A Forecast Cycle–Based Evaluation for Tropical Cyclone Rapid Intensification Forecasts by the Operational HWRF Model

WEIGUO WANG,[a,c] LIN ZHU,[a,c] BIN LIU,[b,c] ZHAN ZHANG,[d] AVICHAL MEHRA,[d] AND VIJAY TALLAPRAGADA[d]

[a] *SAIC, NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*
[b] *LyNker, NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*
[c] *IMSG, NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*
[d] *NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*

(Manuscript received 12 December 2021, in final form 24 October 2022)

ABSTRACT: An evaluation framework for tropical cyclone rapid intensification (RI) forecasts is introduced and applied to evaluate the performance of RI forecasts by the operational Hurricane Weather Research and Forecasting (HWRF) Model. The framework is based on the performance of each 5-day forecast cycle, while the conventional RI evaluation is based on the statistics of successful or false RI forecasts at individual lead times. The framework can be used to compare RI forecasts of different cycles, which helps model developers and forecasters to characterize RI forecasts under different scenarios. It also can provide the evaluation of statistical performance in the context of 5-day forecast cycles. The RI forecast of each cycle is assessed using a modified probability-based approach that takes the absolute errors in intensity changes into account. The overall performance of RI forecasts during a given period is assessed based on the fractions of the individual forecast cycles during which RI events are successfully or falsely predicted. The framework is applied to evaluate the performance of RI forecasts by the HWRF Model for the whole life cycle of a single hurricane, as well as for each of the hurricane seasons from 2009 to 2021. The metric based on the probabilities of detection and false alarm rate of RI is compared with that based on the absolute errors in the intensity and intensity change during RI events.

SIGNIFICANCE STATEMENT: An evaluation framework for tropical cyclone rapid intensification (RI) forecasts is introduced, focusing on the performance of RI forecasts in each 5-day forecast cycle. The cycle-based approach can help to characterize RI forecasts under different conditions such as certain synoptic scenarios, initial conditions, or vortex structures. It also can be used to assess the overall performance of RI forecasts in terms of the percentages of individual forecast cycles that successfully or falsely predict RI events.

KEYWORDS: Forecast verification/skill; Model errors; Model evaluation/performance; Error analysis; Intensification; Tropical cyclones

1. Introduction

Accurately forecasting the rapid intensification (RI) events of tropical cyclones (TCs) remains a challenge despite recent improvements in the overall performance of TC forecasts (Cangialosi et al. 2020; DeMaria et al. 2014, 2021; Gall et al. 2013). RI is a scenario where the intensity of a TC increases dramatically [e.g., by 30 kt (1 kt $\approx$ 0.51 m s^{-1}) or greater] in a short period of time, such as 24 h (Kaplan and DeMaria 2003; Kaplan et al. 2010) and improving RI forecasts is a high priority of the National Hurricane Center (NHC). The challenge in forecasting RI events stems from a lack of understanding of the physical mechanisms of RI and limitations in forecasting approaches. The underlying cause of RI onset and development is an active area of research, where prior studies have shown RI is influenced by multiscale processes such as inner-core convection (e.g., Callaghan 2017; Willoughby et al. 1982), vortex alignment process in the presence of vertical wind shear (e.g., Finocchio et al. 2016; Rios-Berrios et al. 2018), large-scale conditions (e.g., Hendricks et al. 2010; Kaplan et al. 2010; Knaff et al. 2003), and atmosphere–ocean interaction (e.g., Domingues et al. 2019; Kim et al. 2014, 2022). Given our limited knowledge of the physical processes of RI, empirical statistics–based models have been developed and used in both research and operations (e.g., DeMaria et al. 2021; Kaplan et al. 2010, 2015; Knaff et al. 2018, 2020; Rozoff et al. 2015, and references therein). A weakness of these statistical models is that the fundamental dynamic and physical processes are simplified based on empirical equations, which could limit their applications.

Alternatively, numerical modeling with full dynamics and physics suites can be used to predict RI events. With advances in high-performance computing and improvements in physical and dynamical parameterization schemes, numerical models can better resolve the fine-scale structure of the TC inner core; these advancements have improved the forecasting ability of coupled atmosphere–ocean models such as Hurricane Weather Research and Forecasting (HWRF) Model (Tallapragada 2016) and the Hurricane Multiscale Ocean-coupled Nonhydrostatic (HMON) model (Mehra et al. 2018; Wang et al. 2019). For example, the intensity and track forecast errors by the HWRF Model have been significantly reduced over the last decade as documented in the annual reports and strategic plans of the NOAA Hurricane Forecast Improvement Project (HFIP) (https://hfip.org/), as well as in a recent study by Cangialosi et al. (2020). Case studies on RI forecasts over the northwest Pacific

Corresponding author: Weiguo Wang, Weiguo.Wang@noaa.gov

DOI: 10.1175/WAF-D-22-0007.1

basin in 2013 have shown that the real-time operational HWRF Model had high detection and low false alarm rates of RI, and outperformed other numerical as well as statistical models (Tallapragada and Kieu 2014; Tallapragada et al. 2015, 2016). Cangialosi et al. (2020) showed the detection rate of RI by the HWRF Model has notably increased in the years since 2007, with further improvement since 2015. DeMaria et al. (2021) concluded that operational RI forecasts at NHC have improved by 20%–25% since the 2015–17 baseline period, based on a reduction in absolute intensity errors during forecasted or observed RI events. DeMaria et al. (2021) also found that the HWRF and HMON models have improved model skill in RI forecasts since 2015 and both models provided the leading guidance for the Northern Atlantic (NATL) basin. Nevertheless, the utility of numerical models is still limited for RI forecasts (i.e., having low detection rates and high false alarm rates) and more effort is needed to evaluate and analyze model forecasts to identify the defects of numerical models in RI forecasting and to make further improvements.

The existing RI evaluation frameworks focus on the model performance at separate lead times rather than over the entirety of each forecast cycle (typically 5-day forecasts for mesoscale models). Therefore, current RI evaluations may not be best suited for identifying forecast problems, as they do not reflect whether individual forecast cycles successfully or falsely predict RI events. Numerical model developers and hurricane forecasters have shown interest in the model's ability to capture part or all of RI events during an entire forecast cycle. This is especially true when multiple separate RI events occur during a single forecast cycle (e.g., as in Fig. 1), where an RI analysis could be more useful over the entire cycle than for separate lead times. By analyzing an entire TC forecast cycle, the cycle-based analysis can uniquely identify and categorize forecast challenges related to recurring synoptic scenarios, initial conditions, or TC characteristics. For forecasters, a cycle-based RI analysis can help to diagnose and recognize scenarios where a model is likely to perform particularly well or poorly. In addition, the cycle-based analysis can provide seasonal statistics of RI performance in the context of 5-day forecast cycles, which provides a different perspective from the traditional evaluations focusing on RI performance at individual lead times. Given the advantages of the cycle-based analysis, we propose a new evaluation framework to analyze and compare the performance of RI forecasts in the context of individual forecast cycles.

Two common approaches are used in existing RI evaluation frameworks. One approach calculates the overall probability of detection (POD) and false alarm rate (FAR) indices for forecasted RI events at individual forecast lead times (e.g., 24, 48 h) during one or more hurricane seasons (e.g., DeMaria et al. 2014, 2021; Kaplan and DeMaria 2003; Kaplan et al. 2010). The POD index is the rate of the model's successful forecasts of individual RI events that were observed by measurement, which is defined as

$$\mathrm{POD} = \frac{\text{number of forecasted RI events that were observed}}{\text{total number of observed RI events}}, \tag{1}$$

while the FAR index is the rate of the model's false forecasts of RI events that were not observed at all, which is defined as

$$\mathrm{FAR} = \frac{\text{number of forecasted RI events that were not observed}}{\text{total number of forecasted RI events}}. \tag{2}$$

The other approach calculates the mean absolute error (MAE) of the forecasted intensity at individual lead times during the periods of observed or forecasted RI events (DeMaria et al. 2021). The POD and FAR category evaluation with a cutoff value can explicitly give information on RI events being correctly or falsely detected. However, it can be misleading when the observed and forecasted intensity increases are very close or very different. For example, this method can classify a good forecast of an intensity increase as a failure when it is only slightly smaller than that of the observed RI event. Likewise, a bad forecast of an intensity increase can be classified as a success when it is much larger than that of the observed RI event. This is because the binary classification does not consider the quantitative errors in the forecasted intensity or intensity change (IC). Conversely, the MAE-based approach does quantitatively calculate the absolute errors (AE), but it does not include useful information on how many RI events were correctly or falsely predicted and how they contribute to the total errors. This is because the MAE-based approach combines the total RI events that were either forecasted or observed without distinguishing between correct and false RI forecasts. Both approaches can be used to evaluate RI forecasts over individual forecast cycles.

In this study, the POD/FAR-based approach is adapted to evaluate RI forecasts of each 5-day cycle, with a modification to consider the AEs of 24-h ICs (AEIC) in the calculations of POD and FAR indices to address the weakness mentioned above. The statistical performance in RI forecasts by a model is based on the percentages of individual 5-day forecast cycles during which RI events are successfully or falsely predicted. To illustrate the cycle-based framework, we calculate and analyze the successful and false prediction rates of RI cycles by the HWRF Model. The configuration of the HWRF Model can be found in recent publications (e.g., Biswas et al. 2017). Section 2 describes the details of the proposed framework and definitions for assessing the RI forecasts of individual cycles. Section 3 presents and discusses the results of RI performance for the forecast cycles of a single storm, a season, and multiple seasons using the proposed framework and metrics. A summary of the proposed framework is given in section 4.

2. Data and methods

a. Data

The NHC's post-storm best track analysis data obtained from the Automated Tropical Cyclone Forecasting (ATCF) System (Sampson and Schrader 2000) (https://ftp.nhc.noaa.gov/atcf/archive/) are used to determine the "observed" RI events and to compare with model forecasts. Some of the

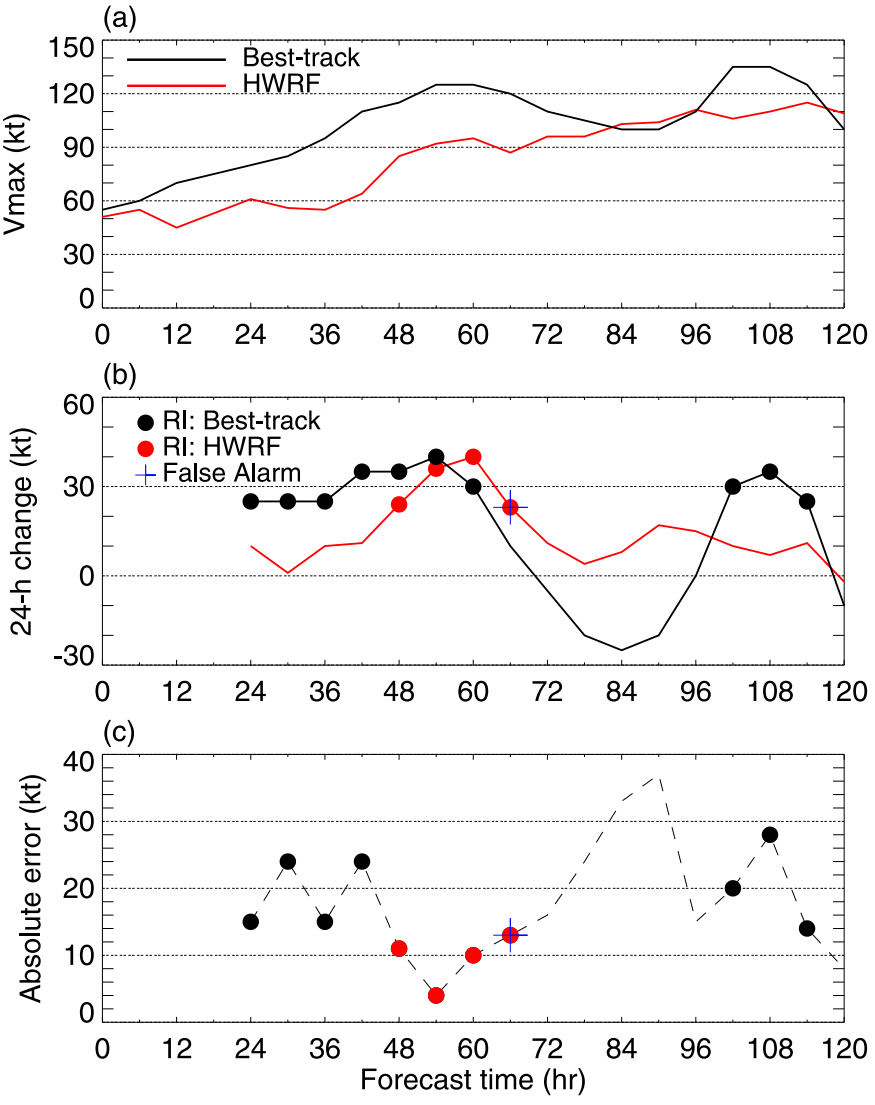

FIG. 1. (a) Time series of the maximum 10-m wind speed (intensity) during the 5 days from an HWRF forecast (red) for Hurricane Lorenzo (2019) initialized at 1800 UTC 24 Sep 2019, and the best track analysis data (black). (b) Time series of 24-h intensity changes (before the lead time), with RI events identified at 10 lead times in the best track analysis (black filled circles), and at four lead times by the HWRF forecast (red). One blue plus symbol denotes the hour when the RI event is forecast, but is not found in the best track analysis (i.e., a false alarm event). (c) AEIC. Note that no evaluations before 24 h are made because RI is relative to the intensity in the past 24 h.

5-day forecast cycles for the TCs in the NATL and eastern North Pacific (EPAC) basins generated by the operational HWRF Model from 2009 to 2021 are chosen based on the following criteria: a forecast cycle is included only if both the HWRF forecast and best track analysis have at least 48 h of intensity data over the ocean spanning the same 5-day period. It is noticed that weak systems can develop and rapidly intensify within 5 days. Therefore, the forecast cycles with initial disturbances (i.e., "invest" systems) are also included to increase the sample size, if the model can generate intensity forecasts longer than 48 h. In addition, when RI is assessed over a 5-day forecast period, it is required that both the HWRF forecast and best track analysis have available data to calculate intensity changes at a given verifying time.

b. *Methodology*

In practice, RI is defined as an increase in the maximum sustained surface wind of a TC equal to or greater than a threshold (RI_c) in a given period, such as 24 h (Kaplan and DeMaria 2003). Different RI thresholds have an impact on the RI evaluation.

Following the HFIP-adopted new performance measure for RI (DeMaria et al. 2021), the term "RI event" hereafter refers to a time when an intensity increase in the past 24 h is equal to or greater than RI_c. To evaluate the RI events produced in a 5-day forecast, a two-step procedure is applied to the time series of intensity (i.e., the maximum 10-m wind speed) derived from the model's 5-day forecast. First, the 24-h IC at a given lead (verifying) time (T) is calculated as the difference in the maximum 10-m wind speeds at T and $T - 24$ h. Second, for a given RI threshold, RI events are identified as binary results (yes or no) at all verifying times. The same two-step procedure is applied to NHC's best track analysis data for the same 5-day period of model integration. Then, the RI events derived from the model output are compared with those from the best track analysis data to determine how many of the observed RI events derived from the best track analysis have been captured or missed by the forecast and how many of the forecast RI events are false alarms during the 5-day integration period. In the evaluation, the AEIC is taken into consideration (see below).

To illustrate the analysis procedure, Figs. 1a and 1b present an example of a 6-hourly time series of the intensity and 24-h IC of Hurricane Lorenzo (2019) forecasted by the operational HWRF Model, initialized at 1800 UTC 24 September 2019. RI events [with the threshold of 20 kt $(24\ h)^{-1}$] are identified every 6 h, denoted by red filled circles for the HWRF Model and black filled circles for the best track analysis (Fig. 1b). The best track analysis suggests that the hurricane went through RI at 10 model forecast lead times (i.e., 24, 30, 36, 42, 48, 54, 60, 102, 108, and 114 h). The HWRF Model predicts RI events at only four lead times (48, 54, 60, and 66 h). Therefore, three RI events in the best track analysis are predicted (at 48, 54, and 60 h) by the HWRF Model, while the HWRF Model produces one false RI prediction at 66 h and does not capture seven RI events shown in the best track analysis. For this forecast cycle, the total numbers of the RI events observed in the best track analysis data (denoted by $N5_{best}$) and forecasted by the HWRF Model ($N5_{mod}$) are 10 and 4, respectively. The POD index in the 5-day period (POD5) is 3/10, i.e., three forecasted RI events that are in the best track analysis, divided by the total number of the RI events in the best track analysis ($N5_{best} = 10$), where the symbol "5" in the abbreviations signifies the calculation over a 5-day period to distinguish it from the conventionally used POD index.[1] Likewise, the FAR index in the 5-day period (FAR5) is 1/4, as one forecasted RI event that is not indicated in the best track analysis is divided by the total number of the forecasted RI events ($N5_{mod} = 4$). POD5 and FAR5 are calculated using the same formula as Eqs. (1) and (2), except that the RI events from the model and analysis are counted during a 5-day cycle, respectively. Note that the AEIC is not yet considered in the above calculation.

Figure 1c shows the time series of the AEIC, indicating that the AEIC is smaller than 15 kt for all three RI events the HWRF Model correctly predicted, and larger than about 14 kt for the seven RI events (at 24, 30, 36, 42, 102, 108, and 114 h) the HWRF Model missed and the RI event (at 66 h) the HWRF Model falsely predicted. As previously mentioned, a weakness in the binary classification based on a cutoff value is that it does not take forecast errors into account. To partially address this issue, RI detection and false alarm rates can be reevaluated with the AEIC being considered. For example, if we set a criterion that a RI event is successfully detected when both $IC \geq RI_c$ and $AEIC \leq 15$ kt are satisfied, there is only one RI event (at 54 h) predicted successfully by the HWRF Model. In this case, the POD5 index would be reduced from 0.3 to 0.1. The inclusion of AEIC into the POD5 calculation prevents the misclassification of a successful detection when the HWRF Model captures an observed RI event but has a very large error in intensity change. Similarly, the number of the missed or falsely predicted RI events could be adjusted if AEIC is very small (e.g., under 10% of the RI_c value); this can address the scenario where a forecasted 24-h IC is very close to the best track analysis but is still classified as a failure or false alarm (if the AEIC is not considered). In the example shown in Fig. 1, the number of missed or falsely predicted RI events could not be reduced because the AEIC is too large (about 70% of the RI_c value).

In summary, the AEIC-integrated evaluation of the forecasted IC at a lead time compared with the best track analysis is as follows:

- The observed RI event ($IC_{best} \geq RI_c$) is detected by the forecast if IC_{mod} is equal to or greater than RI_c, and $|IC_{mod} - IC_{best}| \leq \varepsilon_1$, where IC_{best} and IC_{mod} are the 24-h ICs derived from the best track data and model forecast, respectively; ε_1 is the minimum error for a successful RI detection. Thus, the forecasted RI event is excluded from the POD5 calculation if AEIC is greater than ε_1.
- The observed RI event ($IC_{best} \geq RI_c$) is missed by the forecast if IC_{mod} is smaller than RI_c and $|IC_{mod} - IC_{best}| > \varepsilon_2$, where ε_2 is the minimum IC error for a missed RI forecast. Thus, the forecast is thought to successfully detect the observed RI event and it is included in the POD5 calculation if IC_{mod} is very close to IC_{best}, i.e., $AEIC \leq \varepsilon_2$, even though IC_{mod} is smaller than RI_c.
- The forecast is an RI false alarm if IC_{mod} is equal to or greater than RI_c, but IC_{best} is smaller than RI_c and $|IC_{mod} - IC_{best}| > \varepsilon_3$, where ε_3 is the minimum IC error for a false RI forecast. Thus, the forecast is not classified as a false alarm and it is excluded from the FAR5 calculation if IC_{mod} is very close to IC_{best}, even though $IC_{mod} \geq RI_c$.

There is no impact from AEIC on the calculations of POD5 and FAR5, if $\varepsilon_1 \to \infty$, $\varepsilon_2 = 0$, and $\varepsilon_3 = 0$. In the following calculations, ε_1 is taken to be the value of RI_c and $\varepsilon_2 = \varepsilon_3 = 0.1RI_c$, unless otherwise specified. ε_1 cannot be too small because current model forecasts are not yet sufficiently

[1] The POD and FAR for RI evaluations are usually calculated for the duration of a hurricane season by aggregating the forecasts at each lead time from all storms. To avoid confusion, it should be emphasized that $N5_{best}$, $N5_{mod}$, POD5, and FAR5 in this study are counted or calculated during each single 5-day period or cycle.

TABLE 1. Thresholds for assessing a 5-day forecast cycle by the HWRF Model to predict the observed RI events. Note: P_0 and F_0 are the minimum POD5 and maximum FAR5 values, respectively, if the RI forecast during a 5-day forecast cycle is considered successful.

Performance	Thresholds
Absolute failure	$N5_{best} > 0$ and $N5_{mod} = 0$, POD5 = 0
Absolute false alarm	$N5_{best} = 0$ and $N5_{mod} > 0$, FAR5 = 1
Conditional success (CS)	$N5_{best} > 0$ and $N5_{mod} > 0$, and POD5 $\geq P_0$ and FAR5 $\leq F_0$
Conditional false alarm (CFA)	FAR5 $> F_0$
Conditional failure (CF)	POD5 $< P_0$

accurate. Otherwise, the RI detection rate can be extremely low (see an example in section 3a).

1) PERFORMANCE OF A SINGLE FORECAST CYCLE

The HWRF Model's performance during a single 5-day forecast cycle is evaluated based on the values of POD5, FAR5, as well as $N5_{mod}$ and $N5_{best}$ in the model integration period, as summarized in Table 1. Three scenarios of the model performance are considered as follows:

First, the 5-day forecast is considered an absolute failure in predicting any RI events (regardless of timing) if it does not produce a single RI event in the 5-day period, while the best track analysis indicates that there is at least one RI event in the same period, i.e., $N5_{best} > 0$ but $N5_{mod} = 0$, or POD5 = 0. In this case, the model forecast does not predict any real RI events under any conditions.

Second, the 5-day forecast generates an absolute RI false alarm if at least one RI event during the 5 days is predicted but the best track analysis shows that there is not an RI event in that period, i.e., $N5_{best} = 0$, but $N5_{mod} > 0$, or FAR5 = 1. In this case, all the predicted RI events which are not shown in the best track analysis data are false alarms.

Third, the RI forecast performance depends on the thresholds set by users if both the forecast and the best track analysis show that there is at least one RI event, in the same 5-day period, i.e., $N5_{best} > 0$ and $N5_{mod} > 0$. Users can set the thresholds of POD5 and FAR5 (P_0 and F_0, respectively) to assess whether the model successfully predicts the observed RI events in the 5-day forecast cycle (Table 1). In this scenario, the RI forecast could be a conditional success (CS) if POD5 is equal to or greater than P_0 (e.g., 0.5) and FAR5 is equal to or smaller than F_0 (e.g., 0.5), a conditional false alarm (CFA) if FAR5 exceeds F_0, and a conditional failure (CF) if POD5 is smaller than P_0. The best scenario is when FAR5 = 0 and POD5 = 1, while the worst scenario is when FAR5 = 1 and POD5 = 0.

2) PERFORMANCE OF MULTIPLE FORECAST CYCLES

To statistically evaluate the performance of RI forecasts, all 5-day forecasts by the HWRF Model in a sample set (e.g., a hurricane season) are grouped according to whether RI events can be identified, respectively, in the forecasts and in the best track data for the same 5-day periods. For illustration, the number of 5-day periods (cycles) in each group is shown in a contingency table (Table 2), where n_1, n_2, n_3, and n_4 are the numbers of 5-day periods when $N5_{best} > 0$ and $N5_{mod} > 0$, $N5_{best} = 0$ and $N5_{mod} > 0$, $N5_{best} > 0$ and $N5_{mod} = 0$, and $N5_{best} = 0$ and $N5_{mod} = 0$, respectively. The POD5 and FAR5 indices are then calculated for each of the n_1 5-day forecast cycles in the group of $N5_{best} > 0$ and $N5_{mod} > 0$, and the performance of each forecast cycle is assessed based on the conditions shown in Table 1.

The successful prediction rate (SPR) of the RI-observed cycles[2] is defined as the fraction of the number of RI-observed cycles where all or part of RI events are successfully forecasted by the HWRF Model:

$$\mathrm{SPR} = \frac{n_{CS}}{n_1 + n_3}, \tag{3}$$

where n_{CS} is the number of the conditionally successful forecast cycles (i.e., POD5 $\geq P_0$ and FAR5 $\leq F_0$) among the n_1 cycles, and $n_1 + n_3$ is the total number of the RI-observed cycles.

The false prediction rate (FPR) of the RI-forecasted cycles is defined as the fraction of the number of the RI-forecasted cycles where all or part of the forecasted RI events are not observed in the best track data (i.e., FAR5 $> F_0$):

$$\mathrm{FPR} = \frac{n_{CFA} + n_2}{n_1 + n_2}, \tag{4}$$

where n_{CFA} is the number of the conditional false alarm cycles where FAR5 exceeds F_0 among the n_1 cycles ($N5_{obs} > 0$ and $N5_{mod} > 0$), n_2 is the number of the absolute false alarm cycles, and $n_1 + n_2$ is the total number of the RI-forecasted cycles.

It should be noted that the above cycle-based model performance metrics can vary with differing threshold values of RI_c, AEIC, FAR5, and POD5 as well as differing time matching window sizes over which both the forecasted and observed RI events are checked. Other factors such as the errors in observations and forecasts may affect the evaluation, too.

3. Applications of the cycle-based evaluation and discussion

The performance of a single 5-day forecast cycle by the HWRF Model is quantified by POD5 and FAR5 indices as discussed in the previous section (Fig. 1). Next, we show the applications of the cycle-based metrics during the life cycle of a single storm, a hurricane season, and multiple hurricane seasons.

[2] For simplicity, if there are any RI events forecasted during a 5-day forecast cycle, it is called a RI-forecasted cycle hereafter. If RI events are identified in the best track analysis data during the same period of the forecast cycle, it is called a RI-observed cycle.

TABLE 2. Contingency table of the number of 5-day periods (or cycles) in a sample set based on the number of RI events identified in the forecasts and best track analysis data.

	RI events observed ($N5_{obs} > 0$)	Not observed ($N5_{obs} = 0$)	Total
RI events forecasted ($N5_{mod} > 0$)	n_1	n_2	$n_1 + n_2$
Not forecasted ($N5_{mod} = 0$)	n_3	n_4	$n_3 + n_4$
Total	$n_1 + n_3$	$n_2 + n_4$	$n_1 + n_2 + n_3 + n_4$

a. Performance of RI forecasts for a single storm

The operational HWRF Model was run for Hurricane Laura (13L), initialized every 6 h from 0000 UTC 20 August 2020 to 1800 UTC 27 August 2020, providing 32 5-day forecasts (cycles) in total. Of these, 27 cycles are chosen for evaluation based on the data sampling requirement described in section 2a. Table 3 summarizes the number of 5-day forecast cycles during which RI events [$RI_c = 30$ kt (24 h)$^{-1}$] are identified in the best track analysis data and in the forecasts. The real RI events are identified based on the best track data for 22 5-day cycles (i.e., $n_1 + n_3 = 22$), where the HWRF Model predicts at least one RI event in 21 forecast cycles ($n_1 = 21$). The HWRF Model does not produce any RI events in one forecast cycle (i.e., $n_3 = 1$) initialized at 0600 UTC 26 August 2020, where RI events are observed in the best track data, while it falsely predicts all RI events in four cycles (i.e., $n_2 = 4$) where the best track analysis does not indicate a single RI event. Figures 2a and 2b show the POD5 and FAR5 indices for the 22 forecast cycles where real RI events occurred. The HWRF Model can produce RI events during the early forecast cycles initialized between 0000 UTC 21 August and 1200 UTC 22 August 2020, but all the forecasted RI events do not match the RI occurrence times in the best track data (i.e., POD5 = 0 and FAR5 = 1). In contrast, the HWRF Model captures most real RI events at the correct times during the late cycles initialized between 1800 UTC August 22 and 0000 UTC 26 August 2020, with a high POD5 index (60%–80%) and low FAR5 index (zero except for one cycle). The cycle-to-cycle comparison in Fig. 2 gives us useful information for making hypotheses on possible underlying causes for poor RI forecasts. For this storm, the HWRF Model can predict RI events for most cycles, but the performance of the late cycle RI forecasts is much better than that of the early cycles. This suggests that the model has difficulty in correctly predicting the times of RI occurrence in early cycles of this storm, which might be related to the forecasts of tracks, environment, vortex structure, initial condition, etc. For example, the forecasted tracks from the cycles initialized between 1800 UTC 19 August and 1200 UTC 22 August are shifted to the north of the observed track so that the simulated vortex is not able to interact much with the island of Cuba and experiences favorable conditions for intensifying over water, which could lead to the false RI prediction or an incorrect timing of RI (not shown).

TABLE 3. Contingency table of the number of 5-day forecast cycles for Hurricane Laura (13L) in 2000.

	$N5_{obs} > 0$	$N5_{obs} = 0$	Total
$N5_{mod} > 0$	n_1 (=21)	n_2 (=4)	25
$N5_{mod} = 0$	n_3 (=1)	n_4 (=1)	2
Total	22	5	27

Figure 2c shows the model AEIC values for the RI events predicted by the HWRF Model at correct and incorrect times during each cycle. The AEIC ranges from 10 to 70 kt during the falsely predicted RI events, with a mean value of about 30 kt. The AEIC is substantially reduced during the correctly detected RI events, ranging from 5 to 35 kt, with a mean value of about 12 kt. When the AEIC is considered as described in section 2, the POD5 index is not affected if $\varepsilon_1 = 30$ kt (i.e., AEIC must be under 30 kt for good RI detection) except for one cycle initialized at 1800 UTC 23 August 2020 (see triangles in Fig. 2a). This occurs because AEIC for most of the detected RI events is smaller than 30 kt. If ε_1 is reduced from 30 to 10 kt, the POD5 index is significantly reduced from about 60% to 20% (red dots in Fig. 2a) for most cycles as the AEIC in most of the detected RI events is between 10 and 30 kt, as shown in Fig. 2c. For the cycle initialized at 18 UT 24 August 2020 the forecasted IC at the time of an observed RI event ($IC_{best} \geq RI_c$) is smaller than RI_c but it is very close to IC_{best} (AEIC < 3 kt). Therefore, this observed RI event is thought to be detected with $\varepsilon_2 = 3$ kt rather than missed by the forecast. As a result, the POD5 value for this cycle is adjusted from 0.8 to 1 (see crosses). In this example, the FAR5 index is not affected with $\varepsilon_3 = 3$ kt because all the AEIC values during the falsely predicted RI events are greater than 10 kt (Fig. 2c).

Implementing the thresholds as defined in section 2, the HWRF Model successfully predicts RI events in 13 cycles ($N_{CS} = 13$) if we use $F_0 = 0.5$ and $P_0 = 0.5$ to define success (i.e., POD5 ≥ 0.5 and FAR5 ≤ 0.5). In seven cycles the HWRF Model predicts RI events but the FAR5 index is high, exceeding 0.5 (FAR5 > 0.5). As a result, they are classified as conditional false alarm cycles ($n_{CFA} = 7$). The total number of cycles falsely predicting RI events is $n_{CFA} + n_2 = 11$. Therefore, the SPR is 59% [13 out of 22, see Eq. (3)], while the FPR is 44% [11 out of 25, see Eq. (4)]. SPR is greater than FPR, suggesting that the model generally has a fairly good ability to forecast RI events for this storm.

For comparison, the HFIP standard MAE-based approach is applied to each cycle of Hurricane Laura (13L). It is found that the MAE of the forecasted intensity in early cycles is larger than in late cycles (like Fig. 2c), but the MAE cannot differentiate the sources of the mean error (e.g., how much the errors are from missed, detected, or falsely predicted RI events). Therefore, the MAE approach provides little information to help diagnose the underlying causes of poorly forecasted RI events. By aggregating individual lead times, the

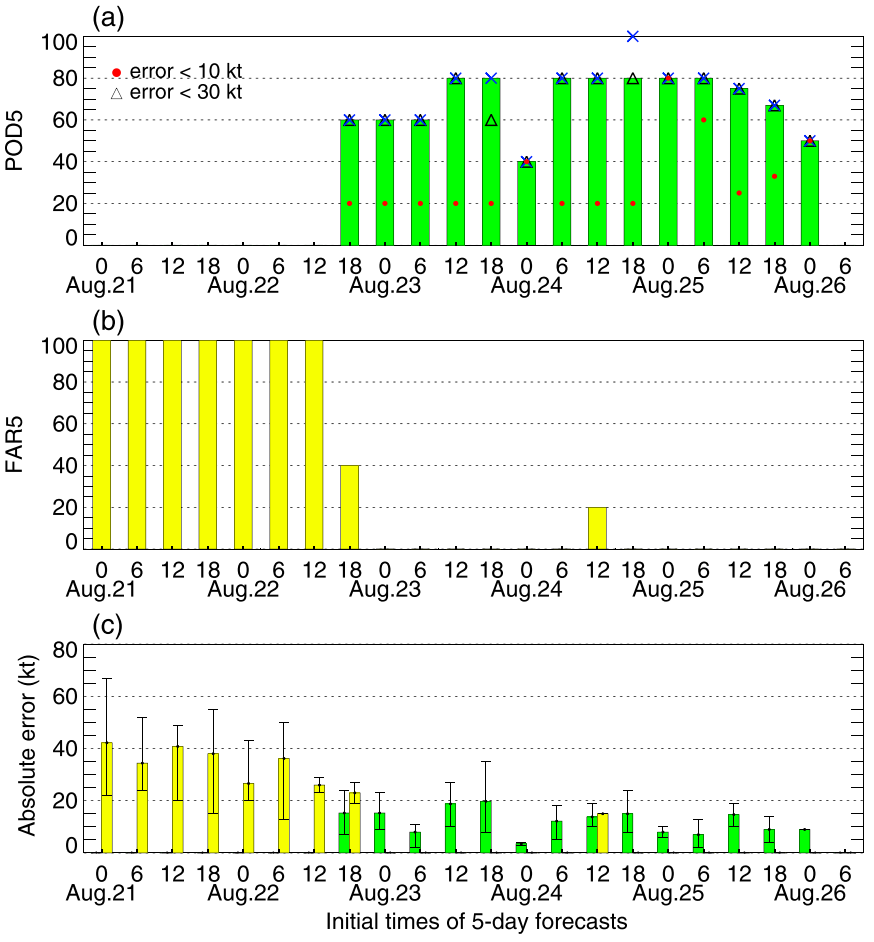

FIG. 2. (a) POD5 index for the 22 forecast cycles during which RI events are identified in the best track data for Hurricane Laura (13L) in 2020. Green bars, red filled circles, and triangles stand for the calculations without considering AEIC, with AEIC < 10 kt, and with AEIC < 30 kt for the detected RI events, respectively. Cross symbols (X) for the calculation with $\varepsilon_2 = 3$ kt to determine if IC_{mod} and IC_{best} are close. (b) FAR5 index. (c) Mean AEIC values during the RI events predicted at the correct (green bars) and incorrect (yellow bars) times. The error bars denote minimum and maximum values. RI_c is 30 kt (24 h)$^{-1}$.

traditional POD/FAR analysis provides useful information such as the statistical performance at different lead times, but it cannot decipher which cycles have issues unlike the cycle-by-cycle analysis. In addition, the cycle-based POD/FAR analysis can compare RI forecasts initialized at the same time by different configurations of a model or different TC models, which can be used to identify potential RI forecasting issues.

b. *Performance of RI forecasts during the NATL hurricane season in 2019*

Like the traditional POD/FAR-based and HFIP standard MAE-based evaluation approaches, cycle-based analysis can also evaluate the statistical performance of RI forecasts over one or more hurricane seasons in terms of percentages of successful and false RI prediction cycles. The analysis of the RI forecasts during the NATL hurricane season in 2019 is shown here as an example.

The real-time operational HWRF Model generated 357 5-day forecast cycles for the NATL hurricanes in 2019, from which 258 cycles are chosen following the sampling requirement. As summarized in Table 4, RI events [RI_c = 30 kt (24 h)$^{-1}$] are identified from the best track data in 62 5-day periods, while RI events are forecasted by the model in 69 5-day periods. There are 35 forecast cycles where RI events can be identified both from the forecasts and from the best track analysis data.

TABLE 4. Contingency table of the number of 5-day forecast cycles over the NATL basin in 2019.

	$N5_{best} > 0$	$N5_{best} = 0$	Total
$N5_{mod} > 0$	n_1 (35)	n_2 (34)	69
$N5_{mod} = 0$	n_3 (27)	n_4 (162)	189
Total	62	196	258

TABLE 5. Number of 5-day forecast cycles in the 2019 NATL basin with POD5 ≥ P_0 and FAR5 ≤ F_0 in the category of $N5_{best} > 0$ and $N5_{mod} > 0$. The results shown in parentheses are for the cases of ±6-h time matching windows (relative to the verification time).

	POD5 ≥ 0	POD5 ≥ 0.25	POD5 ≥ 0.50	POD5 ≥ 0.75	POD5 = 1.00
FAR5 = 0.00	17 (26)	11 (17)	4 (10)	0 (4)	0 (2)
FAR5 ≤ 0.25	18 (27)	12 (18)	5 (11)	0 (5)	0 (3)
FAR5 ≤ 0.50	23 (29)	16 (19)	9 (12)	1 (6)	1 (4)
FAR5 ≤ 0.75	24 (30)	17 (20)	10 (13)	2 (7)	2 (5)
FAR5 ≤ 1.00	35 (36)	21 (21)	13 (14)	2 (7)	2 (5)

The model does not produce any RI events in 27 forecast cycles where the best track data indicate that RI events occur. Therefore, the absolute failure rate is 0.44 (27 out of the total 62 RI-observed cycles). The model predicts RI events in 34 cycles where the best track data do not show any RI. The absolute false alarm rate is 0.49 (34 out of total 69 RI-forecasted cycles).

For the category of $N5_{best} > 0$ and $N5_{mod} > 0$ (35 cycles), model performance in each forecast cycle is dependent on how many RI events are predicted and how well the RI occurrence times of the predictions match the observations during the 5-day model integration period (i.e., quantified by the POD5 and FAR5 indices). Table 5 presents a joint distribution of the number of 5-day forecast cycles with different thresholds of POD5 and FAR5 indices, where the occurrence times of the RI events derived from the forecast and best track analysis exactly match. The number of cycles with successful predictions increases as the POD5 threshold (P_0) decreases or FAR5 threshold (F_0) increases. The HWRF Model can successfully predict at least 50% of the observed RI events with FAR5 less than 50% in nine cycles. There are no forecast cycles where the HWRF Model can predict all the observed RI events without a false alarm prediction. The number of cycles with successful predictions increases when the forecasted RI occurrence time is within ±6 h from the best track analysis, as shown in parentheses in Table 5. This suggests that in some cases the HWRF Model does predict RI events, but they happen too early or too late.

For the 2019 NATL hurricane season, SPR of the RI-observed cycles is 0.21 and FPR of the RI-forecasted cycles is 0.67, where P_0 and F_0 are set to 0.5. Using a ±6-h time matching window, SPR increases to 0.23 and FPR is reduced to 0.58. If the RI threshold is relaxed to 20 kt (24 h)$^{-1}$, SPR and FPR are 0.25 and 0.64, respectively, which performs better than using RI_c = 30 kt (24 h)$^{-1}$. Using a ±6-h time matching window, SPR is increased to 0.38 and FPR is reduced to 0.52.

c. *Performance in different hurricane seasons*

The above analysis is applied to each hurricane season from 2009 to 2021 over the NATL and EPAC basins, respectively, to assess the performance of the HWRF Model for RI forecasts over time. The HWRF Model was upgraded yearly over the last decade, focusing on improvements in spatial resolution, physics parameterization schemes, data assimilation system, initialization techniques, and ocean coupling.

Figure 3 presents how the SPR and FPR indices vary over the years in the NATL basin. With RI_c = 30 kt (24 h)$^{-1}$ and a zero-time window (i.e., exact time matching), the SPR index is generally between 0 and 0.4. There is a clear trend of SPR increasing over time, particularly after 2016, though it is still lower than 0.5. The SPR index increases as the time matching window is relaxed to ±6 (Fig. 3a, blue line) and ±12 h (Fig. 3b, orange line) of the verifying time, suggesting that in some forecast cycles the HWRF Model can correctly predict large intensity increases but not the exact times when RI events begin. The SPR index also increases when a lower RI threshold of 20 kt (24 h)$^{-1}$ is used (Fig. 3a, red line). This implies that the HWRF Model still has difficulty in accurately predicting large increases in intensity like other TC models (Tallapragada and Kieu 2014). This is likely due to insufficient resolution and inadequate physics parameterizations at subgrid scales despite the adjusted parameterization schemes and increased horizontal resolution (from 9 to 3 km in 2010, then to 2 km in 2015, and 1.5 km in 2018). Figure 3b shows that the FPR index of forecasted RI cycles has generally decreased over time. The FPR index can be reduced by using a relaxed time matching window or reduced RI threshold. For comparison, the conventionally used POD and FAR indices are also calculated for each hurricane season based on Eqs. (1) and (2) for all lead times (Fig. 3b, broken lines). The POD and FAR indices have the same time-varying trend as the forecast cycle–based SPR and FPR indices with a zero-time matching window and $P_0 = F_0 = 0.5$ (Fig. 3, solid black lines), despite an approximate 10%–20% difference in magnitude. Note that in addition to the different ways of aggregating data, the SPR and FPR indices can vary with P_0 and F_0 thresholds (Wang et al. 2020), while the traditional POD and FAR indices cannot.

Figure 4 shows SPR and FPR indices in the EPAC basin spanning the analysis period 2009–21. The RI forecast performance in the EPAC basin by the HWRF Model is comparable to or better than in the NATL basin before 2014, but it is worse afterward. Unlike in the NATL basin, the improving trend in SPR and FPR over time is very weak, except for the case of RI_c = 20 kt (24 h)$^{-1}$. Both SPR and FPR indices increase when using a relaxed time window or lower RI_c value, which matches the NATL basin results.

To understand the overall ability of the HWRF Model to predict TC intensification, we analyze the 95th percentile of 24-h intensity increases (W95) for the HWRF Model forecasts and the best track data for each year from 2009 to 2021 (not shown). The mean W95 value (23.4 kt) of the HWRF forecasts is very close to that (22.7 kt) of the best track analysis data in the NATL basin, indicating that the HWRF Model

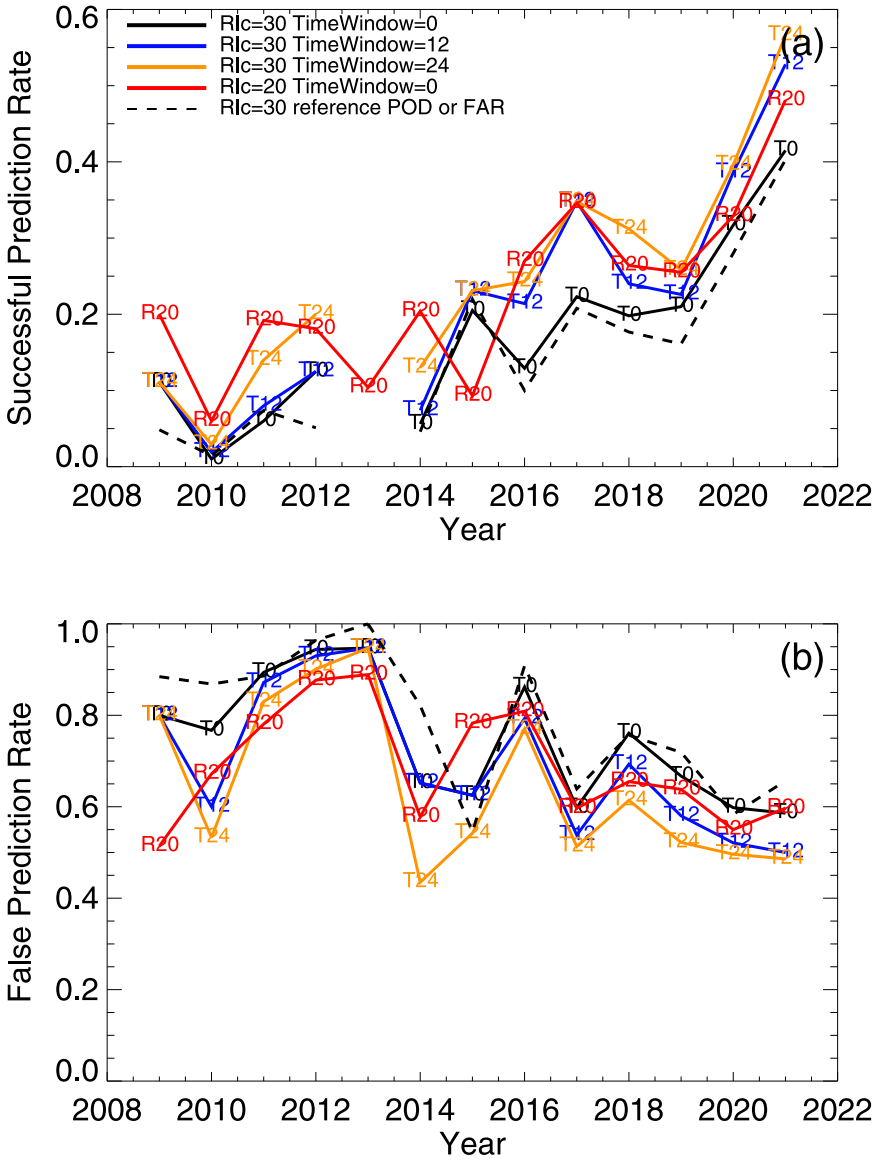

FIG. 3. (a) The SPR index of RI-observed cycles by the HWRF Model over the NATL basin for the years from 2009 to 2021 for Ri_c = 30 kt (24 h)$^{-1}$ and exact time matching (i.e., the same occurrence times) of forecasted and observed RI events (solid black line), RI_c = 30 kt (24 h)$^{-1}$ and ±6-h time matching window (blue), RI_c = 30 kt (24 h)$^{-1}$ and ±12-h time matching window (orange), as well as RI_c = 20 kt (24 h)$^{-1}$ and exact time matching (red). The broken black line shows the conventional POD index with RI_c = 30 kt (24 h)$^{-1}$ calculated by aggregating the forecasts at all lead times. (b) The FPR index of forecasted RI cycles. The calculation is not available in 2013 because no RI events can be identified that year from the best track data with the threshold of 30 kt (24 h)$^{-1}$.

can predict large intensity increases. However, the SPR index remains low (0.1–0.4) and the FPR index is high (0.6–1.0) (Fig. 3), suggesting that the HWRF Model has difficulty in predicting the correct RI occurrence times as both SPR and FAR indices improve with a relaxed time matching window. For the EPAC basin, the mean W95 value (21.3 kt) of the HWRF forecasts is smaller than that (26.5 kt) of the best track data. The HWRF Model predicts large intensity increases much less frequently than does the best track analysis; this could partially explain why SPR increases and FPR decreases significantly with a lower RI threshold value (Fig. 4).

Although the HWRF Model's ability to accurately predict RI events in terms of timing and magnitude still needs improvement, the HWRF Model has performed better over time. It is difficult to attribute specific improvements to specific model enhancements, but the following upgrades could

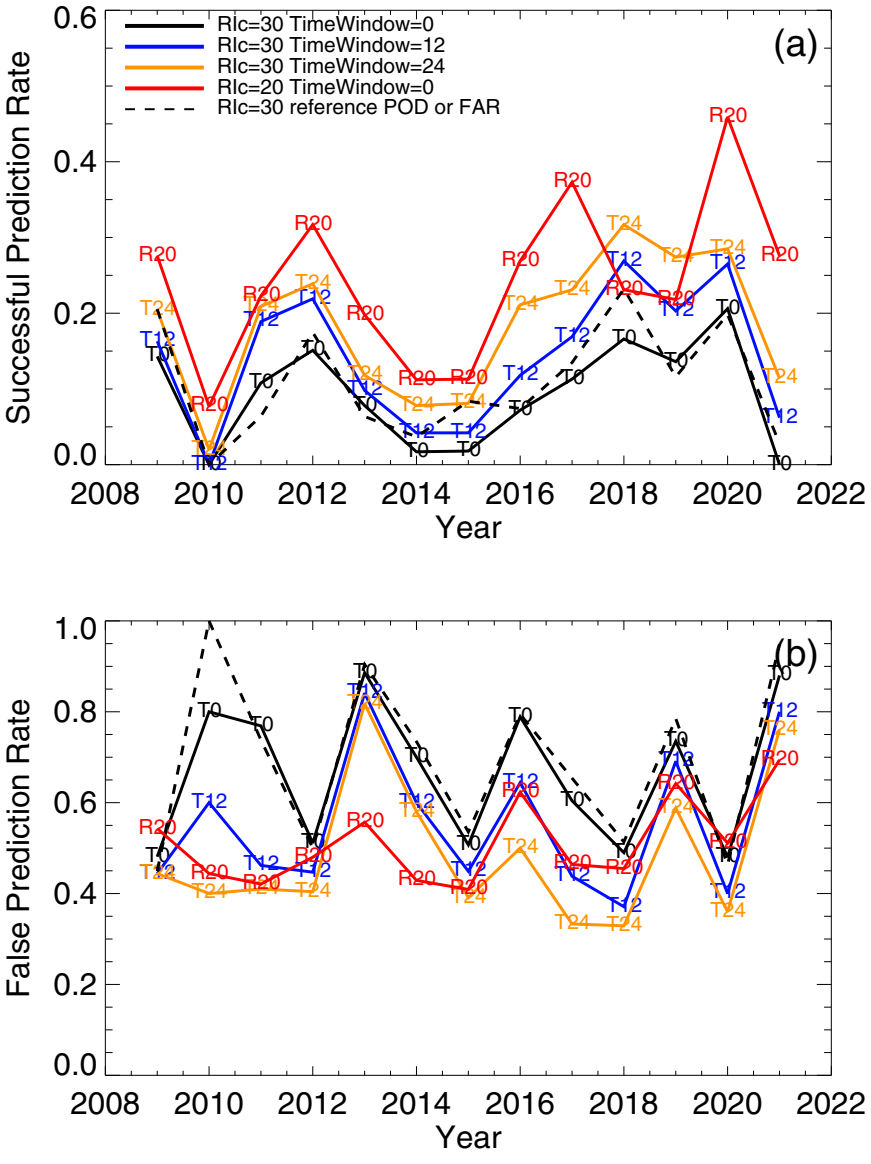

FIG. 4. As in Fig. 3, but for the EPAC basin from 2009 to 2021.

have played an important role. First, the HWRF system increased the spatial resolution three times in the past 12 years, as previously mentioned. Increasing the horizontal resolution benefits the intensity forecast and improves the potential to forecast TC intensification (Gopalakrishnan et al. 2012; Zhang et al. 2011). Second, the physical parameterization schemes have been adjusted and upgraded. For example, the eddy diffusivity in the NCEP global forecast system (GFS) PBL scheme has been adjusted based on observations, which significantly improved the intensity forecasts (Bu et al. 2017; Gopalakrishnan et al. 2013; Wang et al. 2018). Third, the ocean component of the HWRF system contributed to improvements in intensity and track forecasts (Kim et al. 2022). The ocean model upgrades included replacing the one-dimensional column model, Princeton Ocean Model (POM), with a three-dimensional model, the Hybrid Coordinate Ocean Model (HYCOM). Also, more realistic initial conditions are incorporated from the global operational Real-Time Ocean Forecast System (RTOFS) at NCEP (https://polar.ncep.noaa.gov/ofs/index.shtml). The improved coupling of oceanic and atmospheric models likely helped to reduce false RI predictions. Fourth, the initialization processes have been developed and upgraded to include a vortex initialization technique and a data assimilation system that provide more realistic and better-balanced initial conditions. The vortex initialization technique (Liu et al. 2020) was developed to correct the initial location, size, and intensity of the TC vortex. It was further improved by initializing a more realistic composite storm vortex in 2017 and was adjusted to stay consistent with the model's resolution upgrades in 2010, 2015, and 2018. The

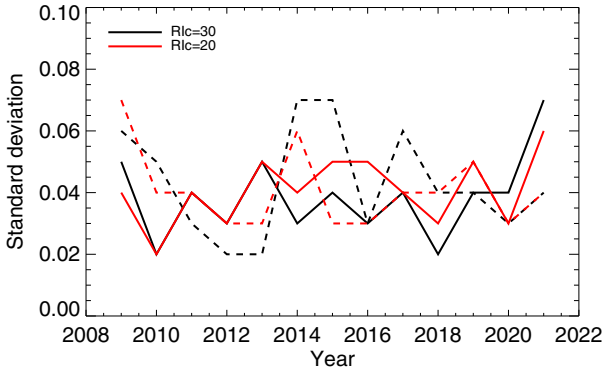

FIG. 5. The estimated standard deviations of SPR (solid lines) and FPR (broken lines) for RI_c = 20 kt (24 h)$^{-1}$ (red) and 30 kt (24 h)$^{-1}$ (black), assuming that relative errors of the maximum 10-m wind speeds from the HWRF Model and the best track analysis are 10%.

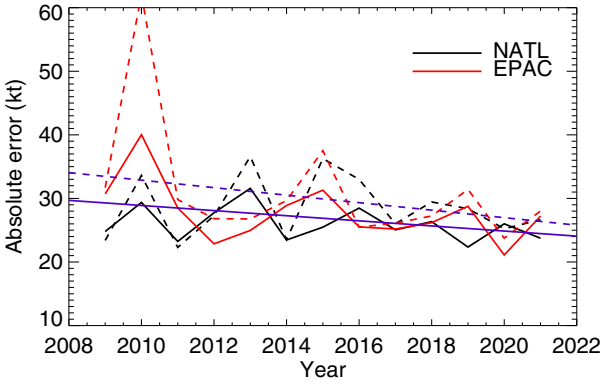

FIG. 6. The MAEs of the intensity (broken line) and 24-h intensity change (solid line) during the observed or forecasted RI events of individual RI cycles in the NATL (black) and EPAC (red) basins each year from 2009 to 2021. The blue lines are linearly fitted trends for the MAEs of the intensity (broken line) and 24-h intensity change (solid line) using the combined data of two basins.

HWRF's data assimilation system adopted the NCEP GFS's gridpoint statistical interpolation (GSI) assimilation technique for nested domains in the early years. With development and upgrades over many years, it has evolved into a more sophisticated stochastic physics-based hybrid GSI ensemble Kalman filter (EnKF) system (Zhang et al. 2020), which can run 40-member HWRF ensembles to provide more accurate data assimilation covariance. The system can assimilate observational data from a variety of platforms, such as satellites, radars, and aircraft. There are still many ongoing efforts, under potential research to operations, to further improve the performance of the HWRF Model or the next-generation hurricane model and provide more reliable guidance to operational TC forecasters (e.g., Lewis et al. 2020; Ma et al. 2020; Wang et al. 2021, 2022).

d. Uncertainty and comparisons

The RI forecast evaluations can be affected by uncertainties in intensity from the best track analysis (Landsea and Franklin 2013; Torn and Snyder 2012) and the HWRF Model forecast (Zhang et al. 2021b). For example, intensity in the HWRF Model is derived from the instantaneous fields every 3 or 6 h using Geophysical Fluid Dynamics Laboratory (GFDL)'s vortex tracker (Biswas et al. 2018; Marchok 2002); this could result in an error of 8% due to large fluctuations in the 10-m maximum wind speed with time (Zhang et al. 2021b). To estimate the impact of data uncertainties on the RI forecast evaluation, 20 sets of normally distributed random errors with a mean of zero and a standard deviation of 10% for the 10-m wind speed were added to the intensity data from the model forecast and best track analysis. The SPR and FPR indices were recalculated for each set of the error-adjusted datasets. The resulting standard deviations of SPR and FPR range from 0.02 to 0.07 (Fig. 5), which is about 5%–10% of the mean SPR and FPR values.

Because HFIP adopted the MAE-based metric as a standard to measure programmatic progress in forecasting RI, we apply the same method to calculate MAEs of the intensity and 24-h IC during observed or forecasted RI events for each cycle. Figure 6 shows decreasing trends for the MAEs for intensity and 24-h IC at rates of 0.59 and 0.40 kt yr^{-1}, respectively. However, the trends in forecast improvement from MAE-based calculations are smaller than in SPR and FPR, as shown in Figs. 3 and 4, which can be explained by two reasons. First, SPR and FPR are based on category classification where it is easier to detect improvements compared to MAEs since the intensity forecast is not yet sufficiently accurate. Second, the HFIP standard method combines data for missed, detected, and falsely predicted RI events, which averages all improvements, degradations, or both.

To examine the relationship between the probability-based metric and MAE-based metric, a synthesized index (SI), defined as POD5 + (1 − FAR5), is used to represent the probability-based metric, with an assumption that POD5 and FAR5 are equally considered in the RI forecast evaluation. A higher SI suggests that RI events could be forecasted better during a cycle, where a zero value indicates the worst forecast and a value of two represents the best forecast. Due to large variability in the data, MAEs are averaged over SI intervals of 0.25. Figure 7 shows that the variations of the bin-averaged MAEs in intensity (broken line) and 24-h IC (solid line) with SI are similar. There is a trend that both the bin-averaged MAEs decrease with SI, which can be fitted by a linear function (green line), MAE = −8.34SI + 30.5, with the coefficient of determination (r^2) of 0.49. The decreasing function suggests that both metrics for RI forecast performance can give similar bulk performance evaluations in that both a higher SI and a lower MAE indicate better performance of RI forecasts. However, it should be kept in mind that the decreasing relation of MAE and SI might not be valid for individual cycles or when the sample size of RI cycles is small due to large variability of data in each bin as shown in Fig. 7. Also, the correlation between MAE and SI is weaker when SI < 1 (i.e., POD5 is smaller than FAR5) than when SI ≥ 1. Therefore, a smaller MAE might not represent a better detection of RI in terms of

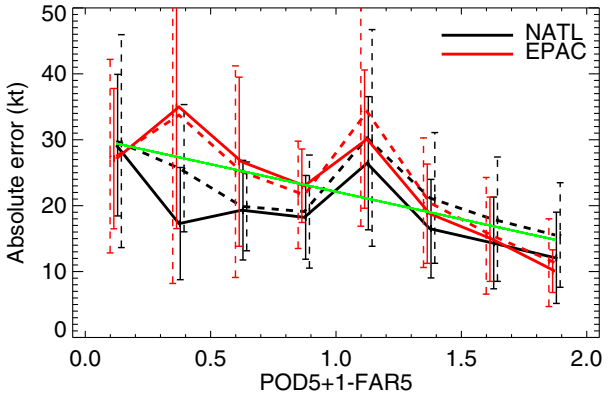

FIG. 7. The bin-averaged MAE in intensity (broken lines) and 24-h intensity change (solid lines) during observed or forecasted RI events in individual 5-day cycles varying with the synthesized index, POD5 + 1 − FAR5, in the NATL (black lines) and EPAC (red lines) basins, with a bin size of 0.25. Error bars denote standard deviations of the absolute errors in the respective bins. The green line is the linearly fitted trend for the bin-averaged MAE and POD5 + 1 − FAR5 using the combined data of two basins and MAEs in intensity and 24-h intensity change.

POD5 and FAR5 and vice versa when we interpret the evaluation results based on different metrics for individual cycles in practice, given that RI sample sizes even in a year are usually small. It is worth further investigation on how to characterize model performance in RI forecasts by synthesizing MAE-based and probability-based metrics, given that both have advantages and disadvantages.

4. Summary and future work

An evaluation framework has been introduced to assess RI forecast performance in individual 5-day forecast cycles rather than individual lead times. This evaluation method is different from the traditional RI evaluation in the following two respects. First, RI evaluation is conducted in the context of individual 5-day forecast cycles. This method can be used to compare RI forecasts initialized at different times from a model, and RI forecasts initialized at the same time from different configurations or different models. The cycle-based analysis allows modelers and forecasters to investigate the underlying causes of good or bad forecasts in the post-storm analyses. In practice, examining the entirety of a 5-day forecast can help forecasters to identify RI forecast challenges related to specific scenarios and to recognize scenarios where a model is likely to perform well or poorly. The cycle-based framework also can provide seasonal RI forecast statistical performance in terms of the percentages of forecast cycles with correctly or incorrectly forecasted RI events. The evaluation of RI forecast performance is traditionally based on the statistics for individual lead times rather than forecast cycles. Therefore, the cycle-based framework provides an evaluation from a different perspective. Second, the evaluation of individual cycles incorporates AEIC and it is not solely based on the POD and FAR indices for intensification rates that are equal to or greater than RI_c. The integration of AEIC into the evaluation can partially address two misleading scenarios when an evaluation is based only on the POD and FAR indices. One scenario occurs when the model forecasts a real RI event with a very large error, but it is still classified as a success. The other scenario is when the forecasted IC is very close to the observation, but the forecast is still classified as a failure if it is below RI_c or as a false alarm if it is larger than RI_c. Like the traditional POD/FAR-based approach, the cycle-based evaluation is dependent on the RI threshold, time window for matching the observed and forecasted RI events, and the uncertainty of data. In addition, the cycle-based evaluation can vary with the choice of thresholds for AEIC, POD5, and FAR5.

The new framework has been applied to analyze RI forecast performance for an individual forecast cycle, the forecast cycle spanning an entire storm, and all forecast cycles from past hurricane seasons by the operational HWRF Model. It has been shown that the overall HWRF Model performance in predicting RI has improved over the years, particularly for the NATL basin. The improvement trends are consistent with those from the existing approaches such as the traditional POD/FAR indices and HFIP standard MAE-based evaluation. In addition, the cycle-based approach provides further information on the percentages of successful and false prediction cycles. While it was initially developed for the deterministic HWRF Model, this cycle-based RI evaluation framework also can be used to evaluate ensemble forecast systems. For a given forecast cycle, the cycle-based analysis can assess how many members of an ensemble system successfully or falsely forecast RI events (Zhang et al. 2021a).

Although the AEIC has been integrated into the calculations of POD5 and FAR5 indices, the cycle-based evaluation still has the limitations of the traditional POD/FAR method, such as a lack of connection to the track forecasts and a dependence on the cut-off value. In addition, there are two main limitations compared to the current existing evaluation approaches for RI forecast performance. First, the analysis for 5-day forecast cycles requires at least 48 h of data during each cycle, which reduces the sample size. The existing approaches alleviate this limitation by sampling data at individual lead times. Second, the cycle-based framework aggregates the successes and failures across the 5-day forecast interval and thus all timing information is lost for storm-total or season-total statistics. Meanwhile, the existing approaches can preserve timing information. Nevertheless, a cycle-based framework can provide a convenient way to compare RI forecasts of different cycles or the same cycles of different models as well as statistical information from a different perspective. In practice, multiple approaches can be used together for a comprehensive evaluation from varying perspectives.

The cycle-based framework can be further improved by using a combination of MAE-based and probability-based metrics to evaluate the success or failure of an RI cycle. To find the relationship between the probability-based and MAE-based metrics, we calculate the MAE of IC (and intensity) during observed or forecasted RI events and a synthesized index based on POD5 and FAR5 for individual cycles. It is

found that both metrics can give similar bulk evaluation results. Nevertheless, the two metrics might not give consistent evaluations for individual cycles or a small sample size due to the large variability of MAEs at a given SI. This should be kept in mind when interpreting in practice the evaluation results with the two different metrics, given that RI sample sizes even in a year are usually small. Therefore, it is challenging to synthesize probability-based and MAE-based metrics to reflect both forecast errors in intensity and the probabilities of RI detections and false alarms. For example, POD and FAR can be integrated into the calculation of MAE in intensity by using different weights for missed, detected, or falsely predicted RI events. Another improvement in the cycle-based framework is to quantify the usefulness of RI forecasts. To characterize the degree to which RI forecasts are useful, SPR and FPR can be combined in different ways, depending on the importance model developers and forecasters put on the detection or false alarms of RI events.

Acknowledgments. This work is supported by NOAA's Hurricane Forecast Improvement Project. Thanks are due to Mary Hart for editing the manuscript and Drs. Junghong Shin and Qingfu Liu for providing insightful comments during the internal review process. The authors also thank three reviewers for their critical and insightful comments and suggestions. We are very grateful to our colleague, Dr. John Steffen, for comments and carefully proofreading the manuscript.

Data availability statement. All datasets and results produced by the HWRF Model used in this study are available for public release upon request.

REFERENCES

Biswas, M. K., L. Bernardet, S. Abarca, I. Ginis, E. Grell, E. Kalina, and Z. Zhang, 2017: Hurricane Weather Research and Forecasting (HWRF) Model: 2017 scientific documentation. NCAR Tech. Note NCAR/TN-544+STR, 111 pp., https://doi.org/10.5065/D6MK6BPR.

——, D. Stark, and L. Carson, 2018: GFDL vortex tracker users' guide version 3.9a. Developmental Testbed Center, 35 pp., https://dtcenter.org/sites/default/files/community-code/gfdl/standalone_tracker_UG_v3.9a.pdf.

Bu, Y. P., R. G. Fovell, and K. L. Corbosiero, 2017: The influences of boundary layer mixing and cloud-radiative forcing on tropical cyclone size. *J. Atmos. Sci.*, **74**, 1273–1292, https://doi.org/10.1175/JAS-D-16-0231.1.

Callaghan, J., 2017: Asymmetric inner core convection leading to tropical cyclone intensification. *Trop. Cyclone Res. Rev.*, **6**, 55–66, https://doi.org/10.6057/2017TCRRh3.02.

Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Wea. Forecasting*, **35**, 1913–1922, https://doi.org/10.1175/WAF-D-20-0059.1.

DeMaria, M., C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, https://doi.org/10.1175/BAMS-D-12-00240.1.

——, J. L. Franklin, M. J. Onderlinde, and J. Kaplan, 2021: Operational forecasting of tropical cyclone rapid intensification at the National Hurricane Center. *Atmosphere*, **12**, 683, https://doi.org/10.3390/atmos12060683.

Domingues, R., and Coauthors, 2019: Ocean observations in support of studies and forecasts of tropical and extratropical cyclones. *Front. Mar. Sci.*, **6**, 446, https://doi.org/10.3389/fmars.2019.00446.

Finocchio, P. M., S. J. Majumdar, D. S. Nolan, and M. Iskandarani, 2016: Idealized tropical cyclone responses to the height and depth of environmental vertical wind shear. *Mon. Wea. Rev.*, **144**, 2155–2175, https://doi.org/10.1175/MWR-D-15-0320.1.

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, https://doi.org/10.1175/BAMS-D-12-00071.1.

Gopalakrishnan, S. G., S. Goldenberg, T. Quirino, X. Zhang, F. Marks Jr., K.-S. Yeh, R. Altas, and T. 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, J. A. Zing, 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.

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.

Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108, https://doi.org/10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2.

——, ——, and J. A. Knaff, 2010: A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **25**, 220–241, https://doi.org/10.1175/2009WAF2222280.1.

——, and Coauthors, 2015: Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Wea. Forecasting*, **30**, 1374–1396, https://doi.org/10.1175/WAF-D-15-0032.1.

Kim, H.-S., C. Lozano, V. Tallapragada, D. Iredell, D. Sheinin, H. L. Tolman, V. M. Gerald, and J. Sims, 2014: Performance of ocean simulations in the coupled HWRF–HYCOM model. *J. Atmos. Oceanic Technol.*, **31**, 545–559, https://doi.org/10.1175/JTECH-D-13-00013.1.

——, J. Meixner, B. Thomas, B. Reichl, B. Liu, A. Mehra, and A. Wallcraft, 2022: Skill assessment of NCEP three-way coupled HWRF-HYCOM-WW3 modeling system: Hurricane Laura case study. *Wea. Forecasting*, **37**, 1309–1331, https://doi.org/10.1175/WAF-D-21-0191.1.

Knaff, J. A., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting*, **18**, 80–92, https://doi.org/10.1175/1520-0434(2003)018<0080:SDTCIF>2.0.CO;2.

——, C. R. Sampson, and K. D. Musgrave, 2018: An operational rapid intensification prediction aid for the western North

Pacific. *Wea. Forecasting*, **33**, 799–811, https://doi.org/10.1175/WAF-D-18-0012.1.

——, ——, and B. R. Strahl, 2020: A tropical cyclone rapid intensification prediction aid for the Joint Typhoon Warning Center's areas of responsibility. *Wea. Forecasting*, **35**, 1173–1185, https://doi.org/10.1175/WAF-D-19-0228.1.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

Lewis, W. E., C. Rozoff, S. Alessandrini, and L. Delle Monache, 2020: Performance of the HWRF rapid intensification analog ensemble (HWRF RI-AnEn) during the 2017 and 2018 HFIP real-time demonstrations. *Wea. Forecasting*, **35**, 841–856, https://doi.org/10.1175/WAF-D-19-0037.1.

Liu, Q., and Coauthors, 2020: Vortex initialization in the NCEP operational hurricane models. *Atmosphere*, **11**, 968, https://doi.org/10.3390/atmos11090968.

Ma, Z., and Coauthors, 2020: Investigating the impact of high-resolution land–sea masks on hurricane forecasts in HWRF. *Atmosphere*, **11**, 888, https://doi.org/10.3390/atmos11090888.

Marchok, T. P., 2002: How the NCEP tropical cyclone tracker works. *25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., P1.13, https://ams.confex.com/ams/25HURR/techprogram/paper_37628.htm.

Mehra, A., V. Tallapragada, Z. Zhang, B. Liu, L. Zhu, W. Wang, and H.-S. Kim, 2018: Advancing the state of the art in operational tropical cyclone forecasting at NCEP. *Trop. Cyclone Res. Rev.*, **7**, 51–56, https://doi.org/10.6057/2018TCRR01.06.

Rios-Berrios, R., C. A. Davis, and R. D. Torn, 2018: A hypothesis for the intensification of tropical cyclones under moderate vertical wind shear. *J. Atmos. Sci.*, **75**, 4149–4173, https://doi.org/10.1175/JAS-D-18-0070.1.

Rozoff, C. M., C. S. Velden, J. Kaplan, J. P. Kossin, and A. J. Wimmers, 2015: Improvements in the probabilistic prediction of tropical cyclone rapid intensification with passive microwave observations. *Wea. Forecasting*, **30**, 1016–1038, https://doi.org/10.1175/WAF-D-14-00109.1.

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

Tallapragada, V., 2016: Overview of the NOAA/NCEP operational Hurricane Weather Research and Forecast (HWRF) modelling system. *Advanced Numerical Modeling and Data Assimilation Techniques for Tropical Cyclone Prediction*, U. C. Mohanty and S. G. Gopalakrishnan, Eds., Springer, 51–106.

——, and C. Kieu, 2014: Real-time forecasts of typhoon rapid intensification in the North Western Pacific basin with the NCEP operational HWRF model. *Trop. Cyclone Res. Rev.*, **3**, 63–77, https://doi.org/10.6057/2014TCRR02.01.

——, and Coauthors, 2015: Forecasting tropical cyclones in the western North Pacific basin using the NCEP operational HWRF: Real-time implementation in 2012. *Wea. Forecasting*, **30**, 1355–1373, https://doi.org/10.1175/WAF-D-14-00138.1.

——, and Coauthors, 2016: Forecasting tropical cyclones in the western North Pacific basin using the NCEP operational HWRF model: Model upgrades and evaluation of real-time performance in 2013. *Wea. Forecasting*, **31**, 877–894, https://doi.org/10.1175/WAF-D-14-00139.1.

Torn, R. D., and C. Snyder, 2012: Uncertainty of tropical cyclone best-track information. *Wea. Forecasting*, **27**, 715–729, https://doi.org/10.1175/WAF-D-11-00085.1.

Wang, W., J. A. Sippel, S. Abarca, L. Zhu, B. Liu, Z. Zhang, A. Mehra, and V. Tallapragada, 2018: Improving NCEP HWRF simulations of surface wind and inflow angle in the eyewall area. *Wea. Forecasting*, **33**, 887–898, https://doi.org/10.1175/WAF-D-17-0115.1.

——, L. Zhu, H. S. Kim, D. Iredell, J. Dong, Z. Zhang, A. Mehra, and V. Tallapragada, 2019: NCEP HMON-based hurricane ensemble forecast system. Research activities in atmospheric and oceanic modelling, CAS/JSC Working Group on Numerical Experimentation, Rep. 49, WCRP Rep.12/2019, WMO, Geneva, Switzerland, 5–15, https://wgne.net/publications/wgne-blue-book/.

——, B. Liu, Z. Zhang, L. Zhu, A. Mehra, and V. Tallapragada, 2020: Ten-year performance of HWRF Model in RI Forecasts—A new metric. Research activities in Earth system modelling, Working Group on Numerical Experimentation Rep. 50, WCRP Rep.4/2020, 10-09, https://wgne.net/publications/wgne-blue-book/.

——, ——, L. Zhu, Z. Zhang, A. Mehra, and V. Tallapragada, 2021: A new horizontal mixing-length formulation for numerical simulations of tropical cyclones. *Wea. Forecasting*, **36**, 679–695, https://doi.org/10.1175/WAF-D-20-0134.1.

——, ——, Z. Zhang, A. Mehra, and V. Tallapragada, 2022: Improving low-level wind simulations of tropical cyclones by a regional Hurricane Analysis and Forecast System. Research activities in Earth system modelling, Working Group on Numerical Experimentation, Rep. 52, WCRP Rep.4/2022, 4-09, https://wgne.net/publications/wgne-blue-book/.

Willoughby, H. E., J. A. Clos, and M. G. Shoreibah, 1982: Concentric eye walls, secondary wind maxima, and the evolution of the hurricane vortex. *J. Atmos. Sci.*, **39**, 395–411, https://doi.org/10.1175/1520-0469(1982)039<0395:CEWSWM>2.0.CO;2.

Zhang, X. J., K. S. Yeh, T. S. Quirino, S. G. Gopalakrishnan, F. D. Marks, S. B. Goldenberg, and S. Aberson, 2011: HWRFx: Improving hurricane forecasts with high-resolution modeling. *Comput. Sci. Eng.*, **13**, 13–21, https://doi.org/10.1109/MCSE.2010.121.

Zhang, Z., and Coauthors, 2020: The impact of stochastic physics-based hybrid GSI/EnKF data assimilation on hurricane forecasts using EMC operational hurricane modeling system. *Atmosphere*, **11**, 801, https://doi.org/10.3390/atmos11080801.

——, W. Wang, B. Liu, L. Zhu, A. Mehra, and V. Tallapragada, 2021a: Performance of HAFS-based ensemble prediction system (HAFSv0.2E) in 2021 Atlantic hurricane season. *2021 HFIP Annual Meeting*, online, National Oceanic and Atmospheric Administration, 15 pp., https://hfip.org/sites/default/files/events/269/1245-zhanzhang-hafsv02e.pdf.

——, J. A. Zhang, G. J. Alaka, K. Wu, A. Mehra, and V. Tallapragada, 2021b: A statistical analysis of high-frequency track and intensity forecasts from NOAA's operational Hurricane Weather Research and Forecasting (HWRF) modeling system. *Mon. Wea. Rev.*, **149**, 3325–3339, https://doi.org/10.1175/MWR-D-21-0021.1.