A Global Climatology of Extratropical Transition. Part II: Statistical Performance of the Cyclone Phase Space

MELANIE BIELI

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

SUZANA J. CAMARGO

Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York

ADAM H. SOBEL

Department of Applied Physics and Applied Mathematics, and Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York

JENNI L. EVANS

Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania

TIMOTHY HALL

NASA Goddard Institute for Space Studies, New York, New York

(Manuscript received 30 January 2018, in final form 15 March 2019)

ABSTRACT

This study analyzes the differences between an objective, automated identification of tropical cyclones (TCs) that undergo extratropical transition (ET), and the designation of ET determined subjectively by human forecasters in best track data in all basins globally. The objective identification of ET is based on the cyclone phase space (CPS), calculated from the Japanese 55-yr Reanalysis (JRA-55) or the ECMWF interim reanalysis (ERA-Interim). The resulting classification into ET storms and non-ET storms underlies the global climatology of ET presented in Part I of this study. Here, the authors investigate how well the CPS classifications agree with those in the best track records calculated from JRA-55 or from ERA-Interim data. According to F1 scores and Matthews correlation coefficients (MCCs), the classification of ET storms in the CPS agrees best with the best track classification in the western North Pacific (MCC $>$ 0.7) and the North Atlantic (MCC $>$ 0.5). In other basins, the correlation between the CPS classification and the best track classification is only slightly higher than that of a random classification. The JRA-55 classification achieves higher performance scores than does the ERA-Interim classification, and the differences are statistically significant in all basins. The lower performance of ERA-Interim is mainly due to a higher false alarm rate, particularly in the eastern North Pacific. Overall, the results show that while the CPS-based classifications are good enough to be useful for many purposes, there is almost certainly room for improvement—in the representation of the storms in reanalyses, in our objective metrics of ET, and in our scientific understanding of the ET process.

1. Introduction

Extratropical transition (ET) is a process in which a tropical cyclone (TC) loses its radially symmetric warm-core structure and becomes an extratropical cyclone with frontal features and a cold core (Jones et al. 2003;

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-18-0052.1.s1.

Corresponding author: Melanie Bieli, mb4036@columbia.edu

DOI: 10.1175/JCLI-D-18-0052.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Evans et al. 2017). To identify the ET of individual storms, forecasters in TC warning centers analyze a wide range of satellite images, model output, and observations. In the TC best track archives, a storm that is determined to have completed ET based on this analysis (and after a poststorm review taking into account all available data) receives an ''extratropical'' label.

The exact procedure for determining whether a cyclone is considered tropical or extratropical varies among different TC warning centers. Usually, the decision is based on a combination of satellite imagery, model forecast fields, and other operational tools such as the CPS; Fogarty (2010) provides an overview of ET-related operational forecast practices in many agencies. Examples of satellite products consulted in ET forecasts include cloud imagery, wind retrievals from scatterometers, or Advanced Microwave Sounding Unit temperature and moisture soundings. These products are used to monitor the defining characteristics of ET: the increasing asymmetry of the cloud pattern, expansion of the wind field, intrusion of dry air from the midlatitude trough, and the erosion of the TC's warm core structure (Fogarty 2010). Sometimes a ''human dimension'' may be included because public perception of a cyclone's threat changes when the system is declared extratropical (Masson 2014). The ''extratropical'' labels thus represent a definition of ET that involves subjective expert judgment. In contrast, the cyclone phase space (CPS) framework proposed by Hart (2003) can be used to define ET in a purely objective, automatable way. The CPS has become widely used and has been applied to operational analysis and reanalysis data (e.g., Hart 2003; Kitabatake 2011; Wood and Ritchie 2014) as well as climate model output (Zarzycki et al. 2017; Liu et al. 2017).

In the first part of this study (Bieli et al. 2019, hereafter Part I), we used two reanalyses, the Japanese 55-yr Reanalysis (JRA-55; Kobayashi et al. 2015) and the European Centre for Medium-Range Weather Forecasts' (ECMWF) interim reanalysis (ERA-Interim; Dee et al. 2011), to locate TCs in the CPS and study ET in seven global ocean basins. For comparison, statistics obtained from the storm type information (i.e., the ''extratropical'' labels) in the TC best track data were included as well. The resulting geographical, seasonal, and temporal characteristics of ET differed between the basins, but also between the two reanalyses and the best track labels. This raises the question to what extent the globally consistent view obtained from a reanalysis is consistent from one reanalysis dataset to another and also with forecaster judgment.

Objective definitions for the onset and completion of ET in the CPS were developed by Evans and Hart (2003) using 61 Atlantic TCs, all of which had been declared by the National Hurricane Center (NHC) to have undergone ET. The study includes a comparison of the timing of ET in the CPS with that in the best track data from the NHC. However, Evans and Hart (2003) did not examine how the classification into ''ET storms'' (i.e., storms that undergo ET at some point in their lifetimes) and ''non-ET storms'' (i.e., storms that do not undergo ET) obtained from the CPS compares to that in the best tracks, when considering a set of TCs with unknown classification. Applying the CPS to identify ET in a set of recurving TCs, Kofron et al. (2010) found that the CPS does not discriminate between ET storms and non-ET storms. However, their definition of ET is not based on the best track labels but on a manual examination of each cyclone's surface pressure field in reanalysis data.

The dependence on the dataset used to locate the TCs in the CPS makes it difficult to isolate the effect of the methodological differences between the definition of ET in the CPS and that in the best tracks. An example of this is the fraction of TCs undergoing ET as presented in Part I: The classification obtained from ERA-Interim diagnoses a larger number of storms as undergoing ET than does the JRA-55 classification. As there is no universal definition of ET, it is not possible to assess the correctness of the two classifications in absolute terms. However, we can evaluate how well the CPS classifications agree with the best track records, and how that agreement depends on whether the CPS is calculated from JRA-55 or from ERA-Interim data. This second part of the study sets out to answer these questions on a global basis.

2. Data and methods

a. TC best track and reanalysis datasets

This study is based on the same data as Part I: The cyclone data are best track datasets from the National Hurricane Center in the North Atlantic (NAT) and in the eastern North Pacific (ENP), from the Joint Typhoon Warning Center (JTWC) in the north Indian Ocean (NI), the Southern Hemisphere (SH), and the western North Pacific (WNP), and from the Japan Meteorological Agency (JMA) in the WNP. Within the SH, we distinguish the south Indian Ocean (SI), the Australian region (AUS), and the South Pacific (SP). Table 1 provides an overview of the basin acronyms and best track datasets used in this study.

In Part I, we considered TCs with tropical storm intensity or higher that occurred in the satellite era 1979–2017. Here, we consider only the years for which the best track data provide the ''extratropical'' labels that denote TCs that have undergone ET, as declared by the

TABLE 1. Definitions and acronyms of the ocean basins examined in this study, including their sources of best track datasets, time period for which "extratropical" labels are available in the best track data, and number of storms in that time period.

Basin	Code	Source of best tracks	Availability of "extratropical" labels	No. of storms
North Atlantic	NAT	NHC	1979–2017	481
Western North Pacific	WNP	JMA, JTWC	1979–2017, 2004–17	994, 331
Eastern North Pacific	ENP	NHC	1988–2017	492
North Indian Ocean	NI	JTWC	2004–17	74
South Indian Ocean	SI	JTWC	2004–17	117
Australian region	AUS	JTWC	2004–17	122
South Pacific	SP	JTWC	2004–17	73

respective operational meteorological agencies. The time periods for which these labels are available vary by basin (Table 1).

We use two reanalysis datasets, the Japanese 55-yr Reanalysis (1.25° × 1.25°) released by the JMA (Kobayashi et al. 2015) and the ECMWF interim reanalysis (0.7° × 0.7°; Dee et al. 2011). Both reanalyses apply a four-dimensional variational data assimilation. A unique feature of the JRA-55 assimilation system is the use of artificial wind profile retrievals in the vicinity of TCs. In this retrieval scheme, three wind models are combined to reconstruct 3D wind profile data at certain locations around the storm center, using TC information from best track data (Fiorino 2002). In the assimilation process, the wind profiles are treated as if they were observations from dropwindsondes (Hatsushika et al. 2006; Ebita et al. 2011). In contrast, ERA-Interim does not assimilate any artificial TC information.

b. Cyclone phase space

We use the cyclone phase space proposed by Hart (2003) to objectively identify storms that undergo ET. In the CPS framework, the physical structure of cyclones is described based on three parameters: the B parameter measures the asymmetry in the layer-mean temperature surrounding the cyclone, and two thermal wind ($-V_T$) parameters assess whether the cyclone has a warm or cold core structure in the upper ($-V_T^U$) and lower ($-V_T^L$) troposphere (with the convention of the minus sign, positive values correspond to warm cores). As in Part I, ET onset is defined here as the first time a TC is either asymmetric ($B > 11$) or has a cold core ($-V_T^L < 0$ and $-V_T^U < 0$), and ET completion is defined as the time when the second criterion is met. This definition allows us to distinguish three pathways of ET in the CPS: $B \rightarrow V_T$ ETs start when the TC becomes asymmetric and end with the formation of a cold core, $V_T \rightarrow B$ ETs start with the formation of a cold core and end when the TC becomes asymmetric, and direct ETs become asymmetric and cold core at the same 6-hourly time step. The reader is referred to Hart (2003) and Evans and Hart (2003) for a comprehensive exposition of the CPS, and to Part I for details on its application to the definition of ET in this study.

After computing the CPS parameters along all best tracks, we applied the CPS criteria to classify each storm either as an ET storm if it completes the transition from a tropical to an extratropical system at some point during its lifetime or as a non-ET storm if it does not. This resulted in two binary classifications, one from the CPS parameters computed using JRA-55 data (the JRA-55 classifier), and one from the CPS parameters obtained from ERA-Interim data (the ERA-Interim classifier). A third is given by the storm type information in the best track archives, whose "extratropical" labels represent the classification proposed by the specialists at the operational warning centers.

c. Statistical performance measures

For the purpose of this study, we treat the best track labels as the "true" classifications of ET storms (see section 4 for a discussion of this assumption). Consequently, the performance of the CPS classifiers is assessed by comparing them to the ET events in the best track labels, both by checking the agreement on individual storms as well as by applying statistical performance measures. Two commonly used statistical performance metrics for binary classification algorithms are precision and recall (e.g., Ting 2010), which are defined as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

TP, FP, and FN are the numbers of true positives, false positives, and false negatives. Thus, precision is the ratio of correctly classified positive observations (here: ET storms) to the total observations classified as positive, and answers the question, "Of all storms a CPS classifier declares to have undergone ET, what fraction actually did?" Recall is the ratio of correctly classified positive observations to the total positive observations,

TABLE 2. Evaluation of the ET events determined in the CPS against those defined in the best track datasets: breakdown into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Values are given as storm counts and as percentages of the total number of storms. Basins for which the values are based on shorter time periods (2004–17 for the NI, SI, AUS and SP; 1988–2017 for the ENP) are marked with an asterisk.

	JRA-55				ERA-Interim			
Basin	TP	TN	FP	FN	TP	TN	FP	FN
NAT	169 (35.1%)	210 (43.7%)	58 (12.1%)	44 (9.1%)	181 (37.6%)	188 (39.1%)	80 (16.6%)	32 (6.7%)
WNP (JMA)	426 (42.9%)	475 (47.8%)	44 (4.4%)	49 (4.9%)	444 (44.7%)	409 (41.1%)	110 (11.1%)	31 (3.1%)
WNP (JTWC)*	96 (29.0%)	168 (50.8%)	20 (6.0%)	47 (14.2%)	102 (30.8%)	139 (42.0%)	49 (14.8%)	41 (12.4%)
ENP	5 (1.0%)	445 (90.4%)	38 (7.7%)	4 (0.8%)	6 (1.2%)	353 (71.7%)	130 (26.4%)	3 (0.6%)
NI*	1 (1.4%)	68 (91.9%)	4 (5.4%)	1 (1.4%)	1 (1.4%)	62 (83.8%)	10 (13.5%)	1 (1.4%)
SI*	16 (13.7%)	75 (64.1%)	11 (9.4%)	15 (12.8%)	14 (12.0%)	67 (57.3%)	19 (16.2%)	17 (14.5%)
AUS*	12 (9.8%)	92 (75.4%)	7 (5.7%)	11 (9.0%)	13 (10.7%)	77 (63.1%)	22 (18.0%)	10 (8.2%)
SP*	18 (24.7%)	30 (41.1%)	10 (13.7%)	15 (20.5%)	17 (23.3%)	23 (31.5%)	17 (23.3%)	16 (21.9%)

answering the question "Of all true ET storms, what fraction does the CPS classifier label as such?" The harmonic mean of precision and recall is called the F1 score and quantifies the overall performance of the CPS classifiers in a single number:

$$\mathrm{F1} = 2\frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

The F1 score, precision, and recall all range from 0 to 1, with higher scores signaling better performances.

The Matthews correlation coefficient (MCC) introduced by Matthews (1975) additionally takes into account the number of true negatives (TN). It is defined as

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}.$$

The MCC can take on a value between −1 and 1, where 1 represents a perfect classification, 0 is equivalent to a random classification, and −1 indicates total disagreement between classification and observation.

d. *Significance test for differences in F1 scores and MCCs*

We use a subsampling method to assess the significance of the differences in the performance metrics (F1 scores and MCCs) achieved by the classifications obtained from JRA-55 and ERA-Interim. The method is based on $n = 1000$ draws of randomly (without replacement) sampled subsets of 5 years. In each draw, the performance metrics of the two classifiers are calculated on the storms that occurred in the sampled 5 years, and the classifier that achieves the higher score is said to have won the draw. Based on the $k_{\mathrm{JRA\text{-}55}}$ times the JRA-55 classifier wins a draw and the $k_{\mathrm{ERAInt}} = n - k_{\mathrm{JRA\text{-}55}}$ draws the ERA-Interim classifier wins, we let $k = \max(k_{\mathrm{JRA\text{-}55}}, k_{\mathrm{ERAInt}})$ denote the number of draws won by the better performing classifier, and we define "success" to be the event that the better classifier wins a draw. Individual draws are treated as Bernoulli trials, that is, as independent random experiments with two possible outcomes ("success" and "failure"), in which the probability of success is the same every time the experiment is conducted.

The null hypothesis is that the JRA-55 and ERA-Interim classifiers are equally likely to win a draw (i.e., that the probability of success p_s equals 0.5). The number k of successes in n Bernoulli trials with probability p_s of success is a binomial(n, p_s) random variable. Thus, the probability of obtaining at least k successes is

$$P(X \geq k | p_s = 0.5) = \sum_{i=k}^{i=n} \binom{n}{i} p_s^i (1 - p_s)^{n-i} = 0.5^n \sum_{i=k}^{i=n} \binom{n}{i}.$$

If this probability is smaller than a significance level of $\sigma = 0.05$, we reject the null hypothesis and conclude that the difference in the performance scores of the JRA-55 and ERA-Interim classifiers is statistically significant.

There is no set rule for determining the appropriate subset size S (Politis et al. 1999). To account for this, the subsampling was repeated with subsets of 7 and 10 years.

3. Results

a. *Spatial distribution of misclassifications*

In our evaluation of the JRA-55 and ERA-Interim classifiers against the best track labels, we distinguish between misclassification of positive samples and negative samples. Misclassified positive samples are false negatives (i.e., actual ET storms that are not identified in the CPS), and misclassified negative samples are false positives (i.e., storms that are classified as ET storms in the CPS but not in the best track data). Similarly, correctly classified storms are either true positives or true negatives. Table 2 gives the complete breakdown for

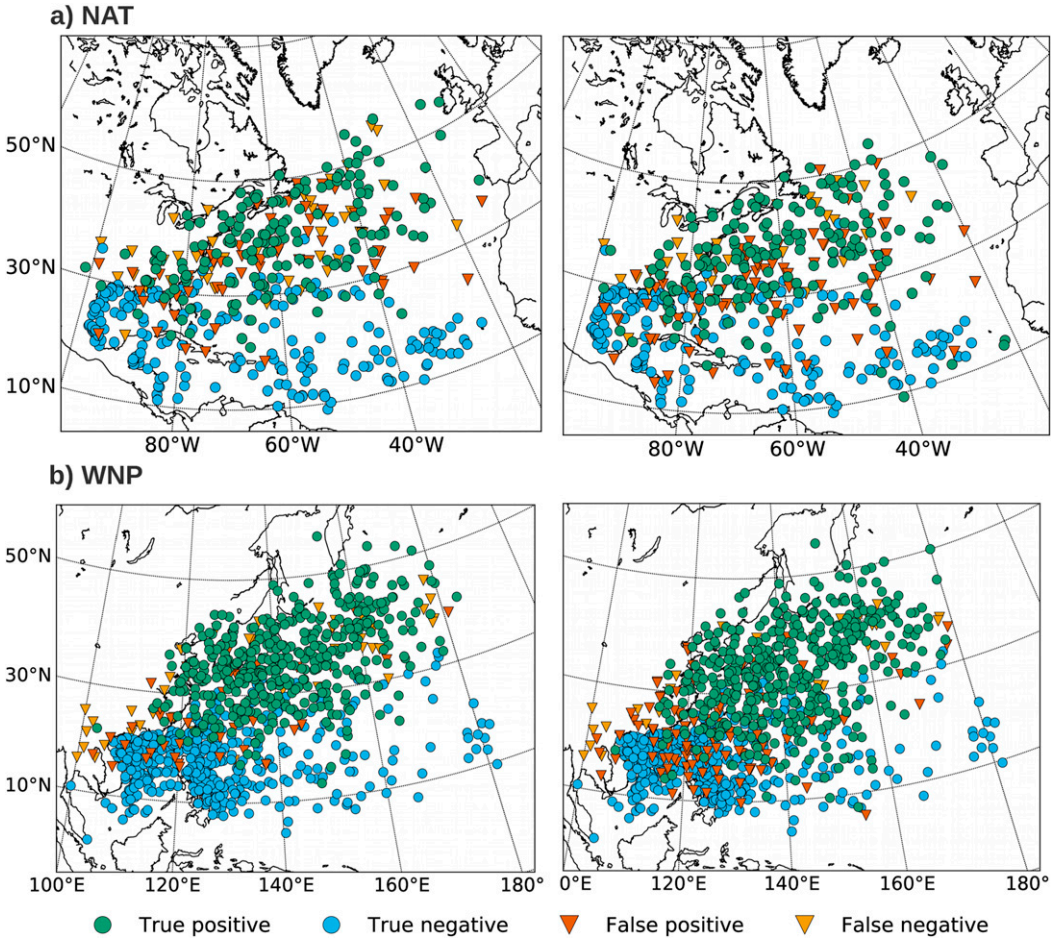

FIG. 1. Comparison of CPS-based ET detection with the best track labels in the NAT and the WNP, using the (left) JRA-55 and (right) ERA-Interim classifications, for the time period 1979–2017. Each symbol represents one storm: green dots mark the position of ET completion for true positive storms (i.e., storms that were classified as ET storms in the CPS-based detection as well as in the best track labels), red triangles denote locations where false positive storms (i.e., storms that were classified as ET storms by the CPS but not by the best track labels) completed ET, and orange triangles show the ET positions of false negative storms (i.e., storms that were classified as ET storms by the best track labels but not by the CPS); here, the ET position is defined as the location where the storm is for the first time considered extratropical in the best track data. Finally, the blue dots mark the locations where the true negative storms (which did not undergo ET in either of the two classification methods) acquire their lifetime maximum intensity.

each basin and reveals that in four of seven basins, false negatives are the dominant source of error in JRA-55, whereas ERA-Interim has more false positives than false negatives in all basins. The classification difference is greatest in the ENP, where ERA-Interim has 130 false positives, compared to 38 for JRA-55 (this finding will be analyzed further in section 3b). It is likely that the wind profile retrievals used in the JRA-55 data assimilation mentioned in section 2a (Hatsushika et al. 2006; Ebita et al. 2011) enhance the tropical characteristics of the cyclones in the reanalysis, reducing the number of false positives while increasing the number of false negatives.

Table 2 demonstrates that a meaningful comparison of the CPS classification with the best track classification has to be based on a storm-by-storm evaluation, not on ET fractions: Part I showed that in the WNP, the difference between ERA-Interim's ET fraction and the ET fraction in the best tracks is smaller for the JTWC data than for the JMA data. However, the percentage of correctly classified cyclones is greater for the JMA data (90.7% in JRA-55 and 85.8% in ERA-Interim) than for the JTWC data (79.8% and 72.8%).

Figure 1 presents the spatial distribution of the storm-by-storm evaluation for the NAT and the WNP, showing the prevailing correct classifications (true positives and

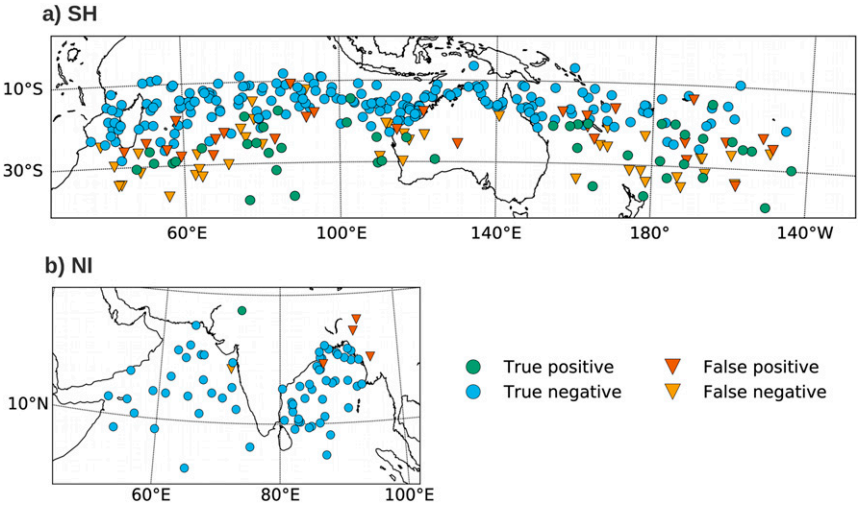

FIG. 2. Comparison of the CPS-based ET detection (using the JRA-55 classification) with the best track labels in the (a) SH and (b) NI, for the time period 2004–17. The meaning of the symbols and colors is the same as in Fig. 1.

true negatives, marked by green and blue dots) compared to the misclassified storms (false positives and false negatives, marked by red and orange triangles). The majority of false positives are located north of 20°N, but they can occur as far south as 6°N (ERA-Interim, WNP). We also note the absence of any obvious systematic differences in the spatial distribution of the wrongly classified storms between the two reanalyses.

For the SH, the distribution as well as the number of wrongly classified storms are similar in the results for JRA-55 (Fig. 2a) and ERA-Interim (not shown). There is a zonal band of true negatives with false positives at its southern edge, which implies that the CPS classifiers tend to declare ET more readily and farther north than the JTWC. At the same time, though, the CPS classification also fails to identify ET events that happen considerably farther south, as indicated by the false negatives poleward of 30°S.

ET in the NI (Fig. 2b) is more difficult to assess due to the blocking effect of the continental landmass, which prevents storms from moving far enough north to undergo ET. From 2004 to 2017, the JTWC only labeled two storms as extratropical. As a result, the evaluation of ET detection in the NI proved most sensitive to changes in the threshold values of the CPS parameters; for example, the JRA-55 classifier misclassifies only a single storm when increasing the asymmetry threshold of the *B* parameter from 11 to 14.

b. A closer look at the ENP

The discrepancy between the ET classifications of JRA-55 and ERA-Interim in the ENP (Table 2) motivates a closer inspection of that basin. Figure 3 confirms that the ET detection in JRA-55 matches the observations better, showing fewer false positives west of Mexico than does ERA-Interim. Hence,

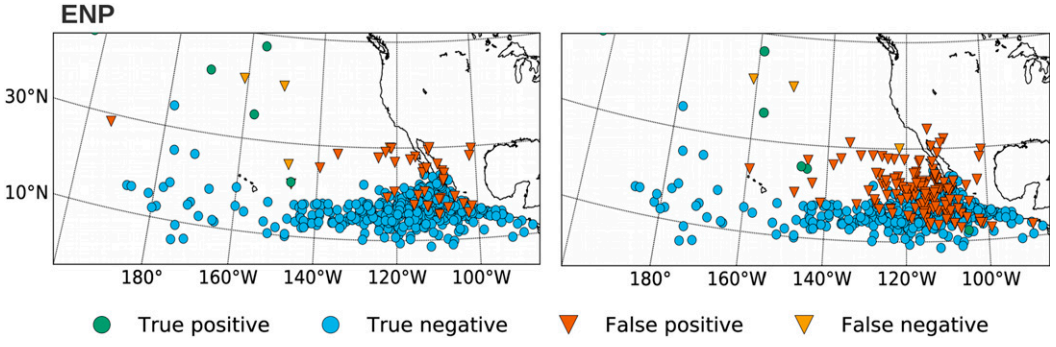

FIG. 3. Comparison of CPS-based ET detection with the best track labels in the ENP, using the (left) JRA-55 and (right) ERA-Interim classifications, for the time period 1988–2017. The meaning of the symbols and colors is the same as in Fig. 1.

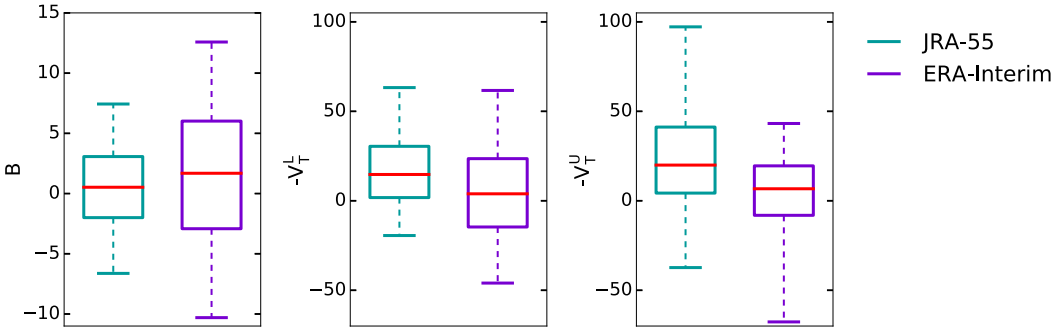

FIG. 4. Box-and-whisker plots of all CPS parameters in the ENP, calculated in JRA-55 and ERA-Interim for all 6-hourly TC positions recorded in the best track data. The box extends from the lower to the upper quartile, with a red line at the median, and the whiskers extend from the 5th to the 95th percentile.

ERA-Interim's overestimation of the ET fraction in the ENP is the result of wrongly classified ET events occurring over the ocean, in the latitude range from about 10° to 30°N. This proneness to false positives is also manifest in boxplots of all 6-hourly CPS parameters in the ENP (Fig. 4)—compared to their counterparts in JRA-55, the distributions of all three parameters in ERA-Interim have larger fractions of their values in the extratropical range (i.e., $B > 11$, $-V_T^L < 0$, $-V_T^U < 0$).

Of all 96 storms that are false positive in the ERA-Interim classification but true negative in the JRA-55 classification, 62 (65%) do not begin ET based on the CPS in JRA-55; that is, they neither exceed the asymmetry threshold $B = 11$ nor exhibit a cold core ($-V_T^L < 0$ and $-V_T^U < 0$) at any point in their lifetimes. In the remaining cases, the JRA-55 classifier diagnoses the onset of ET, but the condition for the completion of ET is not satisfied.

Composite fields of geopotential height (Fig. 5) show the representation of these 96 storms in JRA-55 and ERA-Interim. The composites are the averages of fields centered on the best track storm location, which were extracted in a 20° latitude × 20° longitude box at the time when the ERA-Interim classifier declared ET completion. Both reanalyses feature a cyclone located in the center. Thus, positional differences between the locations of the storm centers in the best tracks and those in ERA-Interim are not the primary reason for ERA-Interim's higher false alarm rate. At the 900- and 600-hPa levels, the composites of JRA-55 show a more radially symmetric and stronger cyclone than those of ERA-Interim. This is consistent with the lower values of the B parameter reached in JRA-55, which leads to fewer storms being diagnosed to have undergone ET.

Weak or dissipating stages at the end of a TC's lifetime may produce CPS signatures similar to those of ET storms, which raises the question if there is a specific type of cyclone in the best track data that tends to be misdiagnosed as ET in the ERA-Interim classification. At the time when ET is completed according to the ERA-Interim classifier, about 45% of the cyclones are labeled "tropical storms" (TCs with an intensity of 34–63 kt; 1 kt $\approx$ 0.51 m s^{-1}) in the NHC best track data. "Tropical depressions" (TCs of intensity $<$ 34 kt) and "lows" (lows of any intensity that are neither tropical, subtropical, nor extratropical cyclones) each account for about 20% of the cases (not shown). Thus, the false alarms in ERA-Interim cannot be attributed to a single type of storm. Instead, they are the result of storms that exhibit a persistent cold-core structure in ERA-Interim throughout much of their lifetimes: On average, a cold core is present at 53% of all time steps along the tracks of the ET storms, while the asymmetry parameter is only exceeded at 15% of the time steps. The median CPS trajectory of the false positives (Fig. S1 in the online supplemental material) only makes a brief excursion into the asymmetric range of the B parameter, but is located in the cold-core region from an early point on. Evidence for a bias in ERA-Interim toward cold-core structures in the representation of TCs was also found by Wood and Ritchie (2014) in their study of ET in the ENP.

The chance of a fluctuation into the $B > 11$ parameter range may be increased because the TCs in the ENP are the smallest of all basins (Knaff et al. 2014); they are about a third smaller than TCs in the NAT or the WNP. For small TCs, the (fixed) radius of 500 km used to calculate the CPS parameters may include less symmetric regions at the outer edge of the storm.

As mentioned in section 2a, JRA-55 uses historical data to produce artificial dropsonde observations in the vicinity of TCs, which are then processed like regular observations (Hatsushika et al. 2006). This is a key difference between JRA-55 and ERA-Interim, which does not apply a special TC treatment in its data assimilation process, and may help to explain the greater strength

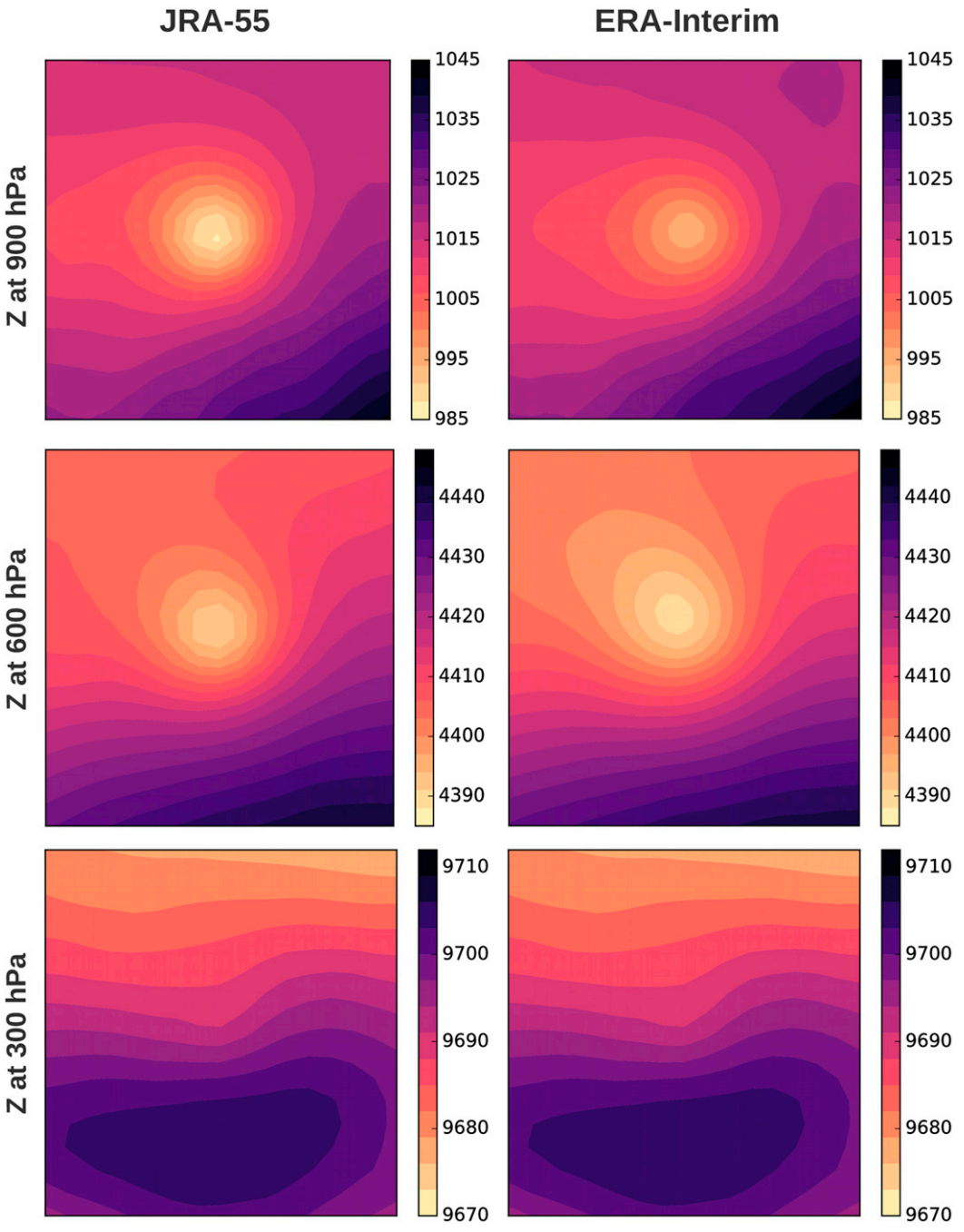

FIG. 5. Composite geopotential height fields (m) of all 96 TCs that are incorrectly labeled as ET storms (i.e., that are false positive) by the ERA-Interim classifier, but that are correctly labeled as non-ET storms (i.e., that are true negative) by the JRA-55 classifier. The composites are calculated from storm-centered geopotential height fields extracted at the time when ERA-Interim declares ET completion, in a 20° latitude × 20° longitude box at three pressure levels (top) 900, (middle) 600, and (bottom) 300 hPa, in (left) JRA-55 and (right) ERA-Interim.

and higher symmetry of the vortices in the JRA-55 composites. Still, it does not explain why the resulting difference in classification skill is greater in the ENP than in the other basins. However, according to the best track classification, there are only nine ENP ET storms between 1988 and 2017. This small sample makes it difficult to analyze whether and how ET may differ in the ENP compared to other basins; thus, our analysis is limited to studying the character of false positives in the reanalysis datasets.

c. *ET time*

To analyze the timing of ET, probability density functions (PDFs) of the differences between the best track ET times (as defined by the operational warning centers) and the times of ET completion in the CPS were calculated using a Gaussian kernel density estimation (Fig. 6). These PDFs are based on the set of all ET events that were identified both in the CPS and in the best track archives (i.e., on the set of all true positives). The distributions in the NAT are broader than those in the WNP and the SH, indicating a higher variance in the declared ET times between the CPS and the NHC than between the CPS and either the JMA or the JTWC. In the NAT, ERA-Interim on average declares ET completion 32 h before the NHC assigns the first "extratropical" label. This is consistent with Evans and Hart (2003), who examined the ET time of 38 cyclones in the NAT and found that the time of ET completion diagnosed by the CPS in the ECMWF's 15-yr Reanalysis (ERA-15; Gibson et al. 1997) occurs on average about 28 h earlier than in the NHC best tracks. In contrast, the mean difference between the ET time in JRA-55 and that of the NHC classification is only 10 h. The JRA-55 ET completion times also agree better with the JMA labels in the WNP than the ERA-Interim completion times do, while the PDFs of the ET time differences to the JTWC labels in the SH are almost identical for the two reanalysis datasets. Based on a *t* test for the sample mean and an *F* test for the sample variance, the interreanalysis differences in the transition time periods are significant in the NAT and the WNP, but not in the SH.

d. *Precision, recall, F1 scores, and Matthews correlation coefficients*

Figure 7a shows the F1 scores of the JRA-55 and ERA-Interim classifiers. The CPS classification agrees best with the observations in the WNP and the NAT, with F1 scores of 0.90 and 0.77, respectively, for JRA-55, and 0.86 and 0.76, respectively, for ERA-Interim. As already indicated in Table 2, the classification in the WNP based on the JTWC best tracks receives a lower F1 score than that based on the JMA best tracks. In Part I, it was shown that the JMA best tracks on average extend farther northeast than the JTWC best tracks. Thus, the operational treatment of ET in the JMA and the JTWC as well as the tracks themselves may contribute to the differences in the F1 scores.

Compared to the F1 scores in the NAT and the WNP, the scores in the ENP, the NI and the SH basins are lower for both reanalysis classifiers, but consistently higher for the JRA-55 classifier than for the ERA-Interim classifier.

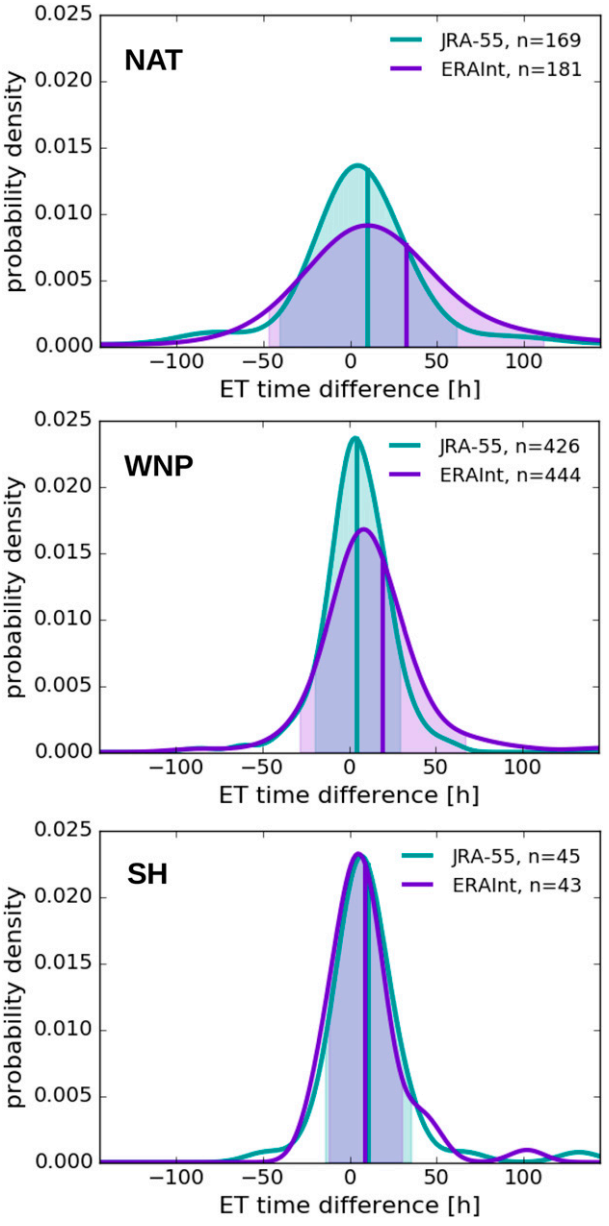

FIG. 6. Probability density functions of the differences between the best track ET times and the ET completion in the CPS, based on the ET events from 1979 to 2017 in the NAT and the WNP, and from 2004 to 2017 in the SH. The vertical lines indicate the means of the distributions, and the shaded regions represent values within one standard deviation about the mean. Positive (negative) time differences indicate that the best track ET time is later than (earlier than) the ET completion in the CPS.

The decomposition of the F1 scores into precision and recall (Fig. 7b) shows that the F1 scores in the NAT, the WNP, and the SH basins are composed of almost equal values of precision and recall—in other words, the CPS ET classification is equally good at avoiding false positives as at avoiding false negatives. The F1 performance

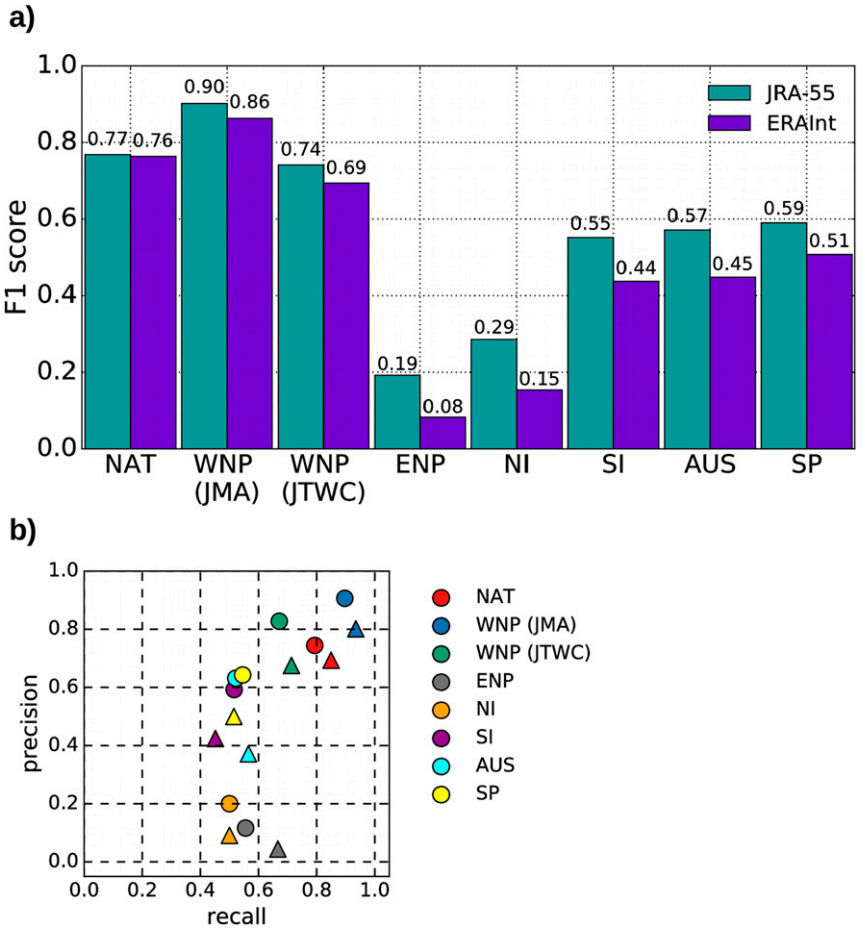

FIG. 7. (a) F1 scores assessing the performance of the CPS classifiers. The time period used to calculate the F1 scores is 1979–2017 for the NAT and the WNP (JMA), 1988–2017 for the ENP, and 2004–17 for the WNP (JTWC), NI, SI, AUS, and the SP. The results for the WNP are shown for the best track archives of JMA as well as JTWC. (b) Precision and recall associated with the F1 scores in (a); scores are marked as circles for JRA-55 and as triangles for ERA-Interim.

in the NI and the ENP is more asymmetric, with a higher recall than precision. This is likely a result of the scarcity of ET events in these two basins, which makes it difficult to identify the rare true ET storms while avoiding false alarms.

As with the F1 scores, the MCCs (Fig. 8) are highest in the WNP and the NAT, and the MCCs of JRA-55 exceed those of ERA-Interim in all basins. The MCCs are greater than zero in all basins, indicating a better than random correlation with the best track classification (recall that the MCC ranges from -1 to 1), although only by a small margin for the ERA-Interim classifications in the SP and the ENP. In the SP, the MCC is considerably lower than in the other two SH basins, despite similar F1 scores. With that exception, the general pattern of the evaluation is robust with respect to the two performance metrics.

However, it is notable that if we used the proportion of correct classifications, also termed accuracy, as a measure of classification skill, the NI would achieve the highest scores (0.93 in JRA-55 and 0.85 in ERA-Interim), and the average score of the two reanalyses in the ENP would be higher than that in the NAT (0.82 compared to 0.78). These results make it clear that accuracy is a misleading performance metric when the two classes (ET storms and non-ET storms) are of very different sizes. To further illustrate this point, consider a hypothetical basin where only 0.1% of all storms undergo ET. A "dummy" classifier that, without performing any analysis, assigns each storm to the majority class (here: non-ET storms) would achieve an accuracy of 0.99 despite not having any classification skill.

Table 3 presents the results of the significance test described in section 2d, for the F1 score and the MCC.

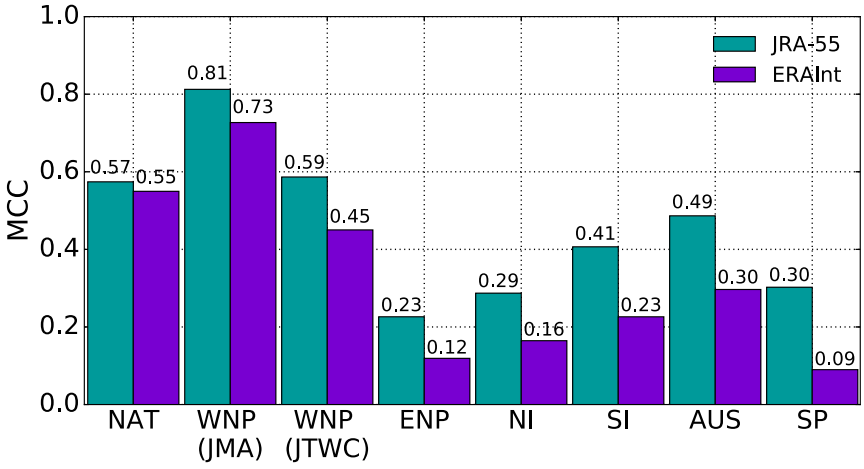

FIG. 8. MCCs assessing the performance of the CPS-based ET classification. The time period used to calculate the F1 scores is 1979–2017 for the NAT and the WNP (JMA), 1988–2017 for the ENP, and 2004–17 for the WNP (JTWC), NI, SI, AUS, and the SP. The results for the WNP are shown for the best track archives of JMA as well as JTWC.

All differences between the performance scores of the JRA-55 and the ERA-Interim classifications are significant. Repeating the test with different subset sizes (S = 7 years and S = 10 years) did not change the significance of the results. Recall that a high statistical significance does not imply that the performance difference is large, but that a (possibly small) difference in classification skill is consistently present on randomly sampled subsets of storms.

e. *Time series of classification skill*

In the NAT and the WNP, the high quality of the best track datasets and the frequency of ET motivate a look at how the agreement between the CPS classification and the best track classification has evolved over time. A possible reason for changes in that agreement is modifications in the operational procedures at TC warning centers; for example, since 2005, the NHC has routinely used model-derived CPS parameters in operational forecast discussions.

Figure 9 shows time series of F1 scores and MCCs in these two basins, and Table 4 summarizes some statistics of these time series. In both basins, the slopes of the linear regression lines are positive, but only those in the WNP are statistically significant for both reanalysis classifiers. In the WNP, the MCCs are almost as high as the F1 scores, indicating that the CPS classifiers perform well both in classifying positive samples and in correctly recognizing negative samples.

The correlations between the time series of JRA-55 and ERA-Interim are high and statistically significant (Table 4). Thus, the two classifiers do not only have similar F1 scores and MCCs on the set of all storms (Figs. 7 and 9), but also on individual 3-yearly subsets of storms.

TABLE 3. Statistical significance of the differences in F1 scores and MCC, evaluated by repeatedly (n = 1000) choosing a random sample of 5 yr and calculating the F1 score and MCC of the JRA-55 and ERA-Interim classifiers on the storms that occurred in the sampled 5 yr. Here, $k_{\text{JRA-55}}$ (k_{ERAInt}) is the number of times the JRA-55 (ERA-Interim) classifier achieves a higher performance score, and $P(X \geq k|p_s = 0.5)$ is the probability of obtaining at least $k = \max(k_{\text{JRA-55}}, k_{\text{ERAInt}})$ successes in n Bernoulli trials, assuming the null hypothesis is true, namely that the probability of success p_s equals 0.5. Statistically significant values are in bold.

	F1 score			MCC		
	$k_{\text{JRA-55}}$	k_{ERAInt}	$P(X \geq k\|p_s = 0.5)$	$k_{\text{JRA-55}}$	k_{ERAInt}	$P(X \geq k\|p_s = 0.5)$
NAT	540	460	**0.006**	576	424	**<0.001**
WNP (JMA)	795	205	**<0.001**	815	185	**<0.001**
WNP (JTWC)	751	249	**<0.001**	929	71	**<0.001**
ENP	627	373	**<0.001**	589	411	**<0.001**
NI	572	428	**<0.001**	550	450	**<0.001**
SI	802	198	**<0.001**	866	134	**<0.001**
AUS	686	314	**<0.001**	737	263	**<0.001**
SP	765	235	**<0.001**	881	119	**<0.001**

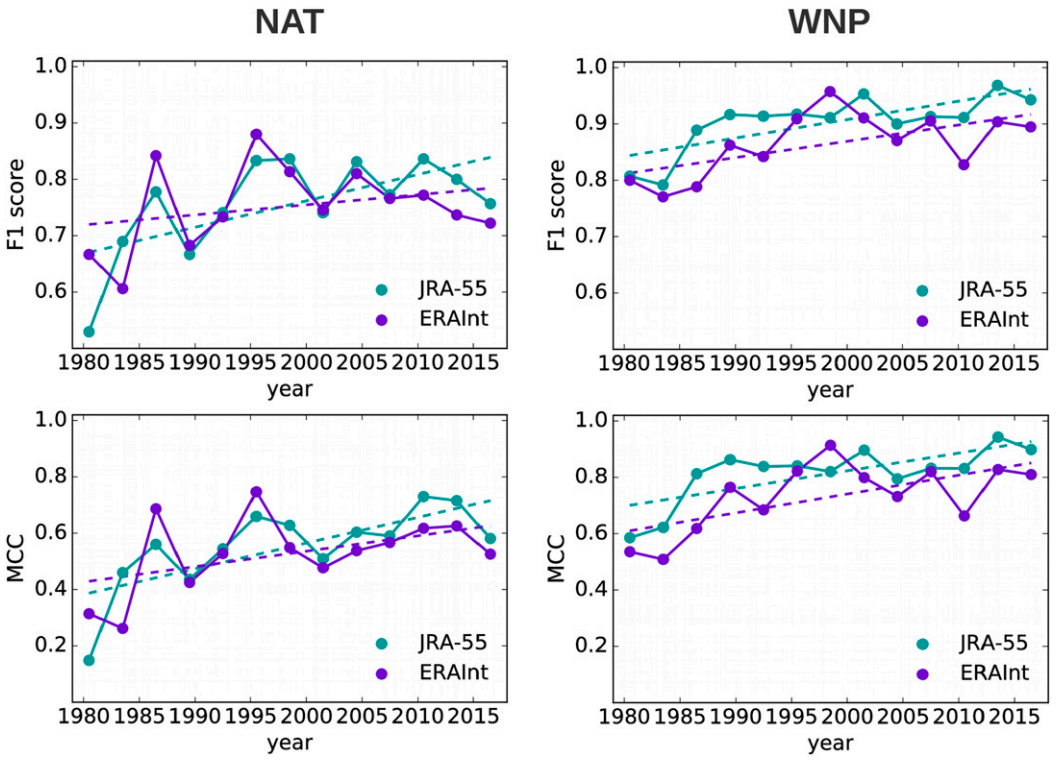

FIG. 9. Time series of (top) F1 scores and (bottom) MCCs in the (left) NAT and (right) WNP, for JRA-55 and ERA-Interim. Each data point represents the classification performance calculated on a 3-yr period. The dashed lines are the linear regression best fits to the time series.

The introduction of the CPS as an operational tool at the NHC does not lead to a jump in the F1 scores and MCCs in the NAT, which may reflect the fact that Evans and Hart (2003) originally built the CPS diagnostics of ET on the NHC classifications.

However, the performance of the CPS classifiers has an upward trend in both basins. Two conceivable reasons are that the increasing number of observations assimilated into JRA-55 and ERA-Interim has made the representation of TCs more accurate over time, or that there have been changes in the operational practices and attention dedicated to the ET designation at the warning centers.

4. Discussion

The fact that the JRA-55 classifier agrees better with the observed ETs recorded in the best track datasets than the ERA-Interim classifier is consistent with the study by Murakami (2014), in which JRA-55 comes out ahead in an evaluation of the representation of TCs in six reanalyses. As mentioned in section 3b, the high rate of false positives we found in the ENP is consistent with Wood and Ritchie (2014), who noted in their study of ET in the ENP that ERA-Interim has a bias toward cold-core values in the 900–600-hPa layer compared

TABLE 4. Statistics of the time series of F1 scores and MCCs: sample mean and standard deviation (JRA-55, ERA-Interim), *p* values of the slope of the linear regression lines (JRA-55, ERA-Interim), Pearson correlation coefficient *R* between the JRA-55 and the ERA-Interim time series, and *p* value of that correlation coefficient. Statistically significant values are in bold.

Basin	Mean	Std dev	*p* value of slope	*R*	*p* value of *R*
			F1 score (1979–2017)		
NAT	0.75, 0.75	0.08, 0.07	**0.022**, 0.356	0.74	**0.004**
WNP	0.90, 0.86	0.05, 0.05	**0.003, 0.029**	0.73	**0.004**
			MCC (1979–2017)		
NAT	0.55, 0.53	0.14, 0.13	**0.006**, 0.105	0.76	**0.003**
WNP	0.81, 0.73	0.10, 0.12	**0.005, 0.018**	0.78	**0.002**

with both JRA-55 and the final operational global analysis (FNL) data from the Global Forecast System. However, deficiencies in the representation of TCs are by no means limited to ERA-Interim but are a well-known issue of reanalyses (including JRA-55; e.g., Schenkel and Hart 2012; Murakami 2014; Hodges et al. 2017) and climate model output (e.g., Randall et al. 2007; Camargo and Wing 2016) in general.

The most prominent problem associated with TCs in reanalyses is the substantial underestimation of the storm intensities. However, the CPS parameters are based on relative comparisons (layer thickness left and right of the storm for B, and vertical profiles of ΔZ for thermal wind parameters) and do not depend in any direct way on storm intensity. This offers the hope that the threshold parameters used to detect ET may not have to be adjusted to the increasing resolution and stronger intensities of cyclones in future reanalyses.

Of course, the performance evaluation of the CPS classifiers presented in this study hinges on the quality of the best track data, in particular on the labels indicating the tropical or extratropical nature of each cyclone. Even though the best tracks are the most accurate and comprehensive archives of historical TC data available, they are still prone to considerable uncertainty, especially the components that are derived from a forecaster's subjective judgment (e.g., Landsea and Franklin 2013). In addition, there may be inhomogeneities in the data quality due to agencies putting less effort into the classification of transitioning storms or stopping the tracking earlier in basins where ET storms do not pose a threat to land.

Given these limitations, it is clear that assessing the CPS classifiers against the best track labels cannot in all cases be interpreted as a comparison with the ''true'' classification. Put simply, when the labels are wrong, high performance scores do not indicate good classification skill, and vice versa. However, the time series of best track ET fractions shown in Part I did not reveal any statistically significant trends at the 0.05 significance level that were robust between the two reanalyses, and neither did time series of the magnitude of the difference between the CPS-based fractions and the best track labels (not shown). Trends were also absent in time series of the annual mean latitude of storm track end points (not shown). Taken together, these results indicate that operational procedures in the tracking and characterization of cyclones have been fairly consistent in the time period 1979–2017, which provides some reassuring evidence that the best track labels can to a reasonable approximation be assumed to represent the ''ET truth.'' In basins where that assumption is less valid, it still provides a means to examine differences in the ET classifications of the two reanalysis datasets, but there is limited value in interpreting the observed differences in terms of classification skill.

5. Summary and concluding remarks

In this study, we analyze the statistical performance of a global classification of tropical cyclones (TCs) that undergo extratropical transition (ET). The classification is used in Part I of this study for an examination of the geographical, seasonal, and temporal characteristics of ET in seven ocean basins. Here, we have investigated how well the ET storms defined in the CPS agree with those defined in the best track records, and how that agreement depends on whether the CPS is calculated from JRA-55 or from ERA-Interim data. At the core of this evaluation is the binary classification into ET storms (TCs that undergo ET at some point in their lifetimes) and non-ET storms (TCs that do not undergo ET) obtained from the CPS analysis using JRA-55 data (the JRA-55 classifier) and ERA-Interim data (the ERA-Interim classifier).

Our results can be summarized as follows:

- According to the F1 score and the Matthews correlation coefficient (MCC), two performance metrics that balance classification sensitivity and specificity, the CPS classification agrees best with the best track classification in the western North Pacific (MCC $>$ 0.7) and the North Atlantic (MCC $>$ 0.5).
- The correlations between the CPS classification and the best track classification are considerably weaker in the other basins. In the South Pacific and the eastern North Pacific, the MCC of the ERA-Interim classification is only slightly higher than that of a random classification.
- The JRA-55 classifier achieves higher performance scores than does the ERA-Interim classifier. The differences are statistically significant in all basins.
- The lower performance of ERA-Interim is mainly due to a higher false alarm rate, which is especially pronounced in the eastern North Pacific. The false positives in the eastern North Pacific are the result of a bias toward cold-core structures in the representation of TCs in ERA-Interim.
- On average, ET completion in the North Atlantic and the western North Pacific occurs earlier in ERA-Interim than in JRA-55, but almost simultaneously in the Southern Hemisphere.
- In the North Atlantic and the western North Pacific, the agreement between the CPS classification and the

best track classification (as measured by the MCC and the F1 score) has increased from 1979 to 2017, but only the trend in the western North Pacific is statistically significant for both the JRA-55 and the ERA-Interim classifier.

Our results show that the CPS computed from reanalysis data can be used to provide a globally consistent dataset that, while by no means in perfect agreement with the diagnoses of ET produced by forecasters, are nonetheless close enough—especially in the basins where ET is most common—to be usable for the purposes of some kinds of climatological studies, as long as the limitations are understood. At the same time, improvement is clearly possible. While we are not certain, it seems plausible that we obtain higher performance scores with JRA-55 than ERA-Interim here due to JRA-55's special procedures to initialize TCs; this suggests that further improvement in the representation of TCs in reanalysis datasets—whether through higher resolution, improved physics, data assimilation, or other TC-specific initialization procedures—might yield further improvements. The CPS itself is also an imperfect measure, and exploration of other objective metrics of ET is warranted, as also suggested by Evans et al. (2017). Since diagnosing ET is in some sense a problem in pattern recognition, machine learning or other advanced statistical approaches might be beneficial; we are exploring a small subset of such methodologies and will report on this in due course.

It is also possible that even the forecaster-generated best track datasets we take here as ground truth are themselves imperfect indicators of ET, and perhaps even that in some cases there might be fundamental scientific uncertainty (i.e., not simply a consequence of inadequate data) as to whether a storm should be considered tropical or extratropical at a given moment, or even whether a binary classification is adequate to describe what might be better thought of as a gradual transition process (Beven 2008, 2012). In cases where different metrics of ET (including CPS from different reanalyses and/or best track datasets) yield strongly different results, in-depth case studies to examine physical mechanisms could be valuable, and could add to our fundamental understanding of the ET process.

Acknowledgments. The funding for this research was provided by NASA Cooperative Agreement NNX15AJ05A, and by NSF under Grant ATM-1322532. The authors also thank the following organizations for making the data used in this study available: ECMWF (ERA-Interim reanalysis data), JMA (JRA-55 reanalysis data and western North Pacific best track data), NHC (North Atlantic and eastern North Pacific best track data), and JTWC (western North Pacific, North Indian Ocean, and Southern Hemisphere best track data).

REFERENCES

Beven, J. L., 2008: Verification of National Hurricane Center forecasts of extratropical transition. Preprints, *28th Conf. on Hurricanes and Tropical Meteorology*, Orlando, FL, Amer. Meteor. Soc., 10C.2, https://ams.confex.com/ams/28Hurricanes/techprogram/paper_138321.htm.

——, 2012: Cyclone type analysis and forecasting: A need to re-visit the issue. Preprints, *30th Conf. on Hurricanes and Tropical Meteorology*, Ponte Vedra Beach, FL, Amer. Meteor. Soc., 2A.3, https://ams.confex.com/ams/30Hurricane/webprogram/Paper205647.html.

Bieli, M., S. J. Camargo, A. H. Sobel, J. L. Evans, and T. Hall, 2019: A global climatology of extratropical transition. Part I: Characteristics across basins. *J. Climate*, **32**, 3557–3582, https://doi.org/10.1175/JCLI-D-17-0518.1.

Camargo, S. J., and A. A. Wing, 2016: Tropical cyclones in climate models. *Wiley Interdiscip. Rev.: Climate Change*, **7**, 211–237, https://doi.org/10.1002/wcc.373.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Ebita, A., and Coauthors, 2011: The Japanese 55-year Reanalysis "JRA-55": An interim report. *SOLA*, **7**, 149–152, https://doi.org/10.2151/sola.2011-038.

Evans, C., and Coauthors, 2017: The extratropical transition of tropical cyclones. Part I: Cyclone evolution and direct impacts. *Mon. Wea. Rev.*, **145**, 4317–4344, https://doi.org/10.1175/MWR-D-17-0027.1.

Evans, J. L., and R. E. Hart, 2003: Objective indicators of the life cycle evolution of extratropical transition for Atlantic tropical cyclones. *Mon. Wea. Rev.*, **131**, 909–925, https://doi.org/10.1175/1520-0493(2003)131<0909:OIOTLC>2.0.CO;2.

Fiorino, M., 2002: Analysis and forecasts of tropical cyclones in the ECMWF 40-year reanalysis (ERA-40). *Extended Abstracts, 25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., 261–264, https://ams.confex.com/ams/25HURR/techprogram/paper_38743.htm.

Fogarty, C., 2010: Forecasting extratropical transition. *Proc. Seventh Int. Workshop on Tropical Cyclones*, St. Gilles Les Bains, La Réunion, France, WMO, 2.5, http://www.wmo.int/pages/prog/arep/wwrp/tmr/otherfileformats/documents/2_5.pdf.

Gibson, J., P. Kållberg, S. Uppala, A. Hernandez, A. Nomura, and E. Serrano, 1997: ERA-15 Description. Vol. 1, ECMWF Re-Analysis (ERA) Project Report Series, European Centre for Medium-Range Weather Forecasts, 72 pp., https://www.ecmwf.int/en/elibrary/9584-era-description.

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

Hatsushika, H., J. Tsutsui, M. Fiorino, and K. Onogi, 2006: Impact of wind profile retrievals on the analysis of tropical cyclones in the JRA-25 reanalysis. *J. Meteor. Soc. Japan*, **84**, 891–905, https://doi.org/10.2151/jmsj.84.891.

Hodges, K., A. Cobb, and P. L. Vidale, 2017: How well are tropical cyclones represented in reanalysis datasets? *J. Climate*, **30**, 5243–5264, https://doi.org/10.1175/JCLI-D-16-0557.1.

Jones, S. C., and Coauthors, 2003: The extratropical transition of tropical cyclones: Forecast challenges, current understanding, and future directions. *Wea. Forecasting*, **18**, 1052–1092, https://doi.org/10.1175/1520-0434(2003)018<1052:TETOTC>2.0.CO;2.

Kitabatake, N., 2011: Climatology of extratropical transition of tropical cyclones in the western North Pacific defined by using cyclone phase space. *J. Meteor. Soc. Japan*, **89**, 309–325, https://doi.org/10.2151/jmsj.2011-402.

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

Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, https://doi.org/10.2151/jmsj.2015-001.

Kofron, D. E., E. A. Ritchie, and J. S. Tyo, 2010: Determination of a consistent time for the extratropical transition of tropical cyclones. Part I: Examination of existing methods for finding ''ET time''. *Mon. Wea. Rev.*, **138**, 4328–4343, https://doi.org/10.1175/2010MWR3180.1.

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.

Liu, M., G. A. Vecchi, J. A. Smith, and H. Murakami, 2017: The present-day simulation and twenty-first-century projection of the climatology of extratropical transition in the North Atlantic. *J. Climate*, **30**, 2739–2756, https://doi.org/10.1175/JCLI-D-16-0352.1.

Masson, A., 2014: The extratropical transition of Hurricane Igor and the impacts on Newfoundland. *Nat. Hazards*, **72**, 617–632, https://doi.org/10.1007/s11069-013-1027-x.

Matthews, B. W., 1975: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta, Protein Struct.*, **405**, 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.

Murakami, H., 2014: Tropical cyclones in reanalysis data sets. *Geophys. Res. Lett.*, **41**, 2133–2141, https://doi.org/10.1002/2014GL059519.

Politis, D. N., J. P. Romano, and M. Wolf, 1999: *Subsampling*. Springer-Verlag, 348 pp., https://doi.org/10.1007/978-1-4612-1554-7.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 589–662.

Schenkel, B. A., and R. E. Hart, 2012: An examination of tropical cyclone position, intensity, and intensity life cycle within atmospheric reanalysis datasets. *J. Climate*, **25**, 3453–3475, https://doi.org/10.1175/2011JCLI4208.1.

Ting, K. M., 2010: Precision and recall. *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds, Springer, 781, https://doi.org/10.1007/978-0-387-30164-8_652.

Wood, K. M., and E. A. Ritchie, 2014: A 40-year climatology of extratropical transition in the eastern North Pacific. *J. Climate*, **27**, 5999–6015, https://doi.org/10.1175/JCLI-D-13-00645.1.

Zarzycki, C. M., D. R. Thatcher, and C. Jablonowski, 2017: 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.