

Identification of Tropical Cyclone Storm Types Using Crowdsourcing

KENNETH R. KNAPP

NOAA/National Centers for Environmental Information, Asheville, North Carolina

JESSICA L. MATTHEWS

NOAA/National Centers for Environmental Information, and Cooperative Institute for Climate and Satellites–North Carolina, North Carolina State University, Asheville, North Carolina

JAMES P. KOSSIN

NOAA/National Centers for Environmental Information, Asheville, North Carolina

CHRISTOPHER C. HENNON

University of North Carolina at Asheville, Asheville, North Carolina

(Manuscript received 11 January 2016, in final form 10 June 2016)

ABSTRACT

The Cyclone Center project maintains a website that allows visitors to answer questions based on tropical cyclone satellite imagery. The goal is to provide a reanalysis of satellite-derived tropical cyclone characteristics from a homogeneous historical database composed of satellite imagery with a common spatial resolution for use in long-term, global analyses. The determination of the cyclone “type” (curved band, eye, shear, etc.) is a starting point for this process. This analysis shows how multiple classifications of a single image are combined to provide probabilities of a particular image’s type using an expectation–maximization (EM) algorithm. Analysis suggests that the project needs about 10 classifications of an image to adequately determine the storm type. The algorithm is capable of characterizing classifiers with varying levels of expertise, though the project needs about 200 classifications to quantify an individual’s precision. The EM classifications are compared with an objective algorithm, satellite fix data, and the classifications of a known classifier. The EM classifications compare well, with best agreement for eye and embedded center storm types and less agreement for shear and when convection is too weak (termed no-storm images). Both the EM algorithm and the known classifier showed similar tendencies when compared against an objective algorithm. The EM algorithm also fared well when compared to tropical cyclone fix datasets, having higher agreement with embedded centers and less agreement for eye images. The results were used to show the distribution of storm types versus wind speed during a storm’s lifetime.

1. Introduction

The best track data record is important for many applications, one of which is understanding how tropical cyclones could be changing over time. Tropical cyclone observations have changed through time, leading to changes in how best tracks were constructed, which is

summarized by Knapp et al. (2010). Best track data are constructed differently in different basins. Even agencies in the same basin construct intensity records in different ways (Knapp and Kruk 2010; Knapp et al. 2013; Ren et al. 2011). Kossin et al. (2013) summarize many of the causes of differences in best track data, which can be generalized as changes in observations [like the cessation of aircraft reconnaissance (Martin and Gray 1993) or newer satellites], improved understanding of storms [e.g., advances in wind–pressure relationships (Knapp and Zehr 2007)], enhanced tools [such as improvements to automated analysis techniques (Velden et al. 2006b)], and advances in technology [for instance, when digital satellite analysis replaced paper faxes (P. Caroff 2009, personal communication)]. For

 Denotes Open Access content.

Corresponding author address: Kenneth R. Knapp, 151 Patton Ave., Asheville, NC 28801.
E-mail: ken.knapp@noaa.gov

DOI: 10.1175/MWR-D-16-0022.1

instance, the end of aircraft reconnaissance in the western North Pacific significantly impacted typhoon analysis.

Changes in typhoon activity have been widely debated (Emanuel 2005; Webster et al. 2005; Wu et al. 2006). Prior to the end of reconnaissance, the best track datasets showed more agreement (Knapp et al. 2013). However, after reconnaissance ended, the record became almost completely dependent on satellite analysis, which differed between agency (Nakazawa and Hoshino 2009). Thus, a uniform or homogenous best track record is needed to truly understand how tropical cyclone intensity, frequency, or distribution may be changing.

A homogenous record could be derived from experts using the Dvorak analysis technique. It is generally understood that expert operational tropical analysts are highly proficient at estimating tropical cyclone intensity from IR imagery; however, it is unrealistic to assign an expert or a few experts the task of going back through 35 years of satellite imagery to create a more homogeneous record. So, in lieu of this, we use crowdsourcing to provide a homogeneous analysis. This estimate must indeed be based on quantity to provide quality, which is the fundamental concept motivating crowdsourcing.

Cyclone Center is a citizen science project designed to provide the best possible reanalysis of satellite-derived tropical cyclone characteristics from a homogeneous historical database composed of infrared satellite imagery with a common spatial resolution for use in long-term, global analyses. Hennon et al. (2015) provide more information on the purpose, background, and scope of the project. The project asks visitors to the website (<http://www.cyclonecenter.org/>) a series of questions that they can answer based on historical tropical cyclone satellite imagery; thus, it is not meant as a means for real-time operational analysis. The resulting data from this analysis (or the project in general) should not replace any best track databases; instead, it could inform a future reanalysis of tropical cyclone data.

Numerous classifiers provide information for each satellite image. The work herein attempts to derive information about an image using the (sometimes) disparate answers from the various classifiers. While the project asks numerous questions that should help gauge a storm's intensity, the following analysis investigates just one question: what type of cloud pattern (eye, curved band, etc.) is apparent in the image?

The goals of this paper are to show the following:

- 1) the accuracy and consistency of classifiers, while identifying which storm types have more precision;
- 2) that the resulting derived storm types agree with the conceptual idea of what that storm type represents; and

- 3) that the distribution of storm types in time and space is conceptually sound.

Such an analysis—determining storm type—will allow further analysis of Cyclone Center data, where more detailed information is provided for each storm type. Therefore, the goal herein is not to derive a storm intensity, but to investigate if the storm type information is consistent with (i) other data and (ii) reality such that it can be used in future analysis that investigates storm intensity. Based on the positive results, the relationships of storm types during a storm's lifetime are quantified.

2. Data

a. Cyclone Center data

The Cyclone Center (CC) website presents satellite imagery [derived from Hurricane Satellite (HURSAT) data from Knapp and Kossin (2007)] and asks a series of questions for each image that can provide information on storm intensity. The answers to these questions are recorded and provided to the CC science team. However, the identity of specific classifiers in the ensuing discussion is kept anonymous. The questions on the website are similar to the type of analysis described in the Dvorak technique (Dvorak 1984; Velden et al. 2006b). While this approach applies the Dvorak technique globally, tendencies and dependencies identified by Velden et al. (2006a) and Knaff et al. (2010) can be accounted for in postprocessing (dependencies on storm translation speed, latitude, etc.).

To determine the storm type (or pattern) from multiple classifications, we analyze each classifier's selection from a set of canonical imagery when asked to "pick the cyclone type, then choose the closest match." This request addresses the concept of the cloud patterns and derives from what Dvorak calls the pattern T number. The classifications used for this study of cyclone types are shown in Fig. 1, which shows the CC canonical images for curved band (CBD), embedded center (EMB), eye (EYE), and shear (SHR) storms. Also, two other categories are included that classifiers can select: no storm (NS) and posttropical (PT).

The CC data used here are from all classifications through 13 September 2015, which represents nearly 3 full years of project data (which started on 26 September 2012). In total, there are 483 334 image classifications from about 26 000 unique individuals. These classifications are of 92 672 storm images from 1704 tropical cyclones. This represents first steps toward the goal of completing the entire HURSAT record of storms. Figure 2 shows the distribution of classifications from project participants. Most people in the project have few classifications. The two most active classifiers have more than 20 000 classifications; conversely, there are about

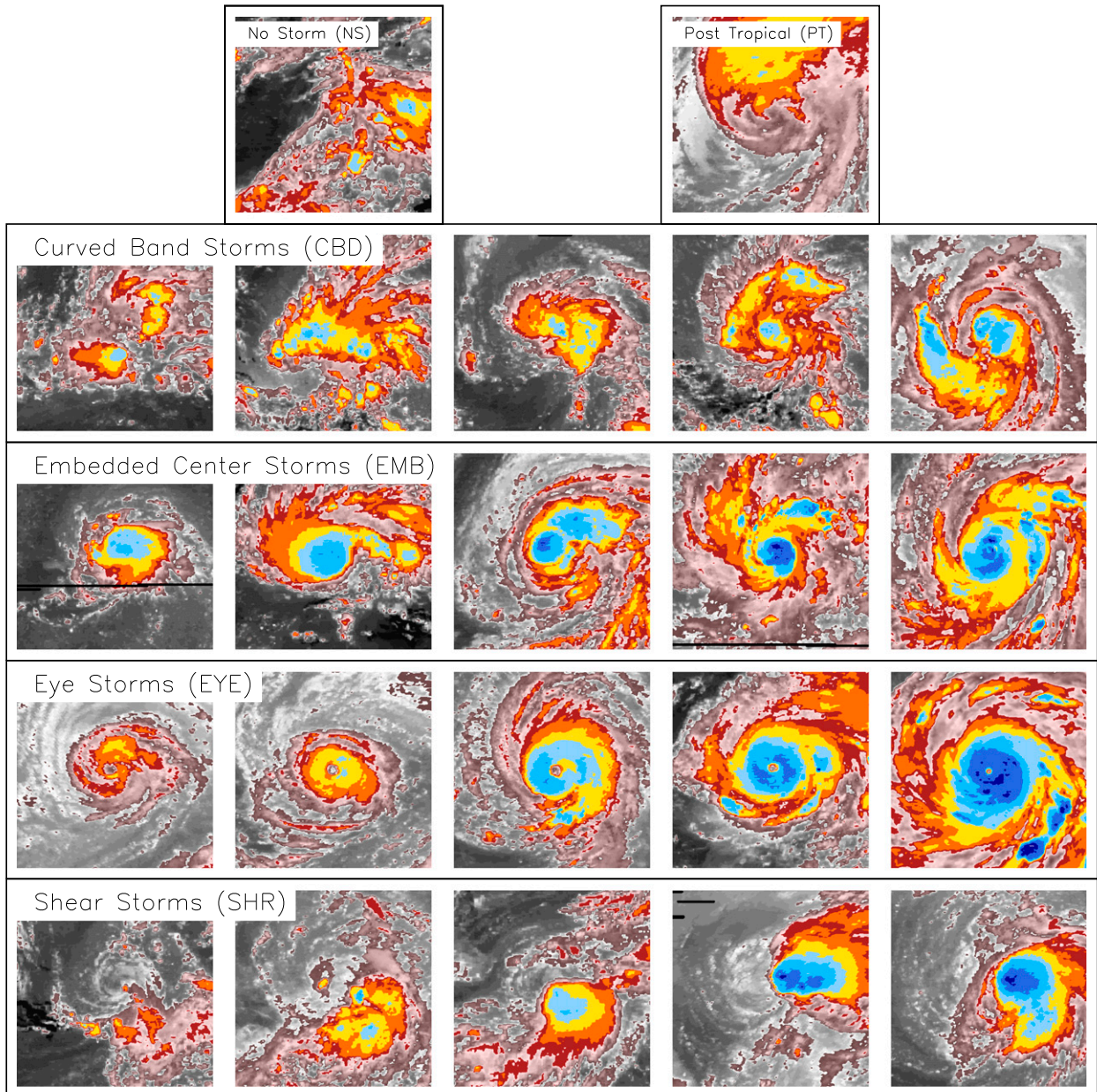


FIG. 1. Canonical images used by the Cyclone Center website (CycloneCenter.org) for classification categories used herein.

10000 classifiers with 1 classification. The challenge, then, is to extract information for each image from multiple classifications when each classifier is unknown and most of the classifiers have few classifications.

b. Comparison data

Storm types derived from citizen science classifications are compared to three sources of tropical cyclone data.

The advanced Dvorak technique (ADT) is an objective algorithm derived for operational analysis of tropical cyclone imagery (Olander and Velden 2007). It has a

long history of development (Velden et al. 2006b) but was only recently applied to the HURSAT dataset (ADT-H) by Kossin et al. (2013). The result is an objective estimation of TC intensity for the entire HURSAT period. One of the products of ADT-H is a characterization of storm type. This will be compared with the storm type derived from the algorithm described below. This should not be confused with the operational ADT output used by forecasters in their analysis; the operational ADT product uses different input data (e.g., higher-resolution infrared satellite

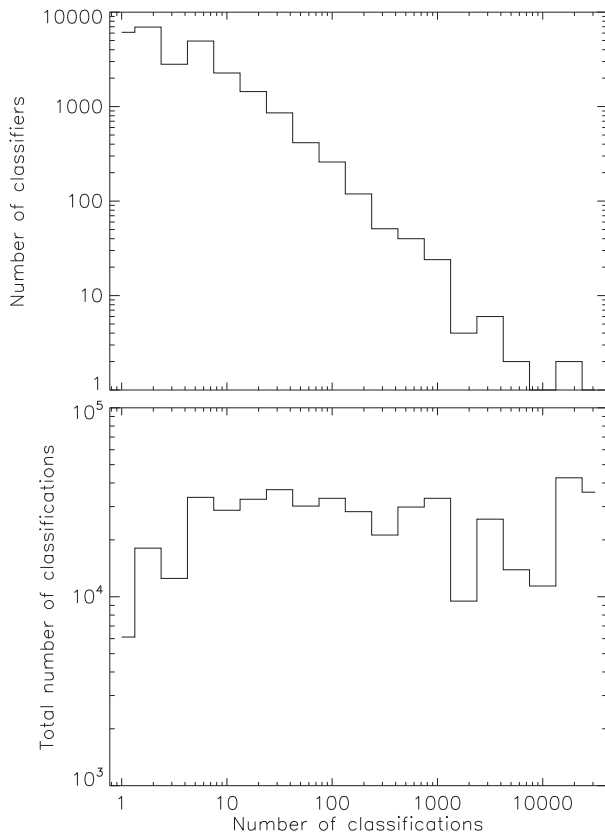


FIG. 2. Distribution of (top) the number of classifications from each classifier and (bottom) the total number of classifications made based on how many classifications a classifier made.

imagery) and continues to evolve toward higher accuracy [e.g., by incorporating satellite microwave imagery (Olander and Velden 2012)].

Professional analysts record tropical cyclone estimates of position and intensity (or tropical cyclone “fixes”) from several different data sources: aircraft reconnaissance, satellite analysis, radar observations, etc. The data are recorded by the Automated Tropical Cyclone Forecast system (Sampson and Schrader 2000) as comma delimited files. The fix data used herein come from the Tropical Analysis Forecast Branch within the National Hurricane Center, the Satellite Analysis Branch within the National Environmental Satellite, Data, and Information Service (NESDIS), and the Joint Typhoon Warning Center, where the results of the Dvorak analysis by tropical meteorologists are recorded. While the storm type is not recorded for each Dvorak analysis, comments are often recorded that document some of their process, which can include information on storm type.

Last, the classifications from a known classifier are used for comparison. These classifications represent the

input of one of the authors, a meteorologist with an understanding of tropical systems, experienced with storm analysis at a forecast center, who is motivated to help the project succeed and has been trained in the Dvorak technique. Classifications from this classifier are denoted as user “Oscar.” Since classifiers are presented storm images randomly and without knowledge of the responses from others and the classifications from Oscar are not used in any of the expectation–maximization (EM) algorithm calculations, the resulting comparisons between the crowd and Oscar are independent.

3. Combining input using crowdsourcing algorithms

One of the primary tenets of crowdsourcing is that multiple answers to a question are more informative than one answer (Cox et al. 2015). An algorithm that learns tendencies of the classifiers is required because

- (i) Users are rarely unanimous in their assessment of an image’s storm type. Only 1% of images with 10 classifications have unanimous storm type selections.
- (ii) Imagery does not always receive a plurality for a given type. Only 40% of images with 10 classifications have one type that gets more than half of the votes.
- (iii) Users have varying capabilities in recognizing and discerning storm types. Similarly, users have varying motives in answering questions correctly.

The approach taken here follows Raykar et al. (2009) and Raykar et al. (2010), which account for the conditions above: when classifications disagree and when users have varying skill in their classifications. They describe an EM algorithm, which we adopt here to iteratively determine the storm type and the tendencies (e.g., specificity) of the classifier, following their modifications for classifications without features and modifying the approach for multiclass (nonbinary) models for the six categories of storm types.

a. Algorithm description

The goal herein is to determine the true type classification c of some image i . While c is unknown, multiple classifications of it exist from several classifiers j . Every image is classified R_i times by different classifiers, where each classifier selects class $k \in [1, K]$ (for a K -class algorithm). A single classification is represented by $y_i^j = k$, which described the j th classifier classifying image i as the k th type. The goal of the algorithm is to characterize the type that has largest probability (μ_{ic}) of being the true class. The algorithm is a three-step algorithm.

First, the probability that image i is type k is initialized with simple majority voting via

$$\mu_{ik} = \frac{\sum_{j=1}^{R_i} \delta(y_i^j, k)}{R_i}, \quad (1)$$

where $\delta(y_i^j, k) = 1$ if $y_i^j = k$ and is 0 otherwise.

Second, each classifier can be characterized using the matrix $\mathbf{A}^j = [\alpha_{ck}^j]$, where the individual elements are

$$\alpha_{ck}^j = \Pr[y^j = k | y = c]. \quad (2)$$

The sum of each column is 1 (i.e., $\sum_{k=1}^K \alpha_{ck}^j = 1$), since α_{ck}^j represents the probability that classifier j will assign a storm type k to an image given that the true storm type is c . This is calculated for Cyclone Center classifiers using

$$\alpha_{ck}^j = \frac{g_{1,ck} + \sum_{i=1}^{N^j} \mu_{ik} \delta(y_i^j, k)}{g_{1,ck} + g_{2,ck} + \sum_{i=1}^{N^j} \mu_{ik}}, \quad (3)$$

Where N^j is the number of classifications from a classifier and $\mathbf{G}_1 = [g_{1,ck}]$ and $\mathbf{G}_2 = [g_{2,ck}]$ are matrices of the beta prior distributions of \mathbf{A}^j , whose derivation is discussed below. The summation terms in Eq. (3) are similar to Eq. (1), but instead of summing each classifier for one image, it sums all classifications from one classifier. The μ_{ik} factor acts as a weight that is maximized when $y_i^j = k$ (i.e., the classifier selects k) and $\mu_{ik} = 1$ (i.e., the image is most likely type k). For classifiers with few classifications (N^j is small) the beta priors provide an a priori estimate. The influence of this beta prior decreases as N^j increases.

Third, the probability of an image being a type k (μ_{ik}) is calculated by combining the prevalence of each type (p_c):

$$p_c \sim p_k = \frac{\sum_{i=1}^N \mu_{ik}}{N}, \quad (4)$$

where N is the total number of images, with the probability from each classifier that an image is of type c (a_{ic}):

$$a_{ic} = \Pr[y_i = c] = \prod_{j=1}^{R_i} \prod_{k=1}^K (\alpha_{ck}^j)^{\delta(y_i^j, k)}. \quad (5)$$

Then

$$\mu_{ic} = \frac{a_{ic} p_c}{\sum_{c=1}^K a_{ic} p_c}. \quad (6)$$

Calculations of the \mathbf{A}^j [Eq. (3)] and μ_{ic} [Eq. (6)] are iterated until convergence. This produces a final description of each classifier (\mathbf{A}^j) and probabilities for each image (μ_{ic}), where the EM-derived storm type satisfies $\mu_c = \max_c \mu_{ic}$.

b. Binary classification: Does the storm have an eye?

Before discussing the complete six-class storm type classification, we present a case study of a binary (i.e., two class) analysis: the results of classifying images as having an eye or not. This instructive example demonstrates the learning algorithm outlined above.

One of the more discernable features in a tropical cyclone is when the system has an eye in the infrared satellite imagery (as will be shown below). The following presents a simplified analysis of the data ($K = 2$), where classifiers that select any of the canonical eye images (Fig. 1) are categorized as “eye storm” ($k = 1$), and all other images in Fig. 1 are classified as “non-eye storm” ($k = 0$). The result is the probability that a storm image has an eye. When the algorithm is complete, each classifier j is characterized by \mathbf{A}^j . Since the perfect classifier would be the identity matrix, classifiers are ranked using trace(\mathbf{A}^j), which sums the diagonal elements. The best active classifier (i.e., who has more than 300 classifications and 25 or more eye images) is Foxtrot, with

$$\mathbf{A}^{j=\text{Foxtrot}} = \begin{matrix} & c = 0 & c = 1 \\ \begin{matrix} k = 0 \\ k = 1 \end{matrix} & \begin{bmatrix} 0.992 & 0.001 \\ 0.008 & 0.999 \end{bmatrix} \end{matrix}$$

So when the algorithm found no eye (590 occurrences), Foxtrot agreed 99.2% of the time. Conversely, Foxtrot’s eye classification matched the EM algorithm on all 118 eye storms classified (n.b., the probabilities are limited to the range 0.001 and 0.999 to avoid numerical singularities in implementation). The worst active classifier (same criteria as above) is classifier Tango:

$$\mathbf{A}^{j=\text{Tango}} = \begin{bmatrix} 0.995 & 0.798 \\ 0.005 & 0.202 \end{bmatrix}.$$

In this case, Tango was good at identifying storms without an eye (99.5% of the time). However, when the algorithm resulted in an eye storm, Tango only selected the eye type about 20% of the time. Nonetheless, Tango’s contributions are still useful since the \mathbf{A}^j matrix provides probabilities based on Tango’s selections. The resulting classification from the algorithm [Eq. (6)] combines the probabilities from each classifier [Eq. (3)].

The results of this binary classification are shown for two cases in Fig. 3. In case 1—Hurricane Ward (1995)—nine classifiers chose “no eye” while four selected “eye,” one of the latter included someone with a top 10 ranked \mathbf{A}^j

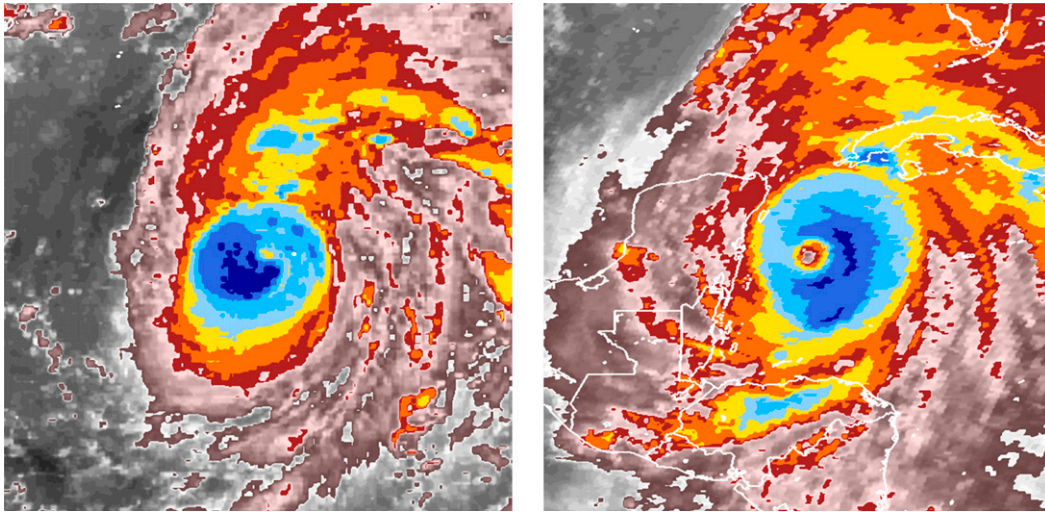


FIG. 3. Images from Cyclone Center for (left) Hurricane Ward (1995) at 0900 UTC 19 Oct 1995 and (right) Hurricane Wilma (2005) at 0300 UTC 21 Oct 2005.

(i.e., not Foxtrot, but a rank very close to Foxtrot). Thus, despite being an image for which the majority vote would be “no eye”, the resulting probability is eye ($\mu_i = 0.9998$). Conversely, 11 of 14 classifiers selected an “eye” storm for case 2—Hurricane Wilma (2005). One of the “no eye” selections came from Tango. In this case, the weight from Tango’s selection was very small and had little bearing on the resulting eye probability ($\mu_i = 1.00$).

The results of the binary classification are compared with Oscar and ADT-H in Table 1. The Heidke skill score (HSS) is used, since it shows skill versus random selections when $HSS > 0$; a perfect score is 1. This comparison uses a subset of all images, which only includes images with 10 or more classifications that are over ocean (and thus have valid ADT-H classifications). When Oscar is compared to each individual classification (49 336 matchups), HSS is only 0.66. Using the classification by the majority vote from the individual classifications increases the skill to 0.82. The

EM algorithm is often the same as the majority, but in some cases the majority may be affected by a classifier with lower skill [i.e., smaller values of $\text{trace}(\mathbf{A}^i)$]. Thus, the EM algorithm provides a slight improvement on skill, increasing HSS to 0.87. The classifications from Oscar and EM are both compared to the ADT-H, with both having a skill score near 0.77. Thus, the EM algorithm in this instance skillfully represents the classifier Oscar.

c. Storm type analysis: Six-class classifications

The analysis of the binary case (eye vs no eye) can be extended to storm type, where there are six types ($K = 6$): no storm, curved band, embedded center, eye, shear, and posttropical; \mathbf{A}^i is now a 6×6 matrix.

Again, the performances of the classifiers are evaluated using $\text{trace}(\mathbf{A}^i)$. The best active classifier for storm typing (now selecting from classifiers with 1000 or more classifications) is Romeo:

$$\mathbf{A}^{i=\text{Romeo}} = \begin{matrix} c = & \text{NS} & \text{CBD} & \text{EMB} & \text{EYE} & \text{SHR} & \text{PT} \\ \left[\begin{array}{cccccc} 0.87 & 0.01 & 0.00 & 0.01 & 0.26 & 0.03 \\ 0.00 & 0.81 & 0.16 & 0.02 & 0.02 & 0.06 \\ 0.00 & 0.08 & 0.77 & 0.02 & 0.03 & 0.00 \\ 0.02 & 0.01 & 0.02 & 0.95 & 0.00 & 0.01 \\ 0.08 & 0.09 & 0.05 & 0.00 & 0.66 & 0.36 \\ 0.02 & 0.00 & 0.00 & 0.00 & 0.01 & 0.54 \end{array} \right] & \begin{array}{l} k = \text{NS} \\ k = \text{CBD} \\ k = \text{EMB} \\ k = \text{EYE} \\ k = \text{SHR} \\ k = \text{PT} \end{array} \end{matrix}$$

On average, Romeo agrees with the EM algorithm nearly two-thirds of the time. Each column sums to one, showing that a classification from Romeo is actually a probability of the true type based on Romeo’s selection.

Romeo’s best categories are discerning an eye storm (agreeing with the EM algorithm 95% of the time) and no storm (87%). Tango again appears as the worst active classifier in the six-class algorithm as well:

TABLE 1. The HSS for various comparisons of the eye vs no-eye classification. All comparisons are when an image has 10 or more classifications and when ADT-H classifications are valid (i.e., circulation center is over water), which is 3727 images. However, there are 49 336 comparisons between Oscar and each individual classification.

Test	HSS
Oscar vs individuals	0.66
Oscar vs majority voting	0.82
Oscar vs EM algorithm	0.87
Oscar vs ADT-H	0.77
ADT-H vs EM algorithm	0.77

$$\mathbf{A}^{j=\text{Tango}} = \begin{bmatrix} 0.02 & 0.01 & 0.00 & 0.00 & 0.00 & 0.03 \\ 0.22 & 0.34 & 0.34 & 0.23 & 0.26 & 0.14 \\ 0.02 & 0.01 & 0.13 & 0.17 & 0.07 & 0.01 \\ 0.00 & 0.00 & 0.01 & 0.21 & 0.00 & 0.00 \\ 0.72 & 0.64 & 0.52 & 0.39 & 0.67 & 0.81 \\ 0.02 & 0.00 & 0.00 & 0.00 & 0.00 & 0.02 \end{bmatrix}.$$

Tango agrees with the EM algorithm less than 1 in 4 times and shows a significant bias toward selecting shear scenes. No matter the true type (column), the type Tango is most likely to select is SHR. Both Tango and Romeo had more than 1000 classifications. Clearly from Fig. 2, very few classifiers have this many classifications. So how many classifications does one need to adequately understand their tendencies (i.e., calculate an accurate \mathbf{A}^j)?

d. *Prior distributions—Characterizing classifiers with few classifications*

About half of all classifiers have fewer than 3 classifications, totaling more than 23 000 classifications (cf. Fig. 2). These are too few to accurately calculate \mathbf{A}^j , so a prior distribution is required. A set of images that had 10 or more classifications was used to calculate the priors; \mathbf{A}^j can then be calculated for the users that contribute to this set of images and then averaged \mathbf{A}^j to estimate the prior. The mean \mathbf{A}^j is

$$\overline{\mathbf{A}^j} = \begin{bmatrix} 0.47 & 0.09 & 0.02 & 0.02 & 0.09 & 0.16 \\ 0.25 & 0.56 & 0.14 & 0.11 & 0.24 & 0.25 \\ 0.05 & 0.18 & 0.67 & 0.21 & 0.27 & 0.05 \\ 0.01 & 0.02 & 0.05 & 0.60 & 0.02 & 0.02 \\ 0.14 & 0.12 & 0.10 & 0.05 & 0.35 & 0.17 \\ 0.08 & 0.03 & 0.01 & 0.01 & 0.03 & 0.33 \end{bmatrix}.$$

The prior is included in Eq. (3) by calculating elements of the beta priors (\mathbf{G}_1 and \mathbf{G}_2) for each element of \mathbf{A}^j following Raykar et al. (2010), where

$$g_{1,ck} = \frac{(-\overline{\alpha_{ck}^j}^3 + \overline{\alpha_{ck}^j}^2 - \overline{\alpha_{ck}^j} \sigma_{\alpha_{ck}^j}^2)}{\sigma_{\alpha_{ck}^j}^2} \quad (7)$$

and

$$g_{2,ck} = \frac{g_{1,ck}(1 - \overline{\alpha_{ck}^j})}{\overline{\alpha_{ck}^j}}. \quad (8)$$

The overbars and sigma denote the mean and standard deviation for each element of \mathbf{A}^j , respectively. While this prior helps estimate performance when classifications are few, how many classifications are needed to calculate an accurate \mathbf{A}^j for an individual?

The number of classifications needed to fully understand a classifier is estimated by simulating their \mathbf{A}^j using a subset of their classifications. The number of classifications in the subset is varied from 1 to 300, calculating \mathbf{A}^j 200 times using a random subset of classifications. This is compared to the \mathbf{A}^j calculated based on all of their classifications. Results are shown for two users: Oscar and Tango. Oscar performed more than 7000 classifications, and Tango (discussed above) performed ~1400. The results are shown in Fig. 4.

For Oscar, the a priori α_{ck}^j already explains 75% of the variance of α with just one classification. This is important because classifications from Oscar were not used in deriving the prior distribution, only in this post comparison. At about 40 and 180 classifications, the explained variance for Oscar passes 80% and 90%, respectively. So the prior describes the performance quite well even when the number of classifications is small. Conversely for Tango, whose \mathbf{A}^j is somewhat of an outlier (having the lowest score in the group of active classifiers), the results show lower explained variance for fewer classifications. This is because Tango's \mathbf{A}^j is very different than the prior. So how many classifications from Tango are needed to recognize that Tango's \mathbf{A}^j was different? The explained variance passes similar thresholds (80% and 90%) at 100 and 180 classifications. This suggests that, initially, the algorithm needs more classifications from classifiers with lower scores to have an accurate \mathbf{A}^j than those with higher scores. However, both Tango and Oscar reached 90% explained variance at roughly 180 classifications. Thus, users with about 180 or more classifications are well described by their \mathbf{A}^j .

e. *How many classifications of an image are enough?*

The goal of crowdsourcing in the Cyclone Center project is to obtain multiple classifications of an image from numerous classifiers in order to estimate the true attributes of an image—in this case, storm type. At the outset, the CC website collected a large number of classifications per image as a conservative approach.

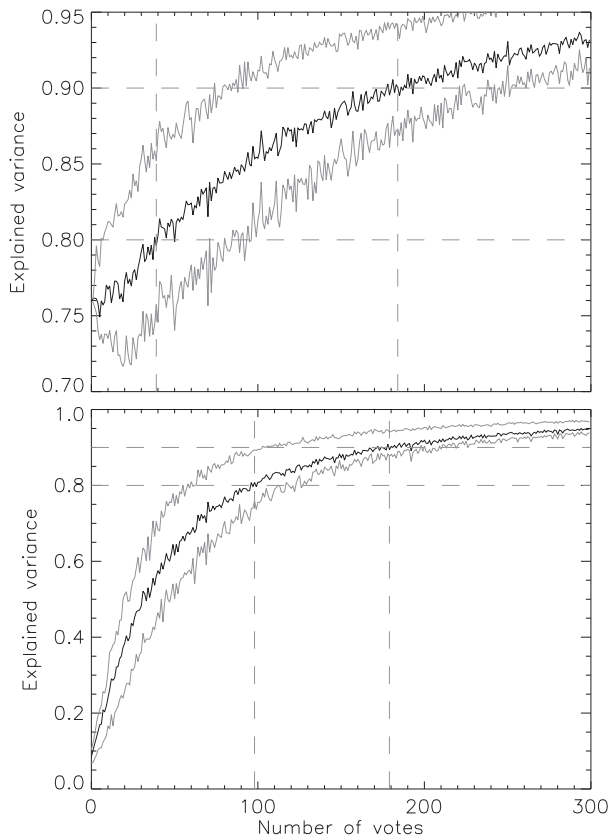


FIG. 4. Simulated explained variance of a classifier's \mathbf{A}^i as a function of the number of classifications, where 200 simulations were made for each number of classifications. The mean explained variance (black) and the 25th and 75th percentiles (gray) are shown for two users: (top) Oscar and (bottom) Tango.

There are 577 images with 30 or more classifications of storm type.

Classifications from a subset of these were used to estimate the number of classifications needed. Storm types were derived by varying the number of classifications from 1 to 29. Results were compared to the type derived from the full set of classifications. The HSS was used to evaluate the skill for each storm type; the result is shown in Fig. 5. The general trend is for less skill at fewer classifications, with HSS approaching 1 at 29 classifications. Near 10 classifications, the HSS values are above 0.9 and above that slowly approach 1. There is some dependence of HSS on storm type. For instance, the eye storm types appear to reach maximum HSS values with only 6 classifications, while the post-tropical storm type begins with no skill ($\text{HSS} < 0$), and shear has the lowest skill at 10 classifications. Given these results, 10 classifications appear to be a sufficient number to determine most storm types, though more classifications could be needed for types with less certainty (e.g., posttropical).

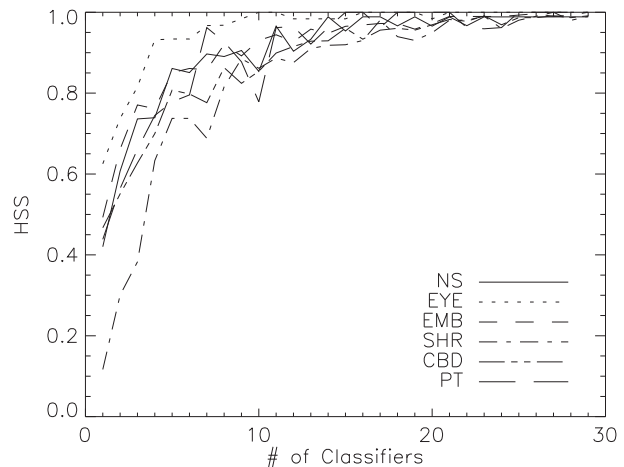


FIG. 5. HSS for each storm type as a function of the number of classifications used in the EM algorithm.

To analyze the consistency of this result, we compared the EM algorithm classifications for images with more than 20 classifications ($\text{Type}_{\text{full}}$) with simulated 10-member subsets for each image (Type_{10} , subsetting each image 10 times), creating 18 560 matchups between an image with 20 or more classifications to just those using the 10-classification subset. The results are provided in Table 2. The classifications of storm type agree $\sim 95\%$ of the time. The skill scores for each type are more than 0.87, the eye type having the largest skill score of 0.97. The highest errors (i.e., off-diagonal values) represent confusion between embedded centers and other types, but even those are small percentages of the total for each type. This suggests that the classification of an image using input from any 10 people is consistent with 10 other classifiers for the same image and is also consistent with a doubling (or more) of the number of classifications. Therefore, in the following discussion a classification of a storm type is considered complete when the image receives 10 or more classifications. It is also possible to identify occurrences when the 10 classifiers have lower

TABLE 2. Demonstration of consistency in classifications through comparisons between classifications of images with 20 or more classifications ($\text{Type}_{\text{full}}$) and 10-member random subsets of those full sets (Type_{10}) with the HSS for each type.

		Type ₁₀						N
		NS	CB	EMB	EYE	SHR	PT	
Type _{full}	NS	0.95	0.02	0.00	0.00	0.01	0.01	1770
	CB	0.02	0.93	0.03	0.00	0.02	0.00	5190
	EMB	0.00	0.02	0.96	0.00	0.02	0.00	7750
	EYE	0.00	0.00	0.04	0.96	0.00	0.00	1510
	SHR	0.01	0.03	0.02	0.00	0.93	0.00	1780
	PT	0.02	0.02	0.00	0.00	0.03	0.94	560
	HSS	0.94	0.92	0.94	0.97	0.87	0.94	

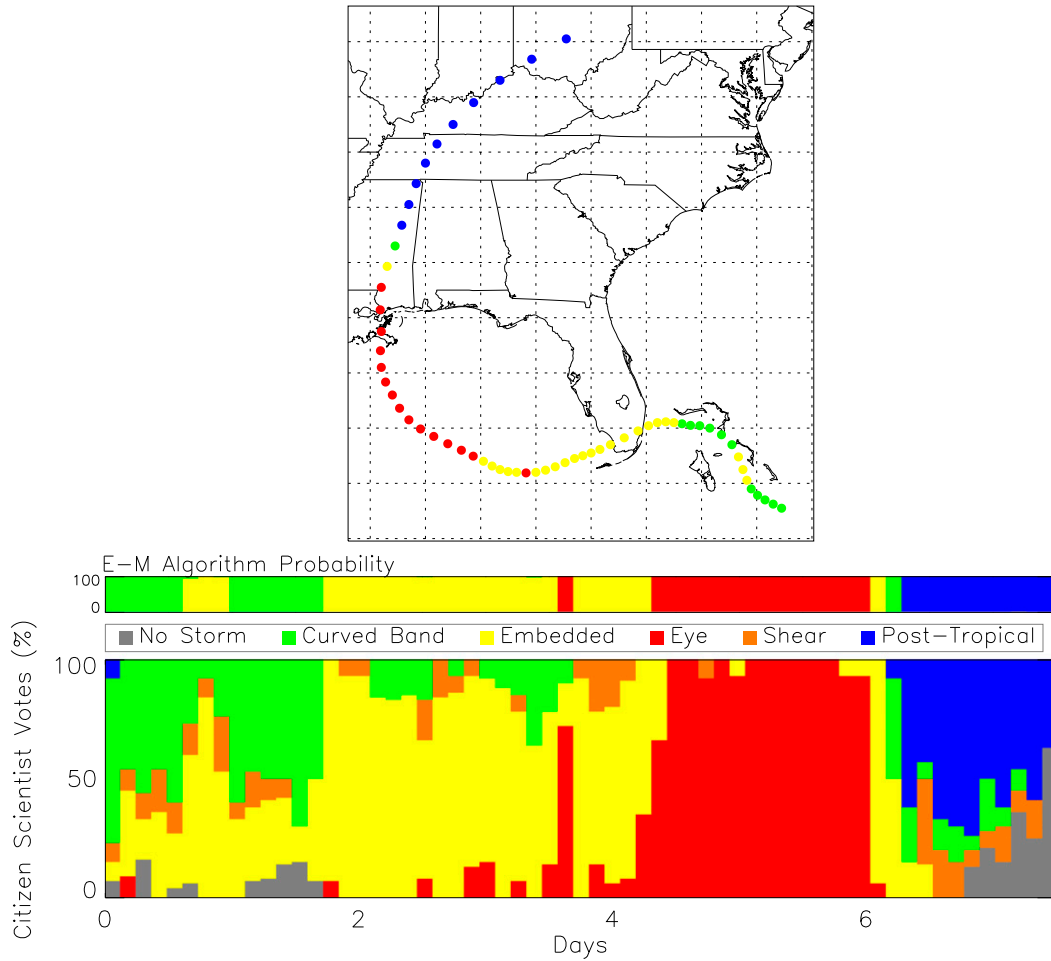


FIG. 6. EM algorithm results for storm types based on imagery from Hurricane Katrina (2005) in the North Atlantic. (bottom) The percentages of the raw classification with the resulting EM probabilities above it. (top) The map plots the storm type along the track of the system at approximately 3-h intervals.

skill and reintroduce those images into the system for more classifications.

f. Assessing algorithm performance

The EM algorithm described above produces a characterization of storm type for each image with 10 or more classifications. An example of the resulting type is provided in Fig. 6 for Hurricane Katrina (2005) in the North Atlantic Ocean. The analysis shows the various individual selections (bottom plot). The EM algorithm converts the individual classifications—using the tendencies of each classifier from \mathbf{A}^i —to estimate the probabilities of each image’s storm type. The analysis classifies Katrina prior to landfall in Florida as oscillating between curved band and embedded center. As the storm emerges from Florida over the Gulf of Mexico, the system remains an embedded center. A brief hint of an eye is noted just prior to day 4, then an eye emerges

near day 5, which lasts until landfall. The storm quickly dissipates over land into posttropical-type images over Mississippi and Tennessee. It should be noted that the EM algorithm operates on each image independently without any prior information on the storm type of nearby times. The EM algorithm reduces the noisiness of the numerous individual storm type selections in the bottom panel into a smooth, interconsistent set of storm types in the top panel. There are some outlier classifications in the first 4 days (with some classifiers selecting eye and shear early on); nonetheless, the algorithm selects the highest probability based on the classifications and the \mathbf{A}^i of each classifier.

The analysis of Katrina provides a qualitative confirmation of the EM algorithm performance, and the consistency of the analysis (e.g., Table 2) provides assurance that the process is repeatable. The following analysis provides quantitative analysis where we compare the

results of the EM algorithm to three other datasets described earlier: a known user (Oscar), ADT-H, and fix data. The results of the comparison are shown in Fig. 7.

1) COMPARISONS WITH A KNOWN CLASSIFIER

The comparisons between classifications from Oscar and those from the EM algorithm (for images with 10 or more classifications) are provided in Fig. 7a. As a reminder, these classifications are not used in the above calculations of A^i or the priors, so they are an independent verification. The maximum percentage in each row agrees with Oscar. The type with the most agreement (and highest skill) is eye types (HSS = 0.87). While there is much agreement between Oscar and EM on post-tropical systems (78%), lower skill (HSS = 0.52) is caused by the EM algorithm classifying many of Oscar's no-storm types as posttropical, resulting in lower skill for no-storm types (HSS = 0.46). This is attributed to the condition that the convection has been largely transported away from the circulation center during the posttropical stage and can appear like no storm is present. There is some agreement for the curved band and embedded center types (HSS values of 0.45 and 0.59, respectively), but some overlap as well (e.g., the EM algorithm flags 19% of Oscar's embedded centers as curved bands). The lowest skill is found in the shear type (HSS = 0.14), which is marginally skillful. For the images classified by Oscar as shear, the EM algorithm agrees 40% of the time but also has large fractions for curved band (22%) and embedded center (19%). Nonetheless, classifications for each type are skillful.

2) COMPARISON WITH ADT-H

We compare the ADT-H types to both the EM algorithm and the classifications from Oscar for perspective. It is worth repeating here that the ADT-H results are not the operational ADT product, but the same algorithm applied to the lower-resolution HURSAT dataset.

The comparisons between the EM algorithm and ADT-H are provided in Fig. 7b. The ADT-H has many of the same storm types with the addition of central dense overcast (CDO) and irregular CDO (IrrCDO), which are combined in Figs. 7b and 7c as EMB/CDO. The number of classifications here is much larger since there is an ADT-H value for all images (except for when the storm center is over land). The skill is not calculated since there is not a one-to-one matchup with EM types. The ADT-H curved band types are classified correctly 48% of the time. Since the ADT-H has no no-storm classification, 11% of the curved band systems were called no storm by the EM algorithm (similarly, 21% of shear systems were called no storm by the EM). The ADT-H embedded center images are largely (70%)

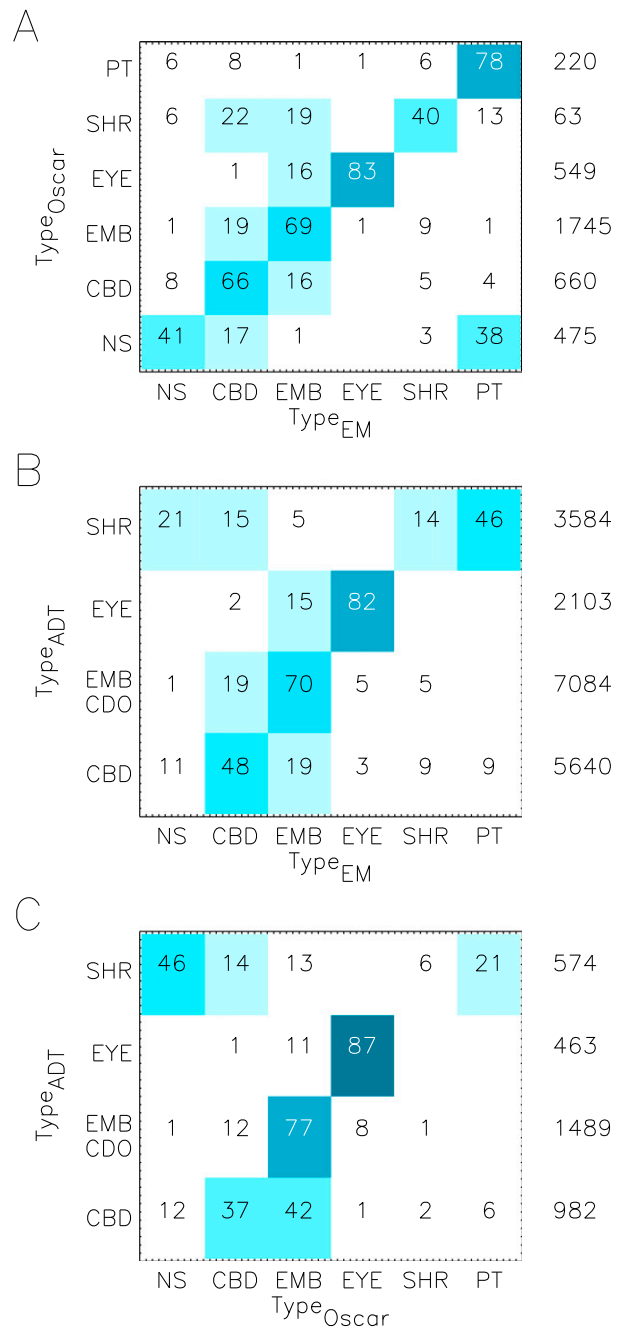


FIG. 7. Heat map distribution of classifications between two classification methods, where the values are the percentages of the total occurrences of that type (which is the number on the right) and percentages are rounded to integer values for (a) EM algorithm vs Oscar, (b) EM algorithm vs ADT-H, and (c) Oscar vs ADT-H.

called the same by the EM algorithm, with 19% being called curved band. There is also high agreement on eye systems (nearly the same as between EM and Oscar). Last, the ADT-H shear type appears to be a combination of two EM types: the no storm and posttropical, since

there is no equivalent of the posttropical or no storm in the ADT-H types. The comparison between the classifications of Oscar and ADT-H (Fig. 7c) provide perspective on the EM/Oscar and EM/ADT-H comparisons.

The comparison between ADT-H and the EM algorithm is consistent with the comparison to classifications from Oscar. Most of the percentages are within $\pm 10\%$ of the EM-ADT-H comparisons. Classifications from Oscar had large agreement for embedded centers. Oscar identified eye images with slightly more agreement than the EM algorithm (87% vs 82%). Oscar also showed similar tendencies to classify ADT-H shear images as either no storm or posttropical. In summary, while there is not complete agreement between the ADT-H and the EM algorithm, there are significant similarities between the ADT-H comparisons with the algorithm and Oscar. Thus, it appears that the classifications from the EM algorithm are consistent with Oscar, and both Oscar and the EM algorithm are consistent when compared to the objective ADT-H algorithm.

3) COMPARISON WITH FIX DATA

The comments portions of the fix data often provide some hint of the storm type. We parsed the comments to characterize storm type, only using fixes based on the subjective Dvorak technique. These are largely systems over the North Atlantic and eastern Pacific. We ignored any systems that referenced microwave data in the comments (e.g., “Center location strongly influenced by 1000 UTC SSMI”), as that suggests other information helped with the classification. In general, the same classes are noted in the fix data as the six Cyclone Center types, but the fix data denotes many storms as too weak to classify (TWTC), so we retain that term. It should be noted, that analysts were working with very different satellite data. Often, they have access to 1-km visible imagery (compared to the 8-km imagery in HURSAT) and other channels (visible imagery, etc.). They also have access to microwave data, which often affects their analysis (as noted above). While the times are somewhat coincident (within 15 min), the analysts had access to more than one satellite image and the ability to animate images through time and zoom in or out, all of which are capabilities not available to Cyclone Center classifiers.

There is general agreement between the EM algorithm results and the classifications in the fix data (Fig. 8), which is consistent with the previous comparisons. The TWTC type is classified by the EM algorithm as no-storm type (33%) or posttropical type (31%), which is consistent with comparisons to Oscar. The EM algorithm also tends to classify fix data curved bands as embedded centers (85%) than in previous comparisons.

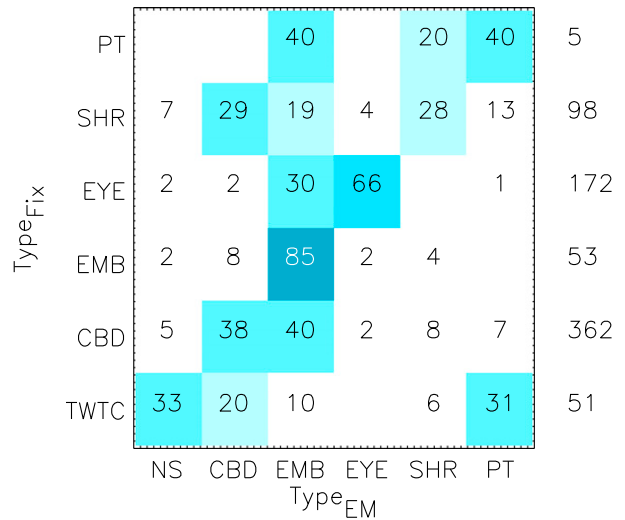


FIG. 8. As in Fig. 7, but for comparisons between EM and fix dataset, where the fix category too weak to classify is encoded TWTC.

Conversely, the eye type has less agreement than in previous comparisons. This can likely be attributed to the differences in the underlying satellite imagery available to the analysts (with the ability to zoom, animate, and view other satellite imagery) versus the capabilities of the Cyclone Center website interface (one image, one color scale at 8-km resolution). Thus, it is understandable that the EM algorithm only recognizes 65% of the eyes from the fix data. Again, the shear system has the most confusion, expressed in low values across the board. When analysts call a system shear, the EM algorithm tends toward curved band (29%), shear (28%), or embedded center (19%).

Given the success of the algorithm’s performance versus a known classifier (Oscar), an objective algorithm, and fix data from analysts (when available), we investigate the distribution of these storm types as determined by the EM algorithm.

4. A climatology of storm types

The climatology of storm types is categorized herein as the distribution of the storm types, where they occur, and their relationship to storm evolution. In these cases, the storm types are compared with maximum sustained wind speeds that derive from best track data. The wind speed data are completely independent of the storm type information derived by the EM algorithm. The best track data used here are the IBTrACS v03r02 data (Knapp et al. 2010), from which the HURSAT data were derived. To avoid discrepancies in wind speed averaging periods and other practices (Knapp and Kruk 2010;

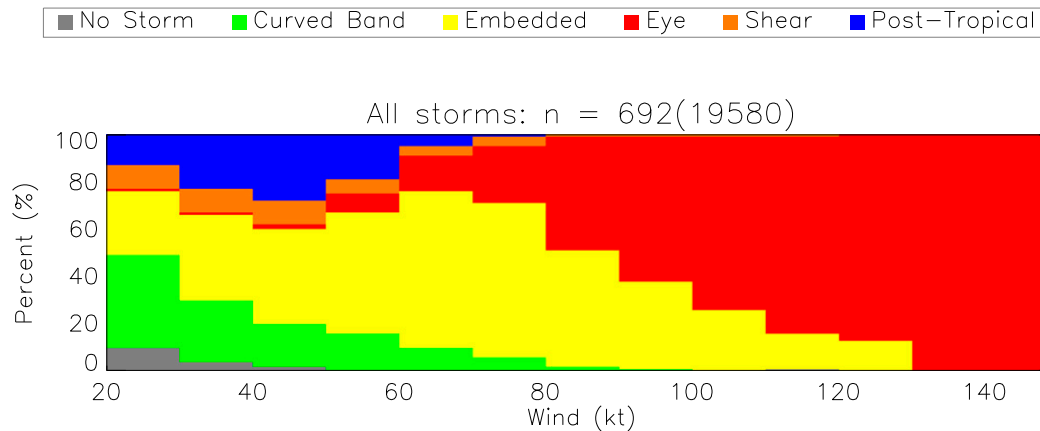


FIG. 9. Distribution of storm type by wind speed, where the storm type is determined from the EM algorithm and the wind speed is from the best track data. The distribution is based on 19 580 images from 692 individual storms. Only images with 10 or more classifications are included in this analysis.

Schreck et al. 2014), we focused on the maximum wind speed available from all agencies reporting on a storm, in order to represent the highest likely storm winds at any time. Winds are reported in the international standard unit for tropical cyclone intensity of knots (kt; $1 \text{ kt} = 0.5144 \text{ m s}^{-1}$).

a. Frequency of storm types

The distribution of the EM algorithm storm types with the wind speeds from best track are shown in Fig. 9 (in 10-kt increments). This shows no-storm types having a maximum percentage for the weakest wind speeds, decreasing as wind speeds increase and becoming negligible beyond 50 kt. The curved band types have the largest percentage at the weakest winds (39%), which also decrease in fraction with increasing wind speed. Curved bands appear associated with stronger winds than no storms. Curved bands, though, are rare when winds are above 90 kt. The embedded centers are present at nearly all wind speeds (though not above 140 kt) with a peak (66%) at 75 kt. The eye type is dominant for images with wind speeds above 90 kt. The eye type is less frequently observed at lower wind speeds and becomes rare below 50 kt. The shear type has the lowest percentage of the EM types. It has a maximum at lower wind speeds, is only present through 80 kt, and is never more than 10% of images at any wind speed. Last, the posttropical type represents the storm's transition away from the tropics and often its cyclolysis. Thus, it peaks in percentage near 45 kt and is not very prevalent above 70 kt. We conclude that the separate types as classified by EM appear to be related to best track winds.

Another analysis of storm type is provided in Fig. 10, where the fractional portions of images are distributed in time relative to three significant points in a storm's

lifetime: 1) the storm's first image (labeled genesis), 2) the first occurrence of its lifetime maximum intensity (LMI), and 3) the last image of the storm (termed here cyclolysis). The top plot shows the distribution of storm types for all available completed storm images; it is then separated by the type of storm: LMI < 65 kt (tropical storms), hurricanes ($65 < \text{LMI} < 115 \text{ kt}$), and intense hurricanes ($\text{LMI} > 115 \text{ kt}$). The analysis uses all completed images (19 580), which derive from 692 separate storms. Not all of these storms are complete, but any complete image can be placed relative to the three times listed above.

The overall distribution (Fig. 10, top) shows each storm type to varying degrees. The no-storm type is present primarily at the start and near the end of a storm's lifetime. The shear type is present for a small fraction throughout the storm's life cycle. The largest fraction of curved bands is at the storm genesis and decreases with time, while the embedded centers peak in fraction shortly before LMI, and the eye types show a sharp peak at LMI. These patterns combine all storm strengths.

Tropical storms have a small fraction of eye types, while eyes dominate the intense hurricanes near LMI. The maximum fraction of the eye type at any given time of a storm increases from only 2% for tropical storms to 32% for hurricanes and 80% for intense hurricanes. This is consistent with Vigh et al. (2012), who characterize the median intensity for initial satellite eye formation around 60 kt. Most storms forming an eye are reaching hurricane intensity. Curved bands are present in large percentages for each intensity level. While it remains a significant fraction throughout a tropical storm's lifetime, curved bands disappear near LMI for intense hurricanes. The shear type is more prevalent

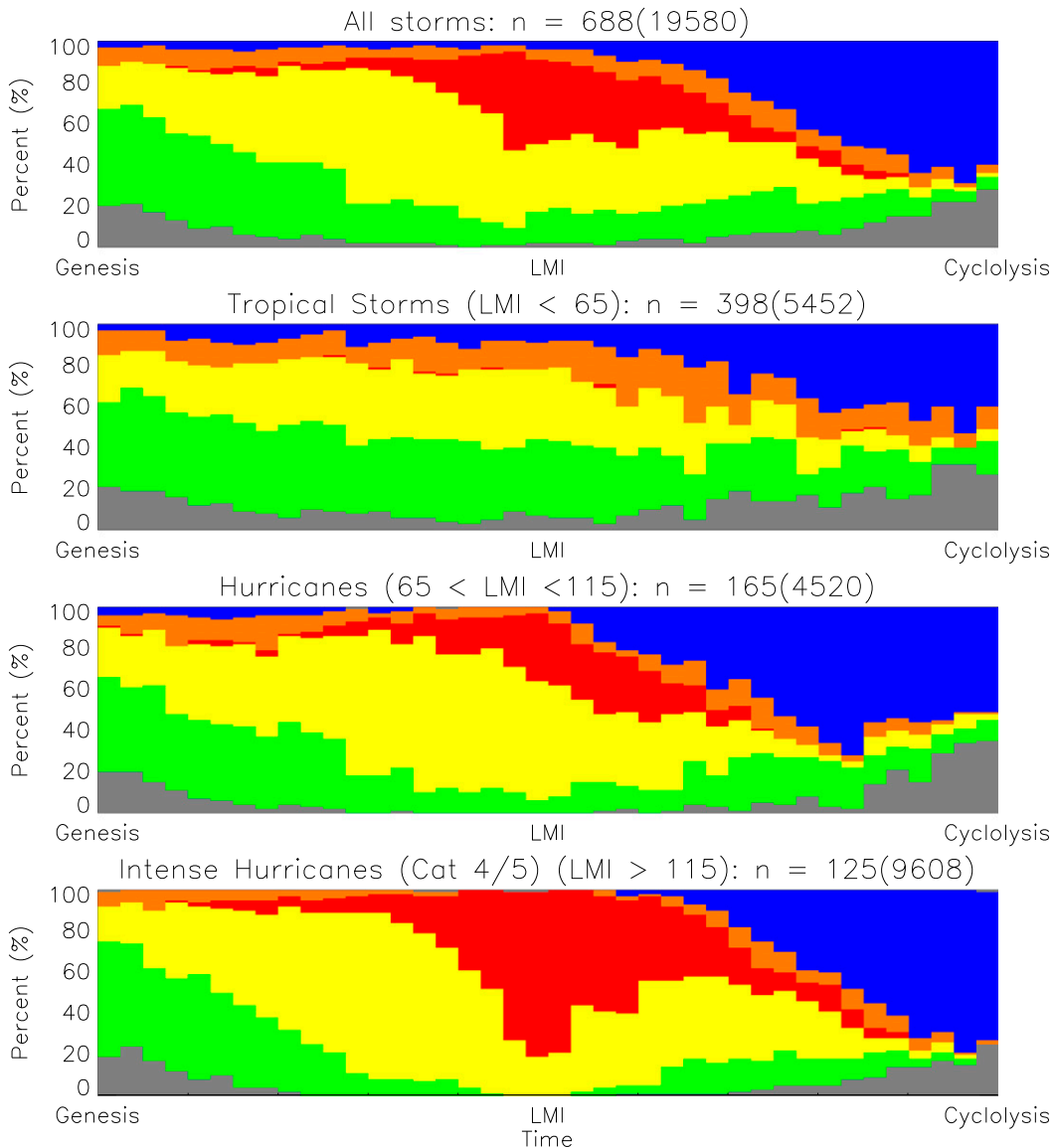


FIG. 10. As in Fig. 9, but showing storm type distribution by age of the storm at the time of the image relative to three points: storm genesis, first occurrence of LMI, and storm cyclolysis for (from top to bottom) all storms (19 580 images of 692 different storms), those storms that only reach tropical storm (TS) strength (5452 images of 398 systems), and so on for hurricanes and intense hurricanes.

for tropical storms than hurricanes and intense hurricanes. Last, the fraction of posttropical type increases near cyclolysis and appears more frequently for more intense systems.

The frequency of eye images is further investigated in Table 3. While numerous storms have a portion of their images complete (10 or more classifications), few storms have all their images completed (we define an entire storm as complete when 80% of a storm’s images are complete). However, from this small sample it is apparent that more intense storms have significantly more

eye images. All completed intense hurricanes have at least 1 eye image, with the mean being 25 eye images (the equivalent of 3 days with an eye). A total of 9 of 13 completed hurricanes had eyes, with about 8 eye images per hurricane. Only 6 tropical storms are complete, so the occurrence of one eye image for one of the storms is not likely statistically significant. Even though there are few completed images, the resulting distribution of eye images does shed light on the fraction of eye images by storm intensity, but clearly more classifications are needed to draw conclusions.

TABLE 3. Frequency of eye images based on storm type, where completed storms are defined as having 80% or more of their images complete.

	All storms	Tropical storm	Hurricane	Major hurricane
Completed storms	57	6	13	38
Completed storms with 1 or more eye scenes	48	1	9	38
Fraction with eye scenes		16%	69%	100%
Mean No. of eye images when at least 1 is present		1	8	25

b. Location of storm types

The fraction of storm images complete varies by basin, so there are not yet enough results to investigate patterns in a specific basin. We can, however, look at the zonal distribution of storm types, as shown in Fig. 11. This shows the distribution of types with latitude; the hemispheric medians are also provided.

The posttropical type is unique, with peaks outside of the tropics. The other types have their peaks (and their hemispheric medians) in the tropics. While the distributions are somewhat similar, the medians change in a logical fashion: types that occur earlier in the storm's life cycle (e.g., curved band and embedded center) occur closer to the equator, while those that occur later are farther poleward (e.g., eye). The exception is the distribution of no-storm types. This can be explained by the proclivity of users to assign both posttropical and no-storm near the end of a storm's life (cf. Figs. 7 and 10). More images are needed to further investigate the spatial distribution of types, but this initial analysis does show that the types from the EM algorithm match a conceptual zonal ordering.

c. Evolution of types

Another way to analyze the initial results of the EM algorithm is in how the storm types change in time, which is shown in Fig. 12. Each row contains the percentage distribution of storm type three hours after the occurrence of a particular type. For example, about half of no-storm scenes are followed by no-storm scenes, with 26% becoming curved bands. Embedded centers, however, are more likely to remain the same (75%) and tend to become either curved bands (13%) or eye storms (7%). It is interesting to see that eyes are really quite rare: they only follow about 12% of other images. When they do occur, they tend to happen several images in a row (82% of eyes are followed by eyes). They also tend to form mostly out of embedded centers (since 7% of those create eyes), which are the most prolific eye producers (aside from eye storms).

Results were also separated by different storm intensities (hurricanes and intense hurricanes, not shown), which resulted in minor differences from Fig. 12. However, intense hurricanes did show a slightly greater propensity to maintain eyes (86% vs 82%). Also, 10% of embedded centers produced eyes (vs 3% for hurricanes), which is consistent with all intense hurricanes having at least one eye image (cf. Table 3).

5. Summary

Initial analysis of storm types from a citizen science project was described here. The goal of Cyclone Center is to provide the best possible reanalysis of satellite-derived tropical cyclone characteristics from a homogeneous historical database composed of infrared imagery with the lowest common spatial resolution for

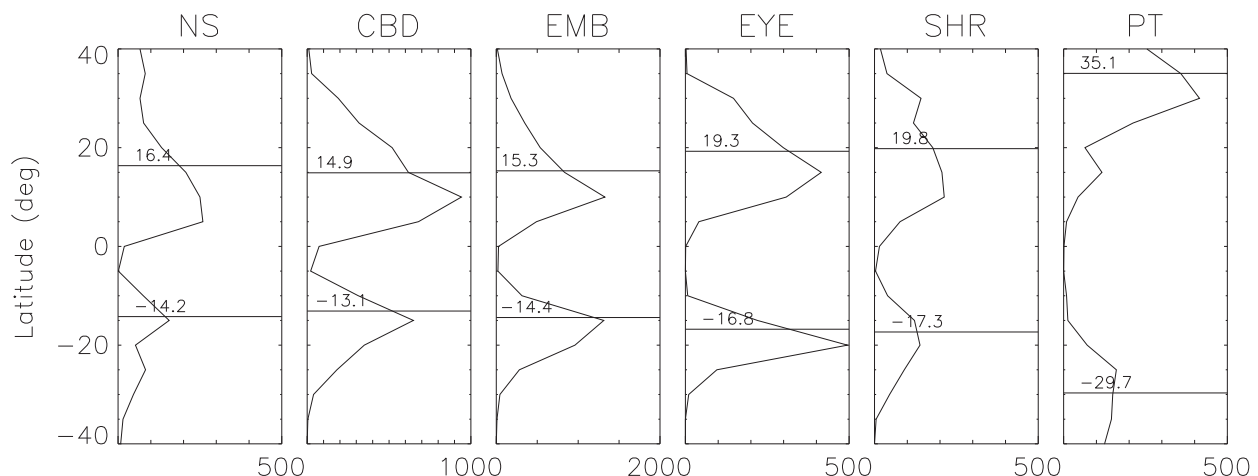


FIG. 11. Zonal distribution of storm types. The hemispheric median latitudes for each type are denoted by the horizontal lines.

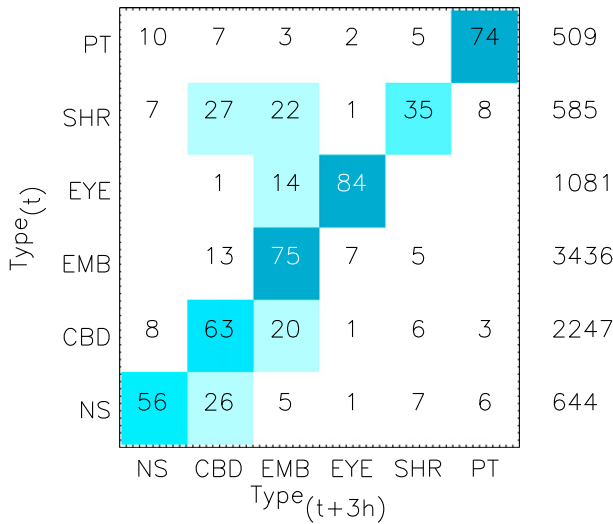


FIG. 12. As in Fig. 7, but showing distribution of types 3 h after any given type.

use in long-term, global analyses. The goal is to provide information for a future synthesis of tropical cyclone data, not to replace any current best track datasets. While the end goal of the project remains far off, this analysis of the data, through the determination of the storm type, reveals some interesting aspects of the project and of tropical cyclones in general.

A statistical algorithm—called the expectation-maximization algorithm (EM)—was used to combine (the sometimes) disparate classifications from numerous classifiers into a consistent and accurate representation of the storm type. In the development of the algorithm, we also determined how many image classifications were needed (10) as well as how many classifications from a classifier were needed to best understand their tendencies (approximately 180). For users with fewer classifications, an a priori distribution provides a starting point.

The results of the EM algorithm were compared with classifications from one of the authors, with an objective satellite analysis algorithm, and with information from satellite analysts (via comments in tropical cyclone center fix data). The results showed that the classifications from the website visitors—analyzed through the lens of the EM algorithm—are consistent, with better agreement for some types (e.g., eye patterns) than others (e.g., shear pattern). The resulting cursory analysis of storm type shows how storm types relate to best track wind speeds in a system and during a storm’s lifetime. While preliminary, the resulting storm types are consistently derived, agree with other means of estimating type, and should prove useful in further analysis of project data. For instance, Hennon et al. (2015) showed a simplistic analysis of Cyclone Center data produced intensities

with a root-mean-square error (RMSE) near 18 kt. It is intended that future analysis can make use of EM-derived information from Cyclone Center questions—such as storm type derived here—to help refine these intensity estimates toward lower RMSE values.

Acknowledgments. The authors acknowledge others on the Cyclone Center science team: Paula Hennon, Michael Kruk, Jared Rennie, Carl Schreck, Scott Stevens, and Peter Thorne. We are also extremely grateful for the support of the Citizen Science Alliance development team. Funding and support for Dr. Hennon was provided in part by the Risk Prediction Initiative of the Bermuda Institute of Ocean Sciences and the Cooperative Institute for Climate and Satellites–North Carolina (CICS-NC). Dr. Matthews is supported by NOAA through the CICS-NC under Cooperative Agreement NA14NES432003. We also appreciate the constructive comments from Christopher Landsea, Matthew Eastin, and anonymous reviewers. Lastly, we are grateful for the contributions from the numerous citizen scientists who have contributed countless hours in providing more than half a million classifications. In particular, we thank baha23, bretarn, shocko61, and skl6284, who have each provided more than 6500 classifications—the equivalent of 1 year of HURSAT data.

REFERENCES

Cox, J., E. Y. Oh, B. Simmons, C. Lintott, K. Masters, A. Greenhill, C. Graham, and K. Holmes, 2015: Defining and measuring success in online citizen science: A case study of Zooniverse projects. *Comput. Sci. Eng.*, **17**, 28–41, doi:10.1109/MCSE.2015.65.

Dvorak, V. F., 1984: Tropical cyclone intensity analysis using satellite data. NOAA/NESDIS Tech. Rep. 11, 47 pp.

Emanuel, K., 2005: Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, **436**, 686–688, doi:10.1038/nature03906.

Hennon, C. C., and Coauthors, 2015: Cyclone Center: Can citizen scientists improve tropical cyclone intensity records? *Bull. Amer. Meteor. Soc.*, **96**, 591–607, doi:10.1175/BAMS-D-13-00152.1.

Knaff, J. A., and R. M. Zehr, 2007: Reexamination of tropical cyclone wind–pressure relationships. *Wea. Forecasting*, **22**, 71–88, doi:10.1175/WAF965.1.

—, D. P. Brown, J. Courtney, G. M. Gallina, and J. L. Beven, 2010: An evaluation of Dvorak technique–based tropical cyclone intensity estimates. *Wea. Forecasting*, **25**, 1362–1379, doi:10.1175/2010WAF2222375.1.

Knapp, K. R., and J. P. Kossin, 2007: New global tropical cyclone data set from ISCCP B1 geostationary satellite observations. *J. Appl. Remote Sens.*, **1**, 013505, doi:10.1117/1.2712816.

—, and M. C. Kruk, 2010: Quantifying interagency differences in tropical cyclone best-track wind speed estimates. *Mon. Wea. Rev.*, **138**, 1459–1473, doi:10.1175/2009MWR3123.1.

—, —, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, doi:10.1175/2009BAMS2755.1.

- , J. A. Knaff, C. R. Sampson, G. M. Riggio, and A. D. Schnapp, 2013: A pressure-based analysis of the historical western North Pacific tropical cyclone intensity record. *Mon. Wea. Rev.*, **141**, 2611–2631, doi:10.1175/MWR-D-12-00323.1.
- Kossin, J. P., T. L. Olander, and K. R. Knapp, 2013: Trend analysis with a new global record of tropical cyclone intensity. *J. Climate*, **26**, 9960–9976, doi:10.1175/JCLI-D-13-00262.1.
- Martin, J. D., and W. M. Gray, 1993: Tropical cyclone observation and forecasting with and without aircraft reconnaissance. *Wea. Forecasting*, **8**, 519–532, doi:10.1175/1520-0434(1993)008<0519:TCOAFW>2.0.CO;2.
- Nakazawa, T., and S. Hoshino, 2009: Intercomparison of Dvorak parameters in the tropical cyclone datasets over the western North Pacific. *SOLA*, **5**, 33–36, doi:10.2151/sola.2009-009.
- Olander, T. L., and C. S. Velden, 2007: The Advanced Dvorak Technique: Continued development of an objective scheme to estimate tropical cyclone intensity using geostationary infrared satellite imagery. *Wea. Forecasting*, **22**, 287–298, doi:10.1175/WAF975.1.
- , and —, 2012: The current status of the UW-CIMSS Advanced Dvorak Technique (ADT). *32nd Conf. on Hurricanes and Tropical Meteorology*, Ponte Vedra Beach, FL, Amer. Meteor. Soc., P75. [Available online at <https://ams.confex.com/ams/32Hurr/webprogram/Paper292775.html>.]
- Raykar, V. C., S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, 2009: Supervised learning from multiple experts: Whom to trust when everyone lies a bit. *Proc. 26th Annual Int. Conf. on Machine Learning*, New York, NY, Association for Computing Machinery, 889–896.
- , —, —, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, 2010: Learning from crowds. *J. Mach. Learn. Res.*, **11**, 1297–1322.
- Ren, F., J. Liang, G. Wu, W. Dong, and X. Yang, 2011: Reliability analysis of climate change of tropical cyclone activity over the western North Pacific. *J. Climate*, **24**, 5887–5898, doi:10.1175/2011JCLI3996.1.
- Sampson, C. R., and A. J. Schrader, 2000: The Automated Tropical Cyclone Forecasting System (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, doi:10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2.
- Schreck, C. J., K. R. Knapp, and J. P. Kossin, 2014: The impact of best track discrepancies on global tropical cyclone climatologies using IBTrACS. *Mon. Wea. Rev.*, **142**, 3881–3899, doi:10.1175/MWR-D-14-00021.1.
- Velden, C., and Coauthors, 2006a: Supplement to: The Dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bull. Amer. Meteor. Soc.*, **87** (Suppl.), S6–S9, doi:10.1175/BAMS-87-9-Velden.
- , and Coauthors, 2006b: The Dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bull. Amer. Meteor. Soc.*, **87**, 1195–1210, doi:10.1175/BAMS-87-9-1195.
- Vigh, J. L., J. A. Knaff, and W. H. Schubert, 2012: A climatology of hurricane eye formation. *Mon. Wea. Rev.*, **140**, 1405–1426, doi:10.1175/MWR-D-11-00108.1.
- Webster, P. J., G. J. Holland, J. A. Curry, and H.-R. Chang, 2005: Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, **309**, 1844–1846, doi:10.1126/science.1116448.
- Wu, M. C., K. H. Yeung, and W. L. Chang, 2006: Trends in western North Pacific tropical cyclone intensity. *Eos, Trans. Amer. Geophys. Union*, **87**, 537–539, doi:10.1029/2006EO480001.