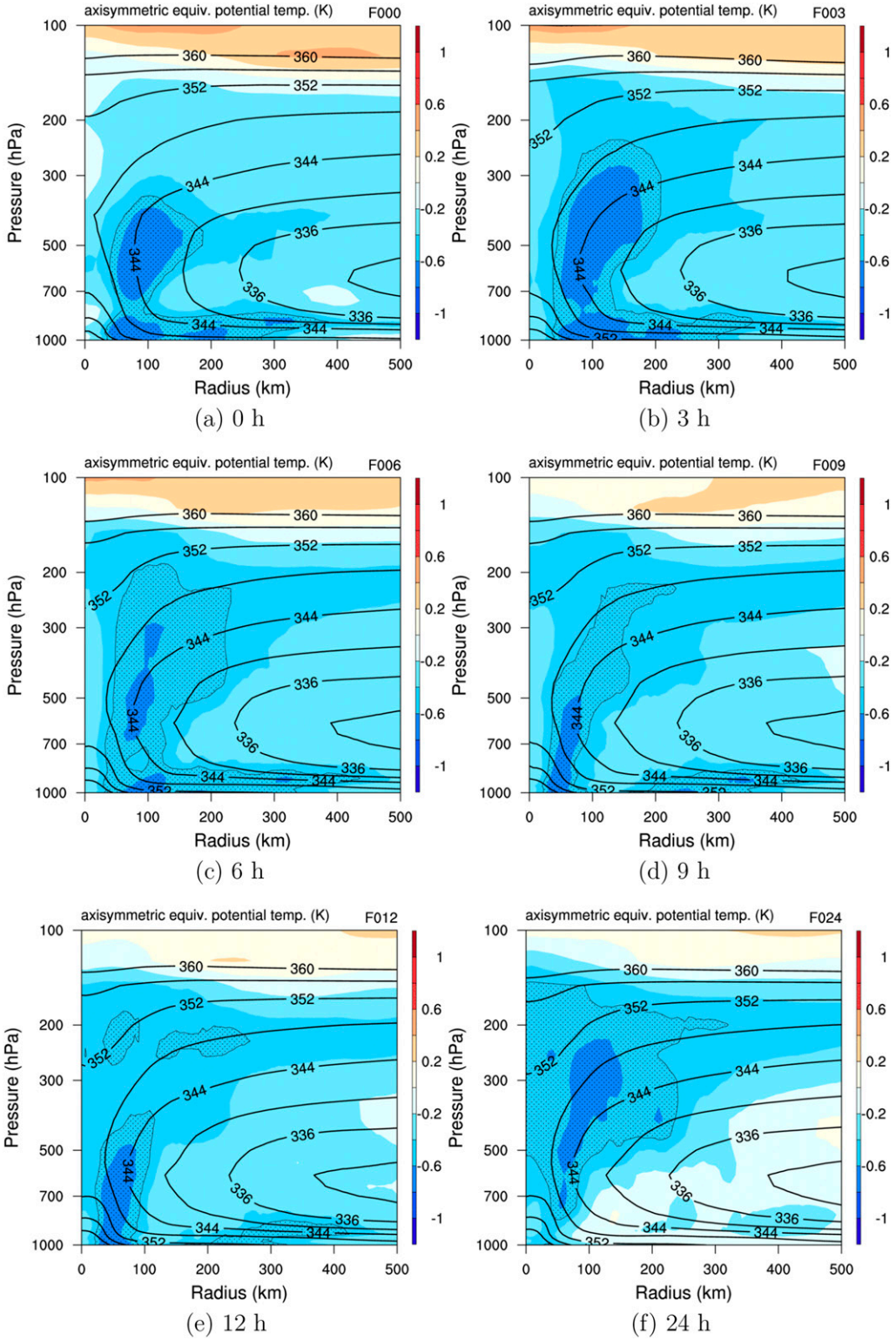



FIG. 23. As in Fig. 11, but for the NATL basin.

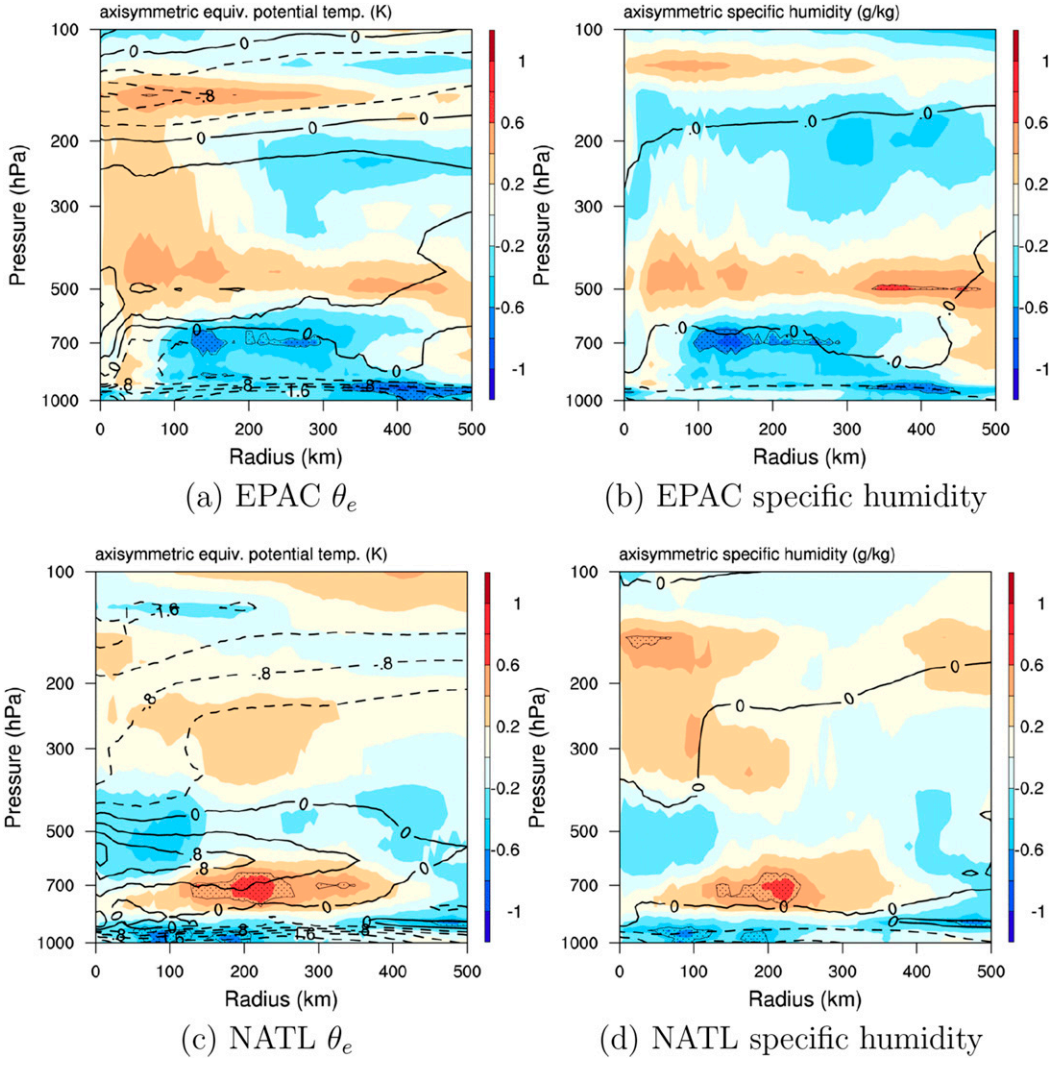



FIG. 24. Composite differences between bad and analog forecasts of the difference between the analysis and 6-h forecast from the previous cycle. Shown are results for (a) EPAC $\theta_e$, (b) EPAC specific humidity, (c) NATL $\theta_e$, and (d) NATL specific humidity.

and shallower in the bad forecasts compared to the analog forecasts. Smaller near-surface $\theta_e$ values also occurred in the bad forecasts. However, unlike the EPAC, these differences were found primarily outside of the TC core (i.e., radius $>$ 100 km) over the NATL. A comparison of the radius of the outer closed isobar for bad and analog forecasts revealed that the bad forecasts are statistically smaller than the analog forecasts over the NATL. This size difference may explain why differences in TC structure are observed at larger radii over the NATL.

This methodology easily can be modified to analyze different models, time periods, definitions of "bad" forecasts, and to control for different factors (i.e., best-track intensity, shear, RMW, etc.). For example, rapid intensity changes were not considered here, but would be an interesting topic for future research. Overall, we believe this methodology can be applied more generally to diagnose systematic model biases.

*Acknowledgments.* This research was supported by the NOAA Hurricane Forecast Improvement Program via Grant NA14NWS4680027. This manuscript benefited from the constructive comments from three anonymous reviewers. The authors acknowledge the NOAA Research and Development High Performance Computing Program for providing computing and storage resources that have contributed to the research results reported within this paper (http://rdhpcs.noaa.gov).

## REFERENCES


Bao, J.-W., S. Gopalakrishnan, S. Michelson, F. Marks, and M. Montgomery, 2012: Impact of physics representations in the HWRFX on simulated hurricane structure and pressure–wind relationships. *Mon. Wea. Rev.*, **140**, 3278–3299, https://doi.org/10.1175/MWR-D-11-00332.1.

Chen, H., and S. G. Gopalakrishnan, 2015: A study on the asymmetric rapid intensification of Hurricane Earl (2010) using the HWRF system. *J. Atmos. Sci.*, **72**, 531–550, https://doi.org/10.1175/JAS-D-14-0097.1.

Corbosiero, K. L., and J. Molinari, 2002: The effects of vertical wind shear on the distribution of convection in tropical cyclones. *Mon. Wea. Rev.*, **130**, 2110–2123, https://doi.org/10.1175/1520-0493(2002)130<2110:TEOVWS>2.0.CO;2.

——, and ——, 2003: The relationship between storm motion, vertical wind shear, and convective asymmetries in tropical cyclones. *J. Atmos. Sci.*, **60**, 366–376, https://doi.org/10.1175/1520-0469(2003)060<0366:TRBSMV>2.0.CO;2.

DeHart, J. C., R. A. Houze Jr., and R. F. Rogers, 2014: Quadrant distribution of tropical cyclone inner-core kinematics in relation to environmental shear. *J. Atmos. Sci.*, **71**, 2713–2732, https://doi.org/10.1175/JAS-D-13-0298.1.

DeMaria, M., 2009: A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.*, **137**, 68–82, https://doi.org/10.1175/2008MWR2513.1.

——, M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543, https://doi.org/10.1175/WAF862.1.

Galarneau, T. J., and T. M. Hamill, 2015: Diagnosis of track forecast errors for Tropical Cyclone Rita (2005) using GEFS reforecasts. *Wea. Forecasting*, **30**, 1334–1354, https://doi.org/10.1175/WAF-D-15-0036.1.

Gopalakrishnan, S. G., F. Marks Jr., X. Zhang, J.-W. Bao, K.-S. Yeh, and R. Atlas, 2011: The experimental HWRF system: A study on the influence of horizontal resolution on the structure and intensity changes in tropical cyclones using an idealized framework. *Mon. Wea. Rev.*, **139**, 1762–1784, https://doi.org/10.1175/2010MWR3535.1.

——, S. Goldenberg, T. Quirino, X. Zhang, F. Marks Jr., K.-S. Yeh, R. Atlas, and V. Tallapragada, 2012: Toward improving high-resolution numerical hurricane forecasting: Influence of model horizontal grid resolution, initialization, and physics. *Wea. Forecasting*, **27**, 647–666, https://doi.org/10.1175/WAF-D-11-00055.1.

——, F. Marks Jr., J. A. Zhang, X. Zhang, J.-W. Bao, and V. Tallapragada, 2013: A study of the impacts of vertical diffusion on the structure and intensity of the tropical cyclones using the high-resolution HWRF system. *J. Atmos. Sci.*, **70**, 524–541, https://doi.org/10.1175/JAS-D-11-0340.1.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33, https://doi.org/10.1175/BAMS-87-1-33.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Hazelton, A. T., R. Rogers, and R. E. Hart, 2015: Shear-relative asymmetries in tropical cyclone eyewall slope. *Mon. Wea. Rev.*, **143**, 883–903, https://doi.org/10.1175/MWR-D-14-00122.1.

Hendricks, E. A., M. S. Peng, B. Fu, and T. Li, 2010: Quantifying environmental control on tropical cyclone intensity change. *Mon. Wea. Rev.*, **138**, 3243–3271, https://doi.org/10.1175/2010MWR3185.1.

Jarvinen, B., C. Neumann, and M. Davis, 1984: A tropical cyclone data tape for the North Atlantic basin, 1886–1983: Contents, limitations, and uses. NOAA Tech. Memo. NWS NHC 22, 21 pp., http://www.nhc.noaa.gov/pdf/NWS-NHC-1988-22.pdf.

Jiang, H., 2012: The relationship between tropical cyclone intensity change and the strength of inner-core convection. *Mon. Wea. Rev.*, **140**, 1164–1176, https://doi.org/10.1175/MWR-D-11-00134.1.

Lamberson, W. S., R. D. Torn, L. F. Bosart, and L. Magnusson, 2016: Diagnosis of the source and evolution of medium-range forecast errors for extratropical cyclone Joachim. *Wea. Forecasting*, **31**, 1197–1214, https://doi.org/10.1175/WAF-D-16-0026.1.

McAdie, C., C. Landsea, C. Neumann, J. David, E. Blake, and G. Hammer, 2009: Tropical cyclones of the North Atlantic Ocean, 1851–2006. Historical Climatology Series 6-2, 238 pp., http://www.nhc.noaa.gov/pdf/TC_Book_Atl_1851-2006_lowres.pdf.

NHC, 2016: National Hurricane Center forecast verification: Official error trends. National Hurricane Center, http://www.nhc.noaa.gov/verification/verify5.shtml.

Rappaport, E. N., J.-G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The Joint Hurricane Test Bed: Its first decade of tropical cyclone research-to-operations activities reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371–380, https://doi.org/10.1175/BAMS-D-11-00037.1.

Reasor, P. D., R. Rogers, and S. Lorsolo, 2013: Environmental flow impacts on tropical cyclone structure diagnosed from airborne Doppler radar composites. *Mon. Wea. Rev.*, **141**, 2949–2969, https://doi.org/10.1175/MWR-D-12-00334.1.

Rios-Berrios, R., and R. D. Torn, 2017: Climatological analysis of tropical cyclone intensity changes under moderate vertical wind shear. *Mon. Wea. Rev.*, **145**, 1717–1738, https://doi.org/10.1175/MWR-D-16-0350.1.

——, ——, and C. A. Davis, 2016a: An ensemble approach to investigate tropical cyclone intensification in sheared environments. Part I: Katia (2011). *J. Atmos. Sci.*, **73**, 71–93, https://doi.org/10.1175/JAS-D-15-0052.1.

——, ——, and ——, 2016b: An ensemble approach to investigate tropical cyclone intensification in sheared environments. Part II: Ophelia (2011). *J. Atmos. Sci.*, **73**, 1555–1575, https://doi.org/10.1175/JAS-D-15-0245.1.

Rogers, R., P. Reasor, and S. Lorsolo, 2013: Airborne Doppler observations of the inner-core structural differences between intensifying and steady-state tropical cyclones. *Mon. Wea. Rev.*, **141**, 2970–2991, https://doi.org/10.1175/MWR-D-12-00357.1.

——, J. A. Zhang, J. Zawislak, H. Jiang, G. R. Alvey, E. J. Zipser, and S. N. Stevenson, 2016: Observations of the structure and evolution of Hurricane Edouard (2014) during intensity change. Part II: Kinematic structure and the distribution of deep convection. *Mon. Wea. Rev.*, **144**, 3355–3376, https://doi.org/10.1175/MWR-D-16-0017.1.

Tallapragada, V., C. Kieu, Y. Kwon, S. Trahan, Q. Liu, Z. Zhang, and I.-H. Kwon, 2014: Evaluation of storm structure from the operational HWRF during 2012 implementation. *Mon. Wea. Rev.*, **142**, 4308–4325, https://doi.org/10.1175/MWR-D-13-00010.1.

——, and Coauthors, 2016: Hurricane Weather Research and Forecasting (HWRF) Model: 2015 scientific documentation. NCAR Tech. Note NCAR/TN-522+STR, 119 pp., https://opensky.ucar.edu/islandora/object/technotes%3A535/datastream/PDF/view.

Torn, R. D., J. S. Whitaker, P. Pegion, T. M. Hamill, and G. J. Hakim, 2015: Diagnosis of the source of GFS medium-range track errors in Hurricane Sandy (2012). *Mon. Wea. Rev.*, **143**, 132–152, https://doi.org/10.1175/MWR-D-14-00086.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 676 pp.

Yuter, S. E., and R. A. Houze, 1995: Three-dimensional kinematic and microphysical evolution of Florida cumulonimbus. Part II: Frequency distributions of vertical velocity, reflectivity, and differential reflectivity. *Mon. Wea. Rev.*, **123**, 1941–1963, https://doi.org/10.1175/1520-0493(1995)123<1941:TDKAME>2.0.CO;2.

Zarzycki, C. M., D. R. Thatcher, and C. Jablonowski, 2016: Objective tropical cyclone extratropical transition detection in high-resolution reanalysis and climate model data. *J. Adv. Model. Earth Syst.*, **9**, 130–148, https://doi.org/10.1002/2016MS000775.

Zhang, J. A., and F. D. Marks, 2015: Effects of horizontal diffusion on tropical cyclone intensity change and structure in idealized three-dimensional numerical simulations. *Mon. Wea. Rev.*, **143**, 3981–3995, https://doi.org/10.1175/MWR-D-14-00341.1.