A Method for Identifying Midlatitude Mesoscale Convective Systems in Radar Mosaics. Part I: Segmentation and Classification

ALEX M. HABERLIE[a] AND WALKER S. ASHLEY

Department of Geographic and Atmospheric Sciences, Northern Illinois University, DeKalb, Illinois

ABSTRACT

This research evaluates the ability of image-processing and select machine-learning algorithms to identify midlatitude mesoscale convective systems (MCSs) in radar-reflectivity images for the conterminous United States. The process used in this study is composed of two parts: segmentation and classification. Segmentation is performed by identifying contiguous or semicontiguous regions of deep, moist convection that are organized on a horizontal scale of at least 100 km. The second part, classification, is performed by first compiling a database of thousands of precipitation clusters and then subjectively assigning each sample one of the following labels: 1) midlatitude MCS, 2) unorganized convective cluster, 3) tropical system, 4) synoptic system, or 5) ground clutter and/or noise. The attributes of each sample, along with their assigned label, are used to train three machine-learning algorithms: random forest, gradient boosting, and "XGBoost." Results using a testing dataset suggest that the algorithms can distinguish between MCS and non-MCS samples with a high probability of detection and low probability of false detection. Further, the trained algorithm predictions are well calibrated, allowing reliable probabilistic classification. The utility of this two-step procedure is illustrated by generating spatial frequency maps of automatically identified precipitation clusters that are stratified by using various reflectivity and probabilistic prediction thresholds. These results suggest that machine learning can add value by limiting the amount of false-positive (non-MCS) samples that are not removed by segmentation alone.

1. Introduction

Midlatitude mesoscale convective systems (MCSs)—an aggregation of deep, moist convection (DMC) organized on a scale larger than individual updrafts—are a fundamental component of the conterminous United States (CONUS) hydroclimate (Zipser 1982; Ashley et al. 2003; Houze 2004). In addition, MCSs can produce (or contribute to) atmospheric and hydrological hazards, such as damaging winds, tornadoes, and flash flooding (Fritsch and Forbes 2001; Ashley and Ashley 2008). Because of the influence of these events on many aspects of society and the increasing availability and temporal length of remotely sensed datasets, MCSs have been the focus of intense study over the last half-century (Houze 2004).

An important aspect of many of these studies is the ability to detect MCSs in remotely sensed imagery. MCSs are identified by noting cloud or precipitation clusters that meet certain size, intensity, and duration thresholds (Houze 2004; Table 1). For example, a widely used, dynamically motivated, definition proposed by Parker and Johnson (2000, hereinafter PJ00) describes MCSs as long-lasting (≥ 3 h) precipitation clusters that contain a contiguous or semicontiguous region of DMC with a major axis length that is greater than or equal to 100 km (herein the objective definition of an MCS). In addition, linear MCSs—and, in a similar way, nonlinear MCSs (Lombardo and Colle 2010)—must also show evidence of organization that matches the current understanding of the internal dynamics of these systems (PJ00; Houze 2004; herein the subjective definition of an MCS). Identification of radar-derived organizational patterns that are commonly affiliated with MCSs has largely been accomplished through subjective analysis (Gallus et al. 2008; Mulder and Schultz 2015; Corfidi et al. 2016; Miller and Mote 2017). This approach limits the amount of data that can be processed in a feasible amount of time (Lakshmanan and Smith 2009) and depends on pattern recognition that is "open to [the] judgement" of those performing the manual analysis on a case-by-case basis (Corfidi et al. 2016). As an

---

[a] Current affiliation: Department of Geography and Anthropology, Louisiana State University, Baton Rouge, Louisiana.

Corresponding author: Alex M. Haberlie, ahaberlie1@lsu.edu

TABLE 1. Selection of MCS definitions on the basis of spatiotemporal radar-reflectivity attributes.

Reference	Radar data type	Reflectivity threshold (dBZ)	Min length (km)	Min duration (h)
Hilgendorf and Johnson (1998)	Level-II mosaic	30	100	—
Geerts (1998)	Composite	20	100	4
Parker and Johnson (2000)	Composite	40	100	3
Schumacher and Johnson (2005)	Composite	40	100	3
Cohen et al. (2007)	Composite	—	100	5
Gallus et al. (2008)	Composite	30	75	2
Hane et al. (2008)	Level-II mosaic	20	100	3
Hocker and Basara (2008)	Level-II mosaic	40	50	0.5
Coniglio et al. (2010)	Composite	35	100	5
Lombardo and Colle (2010)	Composite	35	50	0.5

alternative, image segmentation and storm tracking (Lakshmanan and Smith 2010) can be used to identify spatially contiguous regions of specific ranges of instantaneous precipitation rates (herein regions of precipitation) that meet the objective definition of an MCS. This approach alone does not test for the subjective definition of an MCS, however. To test whether events that meet the objective definition of an MCS also meet the subjective definition of an MCS in an automated way, one can use supervised machine learning (Theodoridis and Koutroumbas 2003). Supervised machine learning has been utilized to classify convective organization automatically at a reasonably high accuracy using features derived from hundreds of manually labeled examples (Baldwin et al. 2005; Gagne et al. 2009). This approach results in more predictable and repeatable classification decisions while also significantly reducing analysis time. The ultimate goal of any supervised machine-learning algorithm used to identify MCS events is to accurately discriminate between MCS events that meet the objective and subjective definition of an MCS and those events that only meet the objective definition of an MCS (herein non-MCSs).

This paper describes part of a framework that is used to identify MCSs—specifically those that occur in the midlatitudes—in sequences of mosaicked composite radar-reflectivity images. The framework includes three major parts: segmentation, classification, and tracking. This paper will focus on the segmentation and classification aspects of the framework, and a second, affiliated, paper (Haberlie and Ashley 2018, hereinafter Part II) will discuss tracking as well as give examples of applying the method with observational data. The main contributions of this paper include 1) evaluation of a machine-learning procedure for discriminating between MCSs and non-MCSs, 2) illustration of the sensitivity of spatial event occurrence to image-segmentation parameters, in particular when identifying MCSs, and 3) spatiostatistical description of a novel, manually labeled, dataset of radar-derived

features from regions of precipitation determined to be an MCS or a non-MCS. Training and testing of select machine-learning algorithms is performed on a sample of thousands of hand-labeled precipitation clusters. This approach and the affiliated results are discussed in section 5. The segmentation and classification procedures are applied to radar images that represent cases from two warm seasons (May–September of 2015 and 2016) to perform a subjective validation (section 6). The process detailed herein is scalable and can be used to process multiple decades of reflectivity mosaics in a reasonable amount of time on a desktop computer. New data and machine-learning techniques can also be incorporated into this framework as they become available. All segmentation and classification procedures are completed using open-source packages written in the Python programming language, including SciPy (van der Walt et al. 2011), scikit-learn (Pedregosa et al. 2011), scikit-image (van der Walt et al. 2014), and XGBoost (Chen and Guestrin 2016). The data and Python code used for this paper are available online (https://github.com/ahaberlie/MCS/).

2. Background

a. Mesoscale convective systems

MCSs are organized assemblages of thunderstorms that produce distinct circulations and features at a larger scale than any individual, constituent convective cell (Zipser 1982). These systems are proficient rain producers and important drivers of energy redistribution in the atmosphere (Fritsch and Forbes 2001), producing an assortment of atmospheric hazards, including tornadoes (Trapp et al. 2005), damaging nontornadic winds (Ashley and Mote 2005), and flash floods (Doswell et al. 1996). In addition, MCSs are an important aspect of the central and eastern CONUS hydroclimate, producing a large proportion of warm-season precipitation for many areas in this region (Ashley et al. 2003; Houze 2004; Feng et al. 2016). Because of their multifaceted nature
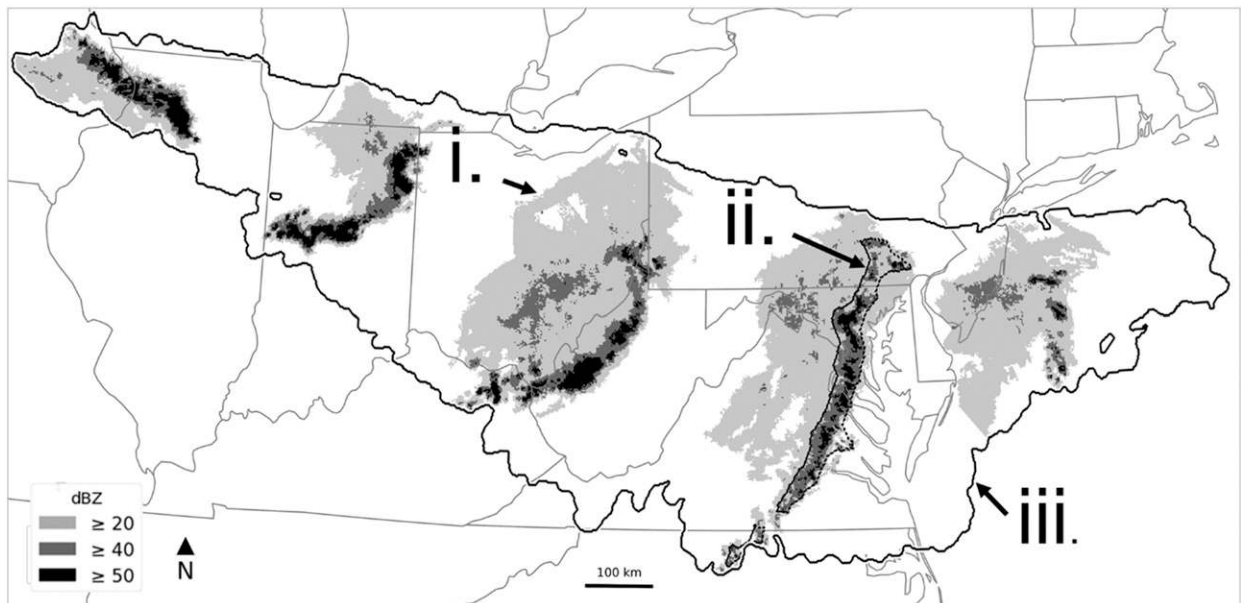
FIG. 1. Demonstration of the data generated during the segmentation and tracking process using the life cycle of a June 2012 derecho (Halverson 2014) as an example. Radar data are valid from 1500 UTC 29 Jun to 1000 UTC 30 Jun 2012. MCS slices (label i points to the middle example) are plotted every 4 h (1500, 1900, 2300, 0300, and 0700 UTC) for visualization purposes. MCS slices are made of MCS cores (e.g., the area inside the dotted black line indicated by label ii) and affiliated stratiform precipitation. MCS slices are then associated over time to generate an MCS swath (the area inside the solid black line indicated by label iii). The shading (from light to dark gray) corresponds to the following reflectivity intensities, respectively: stratiform ($\geq$20 dBZ), convective ($\geq$40 dBZ), and intense ($\geq$50 dBZ).

and meteorological and climatological importance, these events have been (and continue to be) motivating factors for several CONUS-based field projects over recent decades [e.g., PRE-STORM (Cunning 1986), IHOP_2002 (Weckwerth et al. 2004), BAMEX (Davis et al. 2004), the Mesoscale Predictability Experiment (MPEX; Weisman et al. 2015), and the Plains Elevated Convection at Night field project (PECAN; Geerts et al. 2017)].

b. Detecting mesoscale convective systems

MCSs are typically observed, identified, and tracked using radar or satellite imagery (Fritsch and Forbes 2001; Houze 2004). Two kinds of data are generated during an MCS tracking process (e.g., Fiolleau and Roca 2013): 1) "slices" from single radar or satellite images and 2) "swaths" that connect slices through sequences of images (Fig. 1). This paper focuses on the detection of slices in composite reflectivity images, whereas Part II will focus on the generation of swaths. Slices (an example is given by label i in Fig. 1), in the context of this study, are radar-derived objects that represent a contiguous region of instantaneous precipitation. For a slice to be considered a "candidate MCS slice," it must contain an "MCS core" (an example is given by label ii in Fig. 1) that meets intensity and size requirements (e.g., PJ00 criteria). An MCS core is a contiguous or nearly contiguous line or area of convection ($\geq$40 dBZ) that

may or may not be surrounded by affiliated stratiform ($\geq$20 dBZ) precipitation. The existence of an MCS core is an important way to distinguish candidate MCS slices from other radar-derived objects [e.g., mesoscale precipitation features (Rickenbach et al. 2015) or banded-precipitation features (Fairman et al. 2016, 2017)], and these features are noted ubiquitously in schematic and radar-derived examples of MCS slices (e.g., PJ00; Gallus et al. 2008; Lombardo and Colle 2010). Slices that contain qualifying MCS cores can then be organized into swaths (label iii in Fig. 1). A candidate MCS slice is considered to be an MCS slice when it is associated with a swath that persists for a certain amount of time (i.e., an MCS swath).

The main goal of segmentation—the process of extracting candidate MCS slices from radar-reflectivity images—is to identify slices that are likely associated with an MCS. Although a single radar snapshot cannot determine whether a slice is a part of an MCS (PJ00), many studies have noted common sizes, intensities, and patterns of intensity that are indicative of internal mesoscale circulations associated with these events. PJ00 provides a radar-based, dynamically motivated, objective definition of MCSs—namely, that these systems contain contiguous or semicontiguous regions of convective precipitation at least 100 km in length and lasting for 3 h. These temporal ($\tau = LU^{-1}$) and length ($L = U\tau$)
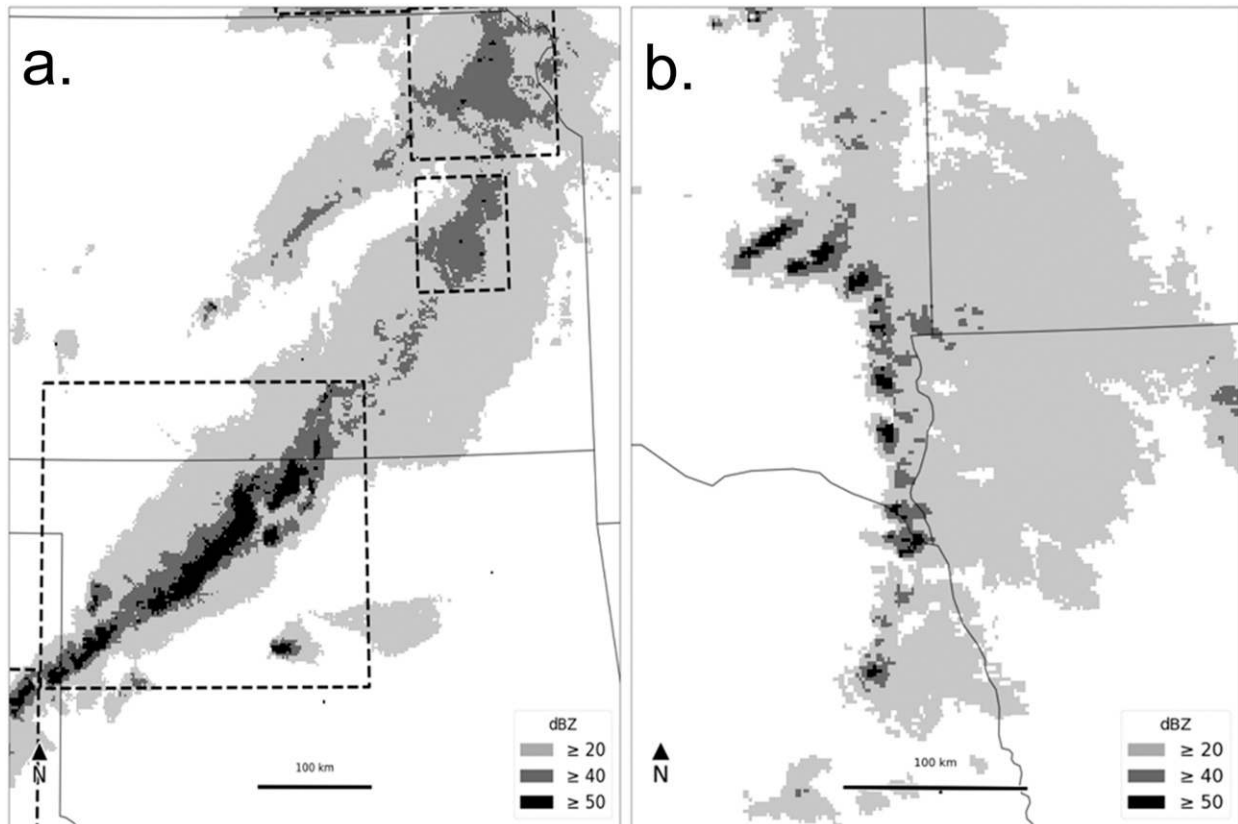
FIG. 2. Two examples [(a) 0400 UTC 2 May and (b) 1100 UTC 7 May 1997] from the literature (Figs. 7c and 6c, respectively, of PJ00) that demonstrate the results of detecting regions of connected pixels with lengths of greater than 100 km when using a convection (≥40 dBZ) threshold (dashed boxes). The dashed lines denote the bounding box of contiguous regions that have a major axis length of at least 100 km.

scales are based on 1) a Rossby number Ro that is indicative of a balance between inertial and Coriolis accelerations (Ro ≈ 1), 2) the characteristic midlatitude Coriolis force f ($f = 10^{-4}$ s^{-1}), and 3) the representative translational velocity U of MCSs and their affiliated cold pools ($U = 10$ m s^{-1}). This definition forms the basis of our radar-based MCS identification process.

The segmentation approach (thresholds, algorithms, etc.) can have a substantial impact on the results of a study and is generally related to the phenomenon of interest (Lakshmanan and Smith 2009). In the case of candidate MCS slice identification, segmentation errors can result in 1) too many candidate MCS slices, 2) missed candidate MCS slices, or 3) incorrect merging of multiple candidate MCS slices. To illustrate the first case, consider an MCS slice from 0400 UTC 2 May 1997, extending from northwestern Missouri to western Oklahoma (see Fig. 7c of PJ00). Although the most intense portion of the MCS slice (with reflectivity values greater than 50 dBZ) is in northwestern Oklahoma, two distinct areas of convective precipitation with major axes greater than 100 km exist to the north and east in northeastern Kansas. On the basis of the objective PJ00 criteria, this single MCS slice would be broken up into three MCS slices (Fig. 2a). A second example of a potential segmentation issue can be illustrated using an MCS slice from 1100 UTC on 7 May 1997 (see Fig. 6c of PJ00). Although the broken group of convective cells formed a line with a length that is greater than 100 km, the individual convective cells are separated by regions of reflectivity of less than 40 dBZ. Thus, even using a stratiform threshold to combine qualifying convective areas would fail when just using a 40-dBZ threshold, despite this event being a legitimate MCS slice (Fig. 2b). Further, using a stratiform shield to combine qualifying regions of convection can sometimes result in the merging of multiple unique candidate MCS slices. For example, when viewing the MCS slice depicted in Fig. 2b from a regional perspective (Fig. 3), it is evident that it is within a larger region of precipitation of at least stratiform intensity extending from South Dakota south and eastward into Nebraska and Iowa (Fig. 3). Although there are at least two unique MCS cores (Fig. 3b) and affiliated stratiform regions (Fig. 3c), using a stratiform
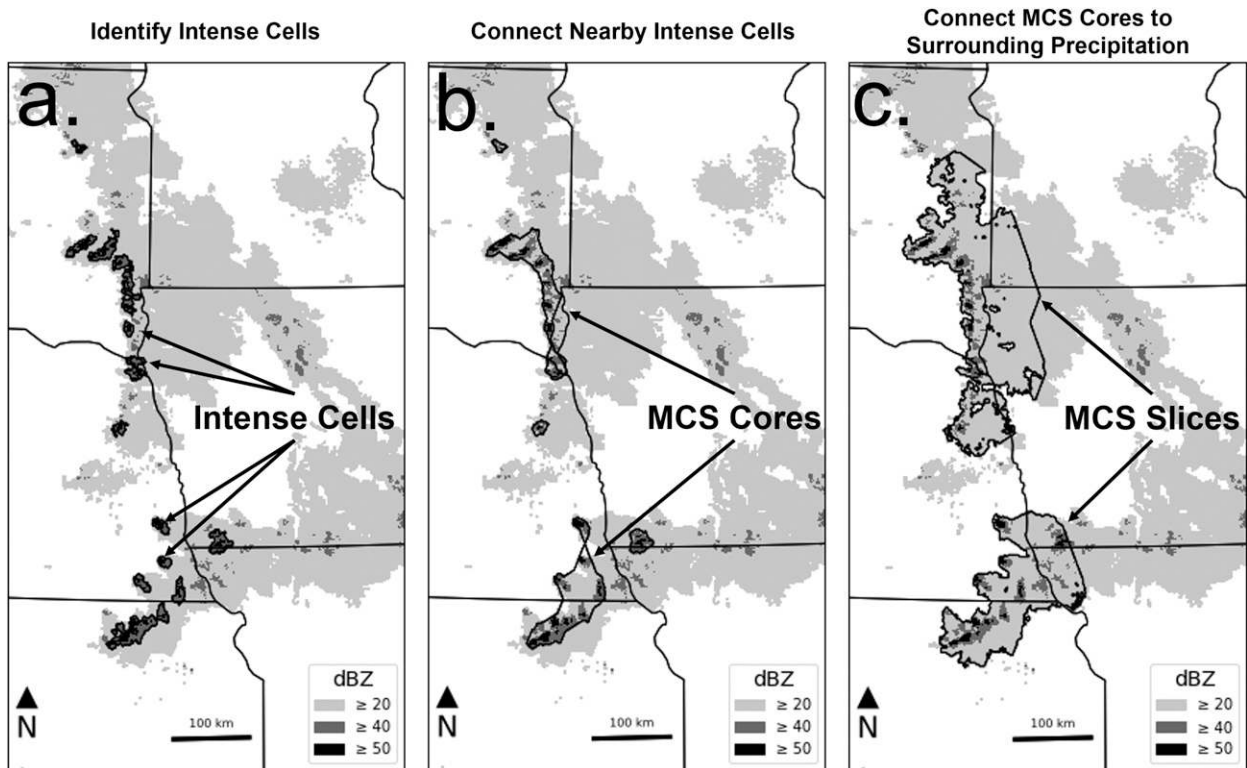
FIG. 3. Demonstration of segmentation steps for candidate MCS slices using reflectivity valid at 1100 UTC 7 May 1997 (cf. Fig. 6c in PJ00): (a) Convective ($\geq$40 dBZ) cells with regions of intense convection ($\geq$50 dBZ) and areas greater than 40 km^2 are extracted (black outlined regions). (b) These cells are then connected if they are within a specified radius (12 km; black-outlined regions). If a connected region has a major axis length of at least 100 km, they are considered to be candidate MCS cores. (c) Stratiform ($\geq$20 dBZ) pixels within a specified radius (96 km) are then associated with their respective cores, and the resulting candidate MCS slices are delineated by the black-outlined areas.

threshold to aggregate qualifying areas of convection would result in one large candidate MCS slice. These issues motivated the development of a more sophisticated segmentation approach that will be discussed in section 4.

c. Classification of convective clusters using supervised machine learning

The process of developing knowledge through observations and then applying it to new, unseen data is called pattern recognition (Theodoridis and Koutroumbas 2003). Although humans excel at pattern recognition, large datasets can make laborious manual classification impractical (Theodoridis and Koutroumbas 2003). Two approaches can be used to automate this task: 1) expert systems and 2) supervised machine learning. Expert systems use domain knowledge to define criteria for classification decisions. An example of an expert system would be an algorithm that labels precipitation clusters using specific size and intensity thresholds (e.g., the objective definition of an MCS). Alternatively, supervised machine learning can generate a generalized model of classification decisions on the basis of the statistical properties of

sample data. Rather than a set of rules, an expert provides training data with categorical labels (e.g., the subjective definition of an MCS), and a machine-learning algorithm develops a way to determine how to sort those data most accurately into categories (e.g., Baldwin et al. 2005; Gagne et al. 2009; McGovern et al. 2017). After this training is complete, a well-performing model would be able to ingest previously unseen data and assign correct labels at rates comparable to, or, possibly better than, those of humans.

Supervised machine learning is used for this study to address the issue of mislabeling non-MCS regions of precipitation as candidate MCS slices (i.e., false positives) when only using the objective definition of an MCS. We have identified three common false-positive classes: unorganized clusters or lines of convective cells (UCC), tropical systems ("tropical"), and synoptic precipitation systems ("synoptic"). In addition, "ground clutter" ("clutter") is included as a class because of its ubiquity in the dataset and will herein be considered as a non-MCS class. Ground clutter is a phenomenon that is not typically associated with precipitation but sometimes appears as a ring of stratiform-intensity pixels with
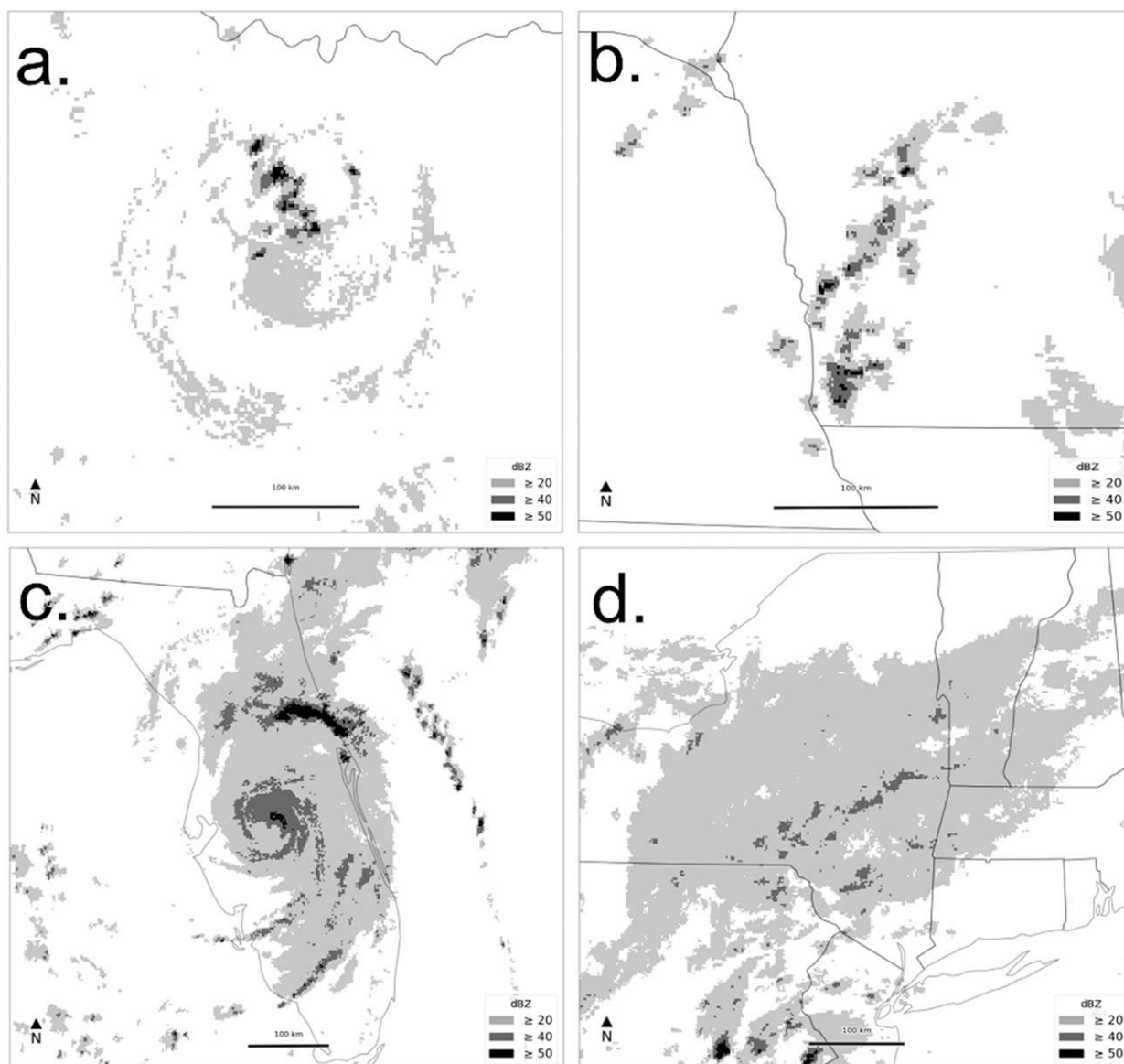
FIG. 4. Examples of non-MCS pixel regions that could qualify as MCSs per PJ00. These include (a) ground clutter near Dallas, Texas, at 2230 UTC 14 Jul 2000, (b) UCCs over western Iowa at 2305 UTC 10 Jun 2005, (c) Hurricane Charley over central Florida at 2350 UTC 13 Aug 2004, and (d) a region of synoptic-scale rainfall over central New York at 1915 UTC 27 Jul 2004. The horizontal line in each image represents 100 km. Intensity is denoted as in Fig. 1.

embedded convective-intensity pixels (Fig. 4a). Precipitating examples of false positives include broken, unorganized, clusters or lines of convective cells [Fig. 4b; see also Fig. 2 of Gallus et al. (2008) and Fig. 2f of Lombardo and Colle (2010)], tropical systems [Fig. 4c; see also Fig. 2 of Matyas (2009)], and synoptic regions of precipitation (Fig. 4d). For true positives (i.e., valid MCS slices), studies like PJ00 and Lombardo and Colle (2010) provide several illustrations and descriptions of various kinds of MCS slice morphologies (parallel, leading, trailing, no stratiform, areal, etc.). Our goal is

not to propose a specific radar-based definition for each of these classes but rather to determine whether the identified events exhibit spatiostatistical attributes that are meteorologically and climatologically reasonable (see section 6).

Although the various false-positive and true-positive examples are visually discernable, manually generating a set of rules to separate the two groups would be difficult (e.g., Fig. 6 of Gagne et al. 2009; McGovern et al. 2017). Supervised machine learning has been used successfully to automate similar problems (McGovern et al. 2017). For

example, Gagne et al. (2009) used several machine-learning algorithms to differentiate among six types of convective precipitation regions. Their approach was to perform hand classification on a set of ~900 images by using a graphical interface that presented various storm attributes (reflectivity, size, shape, etc.). Their manual classification decisions were based on knowledge of convective cluster hierarchies, such as those presented in PJ00 and Gallus et al. (2008). Once the storms were identified, feature extraction was performed to calculate various attributes of the hand-labeled storms (see Table 1 of Gagne et al. 2009). These attributes were then used to train machine-learning algorithms. The best-performing algorithm was the one known as random forest (RF; Breiman 2001), followed closely by various bagging and boosting approaches (see Fig. 8 of Gagne et al. 2009). These algorithms are still popular in many fields (e.g., "XGBoost"; Chen and Guestrin 2016), including the atmospheric sciences (Gagne et al. 2014, 2017; Ahijevych et al. 2016; McGovern et al. 2017). The RF algorithm is an ensemble of fully grown decision trees generated using random split features (Breiman 2001). Alternatively, gradient boosting (GB) is an ensemble of shallow decision trees that are iteratively generated to address deficiencies in previously generated decision-tree predictions (Gagne et al. 2017). In addition, this study uses the XGBoost algorithm (Chen and Guestrin 2016). XGBoost is an extension of GB that uses additional methods that reduce model overfitting. These algorithms can rival the performance of more complex algorithms (e.g., deep neural networks; Krizhevsky et al. 2012) in machine-learning competitions (Chen and Guestrin 2016), with less time spent on tuning model hyperparameters. Gagne et al. (2017, their section 2.4) provide a detailed summary of these algorithms and their application in the atmospheric sciences. This study will focus on the application of RF, GB, and XGBoost to address the false-positive problem, and the approach is described in detail in section 5.

3. Data

a. Radar data

Virtually continuous archiving of remotely sensed precipitation patterns has been ongoing within the CONUS for over 20 years. The associated data archive, a stated goal of the Next Generation Weather Radar program (NEXRAD; Crum et al. 1993), has been increasingly leveraged for climatological studies (Matyas 2010). Reflectivity data (a proxy for instantaneous precipitation intensity) from this archive have greater spatial coverage and temporal resolution than do rain gauge observations (Brooks and Stensrud 2000) and have been used in many

studies to identify the occurrence of atmospheric phenomena. These data are useful for this study because MCSs are generally identified by visual patterns in radar-reflectivity data (see section 2).

National Operational Weather Radar (NOWrad), generated and quality controlled by Weather Services International (now The Weather Company), will be utilized for the purposes of developing and validating the MCS-tracking procedure. NOWrad is a national composite reflectivity radar mosaic product that has a horizontal resolution of approximately 2 km (see Fabry et al. 2017). For the purposes of this study, distance and area calculations are simplified by defining a pixel length as 2 km and a pixel area as $4 \, \text{km}^2$. NOWrad grid (1837 by 3661 pixels) values are calculated by gathering reflectivity data approximately every 5 min from CONUS NEXRAD stations. Because the product is generated from NEXRAD data, most of the caveats associated with composite reflectivity data apply (Smith et al. 1996), with some exceptions. Range-dependent biases are addressed by considering composite data from all radars within 230 km when setting pixel values (Parker and Knievel 2005). Anomalous propagation and false echoes are removed through an automated quality-control procedure (Carbone et al. 2002). These techniques are effective in producing spatially contiguous composite reflectivity fields at both 5- and 15-min intervals, especially in areas with sufficient radar overlap (i.e., much of the central and eastern CONUS), and these data have been used in many climatological studies (e.g., Parker and Knievel 2005; Parker and Ahijevych 2007; Carbone and Tuttle 2008; Tuttle and Carbone 2011; Ashley et al. 2012; Haberlie et al. 2015, 2016; Fabry et al. 2017).

b. Study area and study period

Qualifying slices for May–September in 2015 and 2016 are extracted from composite reflectivity images to assess subjectively the performance of the segmentation and classification procedure. These periods experienced above-average precipitation in the Midwest and Great Plains (McEnery et al. 2005), suggesting that MCS activity was also above average (Houze 2004). Further, the 2015 warm season coincides with the PECAN field project (Geerts et al. 2017), allowing the opportunity for additional verification against an external dataset of events that were identified by researchers with MCS expertise. The machine-learning model was trained and tested with over 3600 labeled objects that were manually extracted from an 11-yr period of NOWrad images (2003–13).

4. Segmentation

Slices are identified in radar-reflectivity images by searching for groups of connected pixels that meet size

and intensity criteria. Three ranges of reflectivity threshold values are used to perform this task, with ad hoc labels of stratiform, convection/convective, and intense (e.g., Fig. 4 in PJ00; Table 2). Further, the term "convective region" is used in this study to refer to an area of pixels that meet or exceed the convective reflectivity threshold, with one or more of those pixels meeting or exceeding the intense threshold. A minimum areal constraint is placed on convective regions to reduce the effects of nonmeteorological radar echoes (i.e., "noise"; Lack and Fox 2012; Lakshmanan et al. 2013) and corresponds to the minimum size of a Byers–Braham cell (Byers and Braham 1948; Miller and Mote 2017). To be considered a candidate MCS slice for this study, the major axis length of the convective region must exceed a certain value (Table 2). If these requirements are met, stratiform precipitation within a specified radius (Table 2) is combined with the qualifying convective region to create a candidate MCS slice object. In the case of broken lines and unconnected cells within a specified distance of one another (e.g., Baldwin et al. 2005; Pinto et al. 2015; Chang et al. 2016; Table 2), temporary connections are made to perform size and intensity qualification checks, and, if necessary, stratiform associations. The decision to connect nearby convective regions, even if there is no convective precipitation "bridge," is based on various examples and conceptual discussions of MCSs that have been provided in the literature [e.g., Figs. 5 and 6 in PJ00 and the discussion in Doswell (2001, p. 8)]. In specific terms, it is thought that DMC updrafts (Byers–Braham cells) should belong to the same MCS if they are close enough to interact. The segmentation process ultimately results in a varying number of unique pixel groups that meet the aforementioned segmentation criteria for each radar image. This definition will encompass the variety of MCS subclassifications found in the literature (e.g., Bluestein and Jain 1985; Parker and Johnson 2000; Lombardo and Colle 2010).

Candidate MCS slices for each NOWrad image are extracted by identifying qualifying convective regions and then associating them with nearby stratiform regions (Fig. 3). First, a binary mask representing the pixel coordinates of the initial convective regions is generated by identifying unique groups of contiguous pixels that contain at least a specified number of convective pixels and contain one or more intense pixels (Fig. 3a). To connect convective regions within a specified distance of one another, a binary closing is applied to the resulting binary mask to form temporary agglomerations of these regions. In contrast to a binary dilation, which indiscriminately extends the binary mask in all directions, a binary closing forms bridges between nearby regions without modifying the size of isolated regions (Fig. 3b).

TABLE 2. Various threshold values used in this study to segment MCSs in composite reflectivity images.

Threshold	Value
Stratiform (dBZ)	20
Convection (dBZ)	40
Intense (dBZ)	50
Convective-region area (km^2)	40
MCS core length (km)	100
Convective-region search radius (km)	6, 12, 24, and 48
Stratiform search radius (km)	48, 96, and 192

The former approach would result in a larger number of small convective regions that meet the size requirement, since their major axis length would be artificially inflated. Next, all stratiform pixels within a specified distance of pixels associated with a qualifying convective region (e.g., label ii in Fig. 1) are identified to generate a second binary mask (Fig. 3c). Last, the corresponding intensity information for each unique group of contiguous pixels in this mask is extracted to generate an MCS slice candidate (e.g., Fig. 1i).

5. Classification

a. Overview

Classification decisions are determined by training selected classifiers using features from thousands of manually labeled intensity images centered on unique slices. Manual classification is performed using subjective judgment (e.g., Gagne et al. 2009; Lack and Fox 2012) to organize the samples into the five distinct categories that are described in section 2c: 1) midlatitude "MCS," 2) nonmeteorological reflectivity objects, or clutter (Fig. 4a), 3) UCCs (Fig. 4b), 4) tropical systems (Fig. 4c), and 5) synoptic systems (Fig. 4d). The choice of these categories is based on observations of false-positive associations using the PJ00 objective definition and previous work that differentiated between precipitation areas associated with MCSs and non-MCSs (e.g., Schumacher and Johnson 2005; Kunkel et al. 2012).

From this labeled population, the samples are organized into two groups of years: 1) the training data (2006–13) and 2) the testing data (2003–05). The training data are used to generate the machine-learning models, whereas the testing data are used to simulate the performance of these models on previously unseen, independent, data. Metrics designed to describe relationships between false positives/negatives and true positives/negatives are used to assess model performance (Gagne et al. 2009; Lakshmanan and Smith 2010; Kolodziej Hobson et al. 2012). The classifiers are tuned for optimal performance using subjective and objective hyperparameter selection

TABLE 3. A list of the features extracted from intensity-image samples and the mean and standard deviation of their values. Importance is derived from the ''best'' binary classifiers discussed in section 5c. The features are sorted from highest to lowest by mean importance values. The abbreviations for best-performing machine-learning classifiers in the importance column are defined in section 5c.

Feature	Definition	Mean	Std dev	Importance		
				RFC	GBC	XGBC
Area	Total area (km^2)	54 131	74 355	0.15	0.16	0.16
Solidity	Ratio of area to convex area	0.50	0.14	0.03	0.15	0.12
Intense area	Area (km^2) covered by pixels exceeding the intense threshold	1414	2419	0.15	0.06	0.05
Minor axis length	Length of the short axis of an ellipse fit to the slice (km)	184.6	140.6	0.06	0.07	0.12
Convex area	Area (km^2) covered by convex hull	120 160	180 403	0.09	0.06	0.10
Intense–stratiform ratio	Ratio of intense area to area	0.05	0.05	0.1	0.05	0.08
Convection area	Area (km^2) covered by pixels exceeding the convection threshold	8174	12 688	0.13	0.04	0.05
Convection–stratiform ratio	Ratio of convection area to area	0.19	0.13	0.04	0.06	0.07
Intensity variance	Variance of pixel intensity	63.8	32.7	0.08	0.05	0.03
Major axis length	Length of the long axis of an ellipse fit to the slice (km)	417	347.8	0.07	0.07	0.02
Mean intensity	Mean intensity of pixels	29.6	4.0	0.02	0.06	0.06
Eccentricity	Noncircularity of the slice	0.84	0.14	0.01	0.04	0.07
Intense–convection ratio	Ratio of intense area to convection area	0.17	0.14	0.04	0.04	0.03
Max intensity	Max intensity of pixels	54.5	9.2	0.02	0.02	0.04

(McGovern et al. 2017). For the purposes of this study, several features (Table 3) are derived from intensity information associated with each manually labeled sample. The selection of features (Baldwin et al. 2005, their section 2) is based on previous related research [see Table 1 in Baldwin et al. (2005), Table 1 in Gagne et al. (2009), and Table 1 in Lack and Fox (2012)] and is performed using existing image-processing functions in scikit-image.

b. Feature extraction and summary statistics

Approximately 4000 composite reflectivity images are randomly selected (without replacement) from a population of 350 000 available images from 2003 to 2013 for June–September. This period is chosen because it does not overlap with the subjective validation period (May–September 2015 and 2016). Visual examination of the selected images is then performed to find cases that fit one of the five possible classifications used in this study. Although this process is subjective, it is based on well-established taxonomies presented in the literature (PJ00; Gallus et al. 2008; Gagne et al. 2009; Lombardo and Colle 2010, etc.). When a case is found, the pixel cluster (or clusters) is circled (features a and b in Fig. 5), and a binary mask is generated by filling in the circle. The pixel coordinates of this binary mask are then used to extract intensity information from the corresponding, unmodified composite reflectivity image. Each cluster and its intensity information are then extracted and saved as an image with dimensions equal to the bounding box of

the affiliated pixel cluster. Various features (Table 3) are then extracted from each image and saved in a table with the affiliated class label. In total, 3659 cases are generated (Table 4).

After feature extraction is complete, summary statistics for each class are created. Figure 6 illustrates several selected features and the distribution of their values for each of the five classifications. MCS samples have areas that are larger than UCCs and clutter but smaller than tropical and synoptic samples (Fig. 6a). The distribution of tropical sample areas matches well with previously reported distributions (e.g., Comstock 2011; Matyas 2014). Hitchens et al. (2012) and Fiolleau and Roca (2013) suggest that MCS objects generally had areas between 10 000 and 1 000 000 km^2, which also matches well with the distribution of MCS sample areas (Fig. 6a) and the theoretical spatial-scale range for MCSs (from 100 km × 100 km to 1000 km × 1000 km; Markowski and Richardson 2011). MCS samples have higher mean and maximum intensity than all of the classes except UCC samples (Fig. 6b). The higher reflectivity variance for MCS and UCC samples is likely due to their propensity to contain strong reflectivity gradients, resulting from intense, convective cells or lines of cells surrounded by regions of less-intense, stratiform precipitation (PJ00). The range of mean reflectivity values for MCS samples agrees with previously reported values (Fritsch and Forbes 2001). Selected synoptic samples are generally large areas of stratiform with isolated regions of embedded convection (e.g., Fig. 4d), and the number of
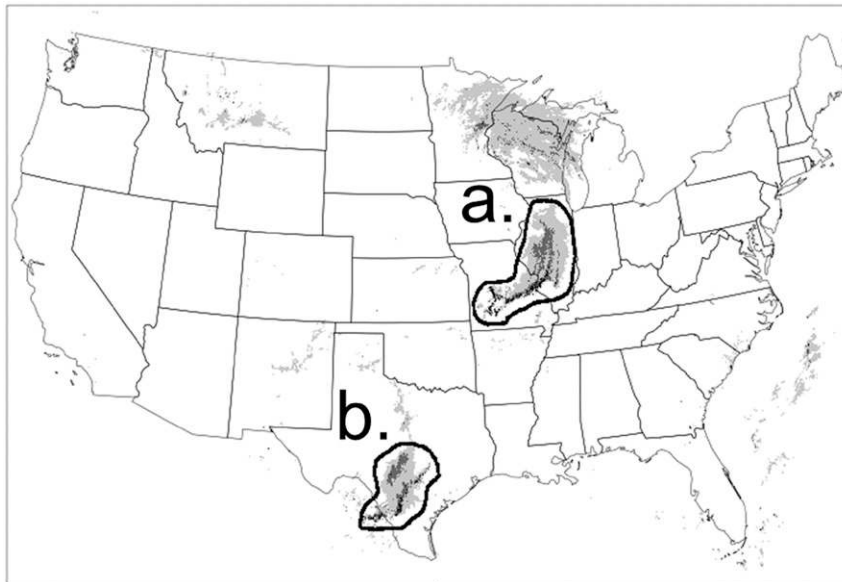
FIG. 5. An example of hand labeling MCS slices in a composite reflectivity image. The black lines are hand drawn around rainfall clusters that are determined to be MCS slices (labeled a and b). The selected composite reflectivity image is from 1205 UTC 10 Jun 2003. Intensity is denoted as in Fig. 1.

stratiform-intensity pixels reduces mean reflectivity values. MCS and UCC samples generally have a higher fraction of pixels meeting or exceeding convective or intense thresholds, relative to their total size (Fig. 6c). MCS samples also tend to be less circular than tropical and clutter samples but have similar eccentricity to that of UCCs and synoptic samples (Fig. 6d). Indeed, high values of eccentricity have been used to identify MCSs and linear systems in previous work (Jirak et al. 2003; Gagne et al. 2009). Overall, the relationships between the distributions for each class are reasonable and agree with the existing literature. Kolodziej Hobson et al. (2012) report a higher mean reflectivity value for organized convective clusters (37 dBZ) than that in this study (26 dBZ), but this is likely attributable to their choice of a 30-dBZ storm identification threshold as compared with the 20-dBZ threshold used in this study. The samples also occur in the CONUS where one would expect. MCS samples are largely gathered from the central plains and Midwest (Fig. 7a), whereas UCC samples are gathered in a higher frequency from the Southeast (Fig. 7b). Tropical samples are clustered around the Gulf and Atlantic coasts (Fig. 7c), in contrast to synoptic samples, which are largely from the northern and northeastern CONUS (Fig. 7d). Clutter samples are mostly located over radar-station locations (Fig. 7e). Overall, the most of the samples are from events that occur over the midwestern CONUS (Fig. 7f), but almost every location east of the continental divide is represented.

c. Classifier training and testing

This study explores the utility of three RF and GB classifier implementations: 1) "RandomForestClassifier" (RFC; scikit-learn 0.18), 2) "GradientBoostingClassifier" (GBC; scikit-learn 0.18), and 3) "XGBClassifier" (XGBC; xgboost-python 0.6). The default settings for each implementation are used unless otherwise specified (http://scikit-learn.org; http://xgboost.readthedocs.io). After feature extraction is completed (see section 5b), the dataset is split into two groups: training and testing data (Table 4).

TABLE 4. Training and testing counts by classification and year.

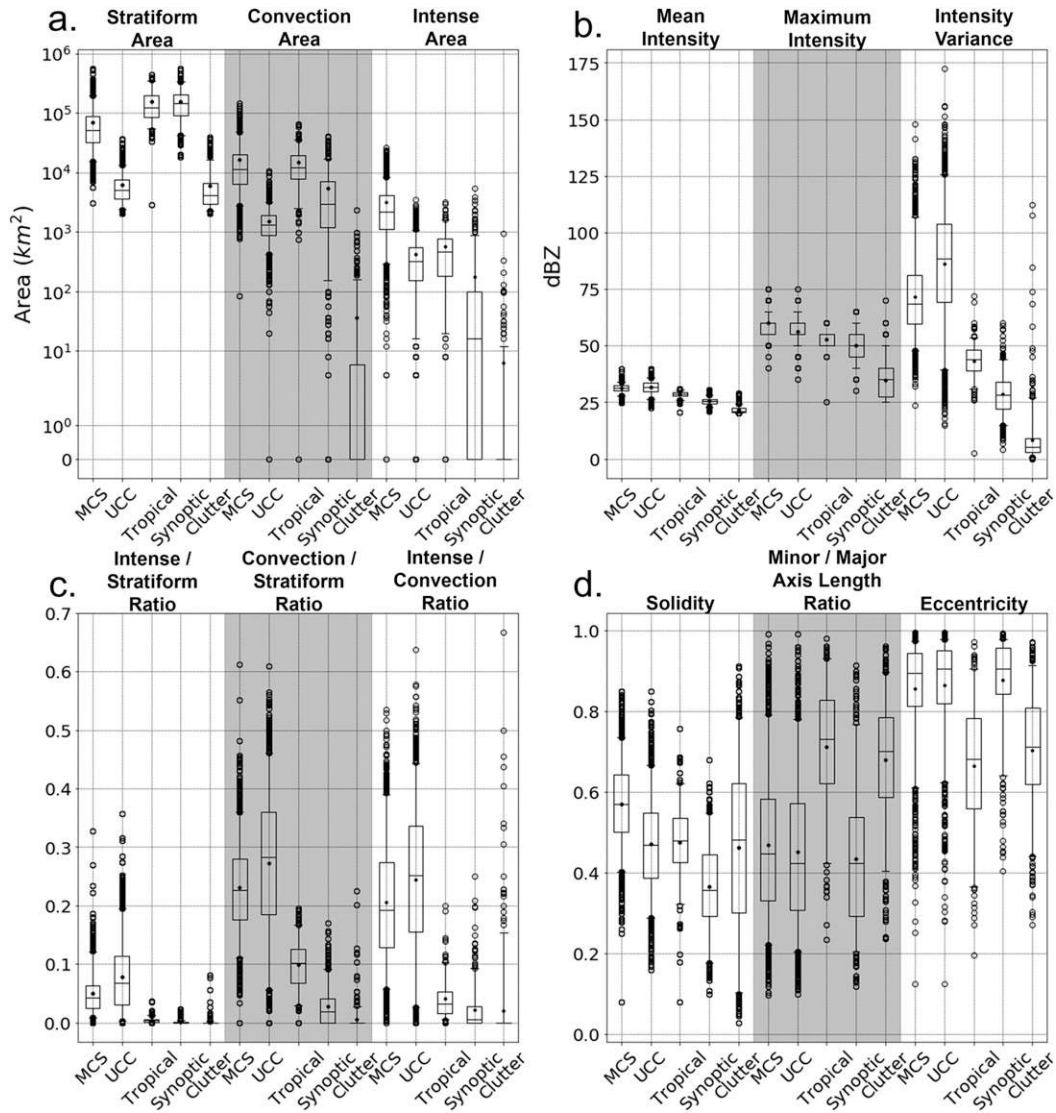|	MCS	UCC	Tropical	Synoptic	Clutter
Testing					
2003	161	208	10	56	38
2004	249	231	37	59	35
2005	168	227	43	44	48
Total	578	666	90	159	121
Training					
2006	97	113	12	61	108
2007	121	96	5	22	112
2008	136	75	56	25	12
2009	110	63	0	50	12
2010	110	92	7	31	7
2011	116	63	7	12	14
2012	74	53	7	14	10
2013	75	49	1	10	7
Total	839	604	95	225	282
Overall	1417	1270	185	384	403

FIG. 6. Comparison of distributions of select features related to (a) area, (b) intensity, (c) relationships between areas of different intensity, and (d) shape metrics for hand-labeled samples from each of the five classifications. All *y* axes use a linear scale except (a), which uses a logarithmic scale. The feature names for each group are located on top of the alternatively shaded areas. The box represents the interquartile range, with the black horizontal line denoting the distribution's median and the black dot denoting the mean. The whiskers represent values between the 5th and 95th percentiles, and the open circles are outliers.

The purpose of this step is 1) to use the training data to generate the classifiers and 2) to use the testing data to determine how well the predictions of the classifiers will generalize to previously unseen data. Class-specific metrics are used to assess the performance of the classifiers because of the unbalanced nature of the dataset. To be specific, the counts for the various labels range from 185 (~5% of the dataset) to 1417 (~39% of the dataset). These metrics are calculated by first reducing the multiclass labels to binary labels (e.g., MCS or non-MCS) and generating a confusion matrix (Table 5).

Then, two performance metrics—probability of detection and probability of false detection—are calculated for each class (Table 6). These metrics assess potential weaknesses in the model that may be missed by reporting prediction accuracy alone, especially for unbalanced datasets (Zheng 2015).

Automatic model tuning is performed by using a grid search with cross validation on training data to explore the parameter space (Elith et al. 2008). The approach builds multiple classifiers using one or more parameters selected from a user-defined range. Each classifier with
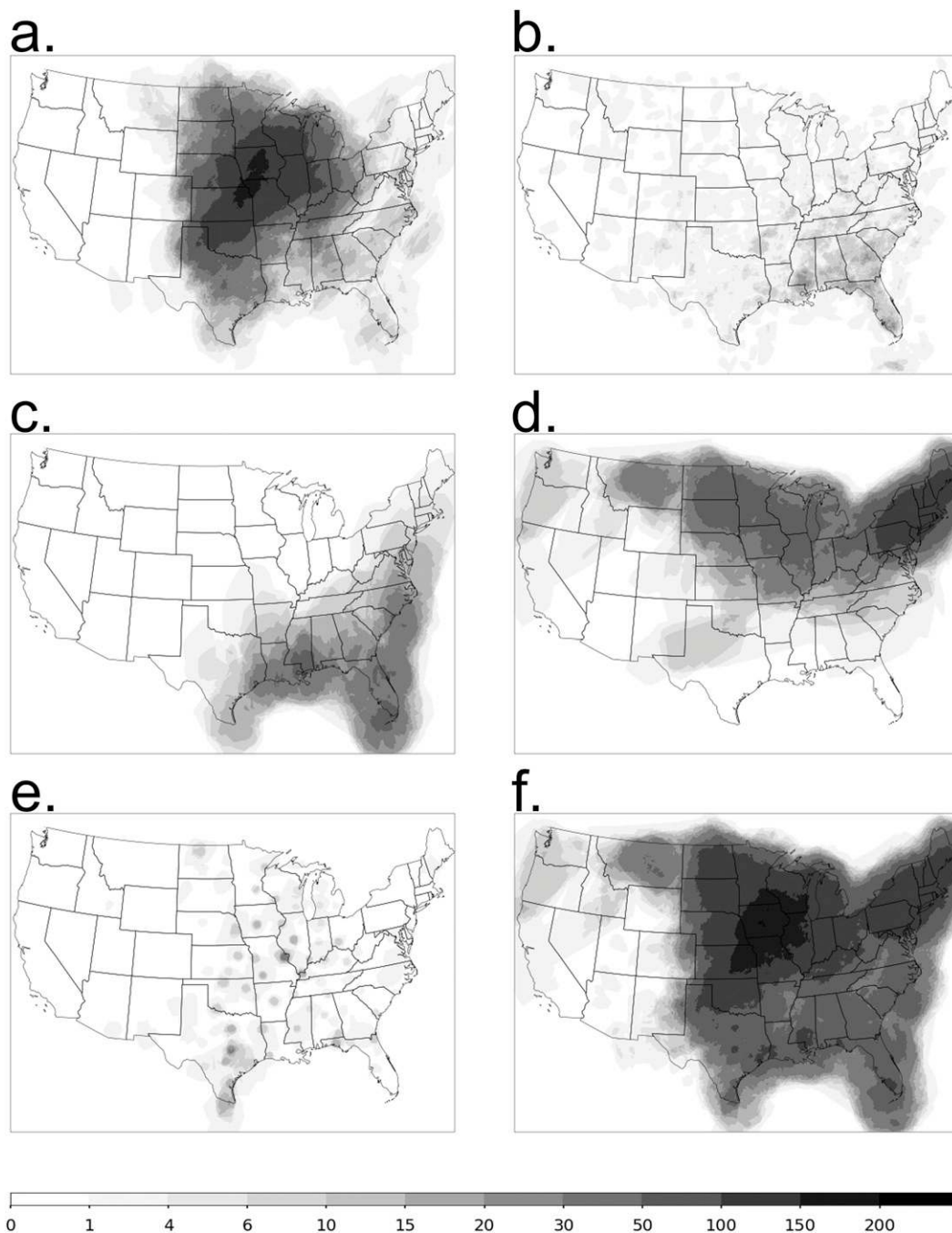
FIG. 7. The spatial occurrence of hand-labeled samples gathered from randomly selected composite reflectivity images for June–September from 2003 to 2013, corresponding to the following labels: (a) MCS, (b) UCC, (c) tropical, (d) synoptic, (e) clutter, and (f) all samples.

the selected parameter value(s) is trained with a portion of the training data missing. In specific terms, this study employs a "leave one year out" approach to determine which data are removed for each iteration (Westerling et al. 2002), where each step of cross validation removes samples from an entire warm season. These data are then used to test each model parameter permutation, after which they are placed back into the training data and a different year is removed. In other words, if a particular iteration of cross validation uses samples from 2009 to test model performance, the models are trained using data from 2006 to 2008 and from 2010 to 2013. This

TABLE 5. Example of a confusion matrix for predictions and actual labels for MCS cases and cases that are not MCS. [Adapted from Table 3 in Gagne et al. (2014).]

	Actual: MCS	Actual: Not MCS
Prediction: MCS	True positive (TP)	False positive (FP)
Prediction: Not MCS	False negative (FN)	True negative (TN)

is repeated until each year is removed for each version of the classifier (Elith et al. 2008), and the goal of this process is to determine which model permutation would generalize the best to data gathered from previously unseen years.

The metric used to decide the best classifier configuration is the Heidke skill score (HSS; Table 6; Wilks 2011) mean for each class averaged over the leave-one-year-out cross validation (Table 7). An average HSS of 0.91 for all three classifiers is achieved using the following model configurations: 1) an RFC that uses 100 estimators, 2) a GBC that uses 500 estimators and a learning rate of 0.01, and 3) an XGBC that uses 500 estimators and a learning rate of 0.01. These values ranged from a low of 0.83 to a high of 0.96, depending on which year was used to test the trained model for the five-class label. To assess class-specific performance, probability of detection and probability of false detection were calculated using predictions on testing data from the best classifiers (Table 8). MCS samples are successfully detected between 91% and 93% of the time, and non-MCS samples are assigned MCS labels between 5% and 9% of the time. All of the approaches struggled to classify tropical samples, with as few as 62% of those samples being assigned the correct label, despite a low occurrence (<1%) of nontropical samples being assigned a label of tropical. Of the misclassified tropical samples, 15% were assigned MCS labels and 21% were assigned synoptic labels (Tables 9 and 10). Slightly better performance was noted for UCC samples, with those samples being labeled correctly 87%–90% of the time, in contrast to non-UCC samples being labeled as UCCs

4%–5% of the time. An ensemble of these three classifiers performs no better than its best member but provides better performance than the worst member.

Using a similar grid search with a cross-validation approach, binary classifiers are generated to differentiate between MCS samples and non-MCS samples. The purpose of this step is to determine whether a binary-classification approach will result in better MCS classifications than will the five-class approach. Again, leave-one-year-out cross validation is used to find optimal binary classifiers. An average HSS of 0.93 for all three classifiers is achieved using the following model configurations: 1) an RFC that uses 100 estimators, 2) a GBC that uses 1000 estimators and a learning rate of 0.1, and 3) an XGBC that uses 1000 estimators and a learning rate of 0.01. The yearly HSS for the binary classifiers ranges from 0.89 to 0.97—an improvement over the five-class approach, in particular for the lower end of that range (Table 7). Binary-classification performance for the testing data (Table 8) produced probability of detection values for MCS similar to those of the five-class approach (91%–92%) but resulted in a reduced probability of false detection (4%–5%). For non-MCS labels, the probability of detection ranges from 95% to 96% and the probability of false detection ranges from 8% to 9%. The binary-classification approach reduces the percentage of tropical samples that were labeled as MCS from 15% to 11% while raising the number of MCS samples labeled as non-MCS from 6% to 9% (Tables 9 and 10).

In addition, the three classifiers used in this study can calculate relative feature importance because of their use of decision trees as ensemble estimators (McGovern et al. 2017). Values of relative importance are determined for each feature by calculating the mean reduction in error rate over all estimators when that feature is used for split decisions (Hastie et al. 2009). These values are reported in Table 3 for each classifier, sorted from most important to least important. The most important features were related to area—in particular,

TABLE 6. Selected model-performance metrics and associated equations. Abbreviations are as described in Table 5. For the Brier loss score, mcs_i denotes the classifier confidence/probability of an MCS prediction for a sample and $actual_i$ is the binary label for that sample.

Metric	Formula	Explanation
Probability of detection	$\dfrac{TP}{TP + FN}$	When the actual label is MCS, how often does the model predict MCS?
Probability of false detection	$\dfrac{FP}{FP + TN}$	When the actual label is not MCS, how often does the model predict MCS?
Heidke skill score	$\dfrac{2[(TP \times TN) - (FN \times FP)]}{(TP + FN)(FN + TN) - (TP + FP)(FP + TN)}$	How well does the model perform relative to the expected random agreement between actual labels and model predictions?
Brier loss score	$\dfrac{1}{N} \sum\limits_{i=1}^{N} (mcs_i - actual_i)^2$	How reliable are probabilistic MCS predictions?

TABLE 7. Heidke skill scores produced by the best-performing model configurations trained using five-class and binary labels for an RFC, a GBC, and an XGBC for each iteration of leave-one-year-out cross validation on training data (2006–13). The standard deviation (std dev) and mean HSS for each of the three models are also reported.

	2006	2007	2008	2009	2010	2011	2012	2013	Std dev	Mean
Five-class label										
RFC	0.96	0.90	0.86	0.95	0.95	0.87	0.90	0.90	0.04	0.91
GBC	0.96	0.90	0.85	0.95	0.93	0.88	0.86	0.95	0.04	0.91
XGBC	0.96	0.90	0.83	0.95	0.94	0.91	0.91	0.96	0.04	0.92
Binary label										
RFC	0.95	0.89	0.93	0.97	0.95	0.91	0.91	0.89	0.03	0.93
GBC	0.95	0.89	0.95	0.96	0.97	0.90	0.90	0.94	0.03	0.93
XGBC	0.94	0.89	0.96	0.97	0.98	0.90	0.90	0.93	0.04	0.93

the total area, convex area, and area covered by pixels with intensities exceeding intense thresholds. This is not surprising, because Fig. 6a illustrates the disparity in area among the different classes. The ratio of the area of intense pixels to the area of stratiform pixels also shows a relatively high feature importance, suggesting that the fractional region covered by intense pixels is an important discriminating factor between MCS and non-MCS samples.

Probabilistic predictions are also used to assess the performance of the classifier. In contrast to "hard" predictions—that is, a label of MCS or non-MCS—a "soft" prediction assigns a probability of every possible class for a given sample (e.g., 49% non-MCS and 51% MCS). For example, only 1% of decision trees/estimators in the best-performing RFC predict MCS as the label for the clutter sample in Fig. 4a, 3% of the decision trees predict MCS for the UCC sample in Fig. 4b, and less than 1% of the decision trees predict MCS for the synoptic sample in Fig. 4d. In contrast, for Fig. 4c, 93% of the decision trees incorrectly predicted MCS for the tropical sample. This mislabeling is likely due to Hurricane Charley's relatively small size (85 768 km^2)

and unusually large area of intense pixel values (1820 km^2) in comparison with other tropical samples.

The testing-data classification probabilities can be used to approximate not only how often a sample is mislabeled but how often a sample is mislabeled with a high (e.g., Hurricane Charley) or low probability. First, the model probabilities must be examined to determine whether they are properly calibrated using the Brier loss score (Brier 1950; Table 6). This score measures the mean square error between the predicted probability of a label and the actual label, and a lower Brier loss score suggests that the model probabilities generally match up well with the actual labels (see Fig. 8a in Gagne et al. 2012). The lowest Brier loss score (0.042) is achieved by an ensemble voting classifier comprising votes from the best RFC, GBC, and XBGC (Fig. 8b). High and low MCS probabilities are reliable, with some overconfidence noted in probabilities around 0.4 and some underconfidence noted around 0.7. Further examination of the relationship between classification confidence/probability and correct labeling can be illustrated by using a "receiver operating characteristic" (ROC) plot and reporting the area under the curve of

TABLE 8. Results for probability of detection and probability of false detection for the testing dataset (2003–05) produced from the best-performing configurations of an RFC, a GBC, an XGBC, and an ensemble classifier of these models (ENS) trained using five classes and binary classes.

	Probability of detection				Probability of false detection			
	RFC	GBC	XGBC	ENS	RFC	GBC	XGBC	ENS
Five-class label								
MCS	0.92	0.93	0.91	0.92	0.06	0.09	0.05	0.06
UCC	0.87	0.87	0.90	0.88	0.05	0.04	0.05	0.04
Tropical	0.62	0.63	0.71	0.68	<0.01	<0.01	<0.01	<0.01
Synoptic	0.92	0.92	0.93	0.94	0.01	0.01	0.02	0.01
Clutter	0.95	0.93	0.94	0.95	0.03	0.01	0.02	0.02
Mean	0.86	0.86	0.88	0.88	0.03	0.03	0.03	0.03
Binary label								
Non-MCS	0.96	0.96	0.95	0.96	0.08	0.09	0.08	0.09
MCS	0.92	0.91	0.92	0.91	0.04	0.04	0.05	0.04
Mean	0.94	0.94	0.94	0.94	0.06	0.07	0.07	0.07

TABLE 9. Confusion matrix for the testing dataset (2003–05) produced from an ensemble model with members trained using five classes.

		Actual:				
		MCS	UCC	Tropical	Synoptic	Clutter
Prediction:	MCS	539	47	12	2	0
	UCC	36	589	0	0	6
	Tropical	3	0	61	6	0
	Synoptic	0	1	17	149	0
	Clutter	0	29	0	2	115

TABLE 10. Confusion matrix for the testing dataset (2003–05) produced from an ensemble model with members trained using binary classes.

	Actual: MCS	Actual: Non-MCS
Prediction: MCS	528	41
Prediction: Non-MCS	50	995

probability of false detection versus probability of detection (AUC; Fig. 8a). The purpose of this plot is to tweak iteratively the probability threshold for hard classifications to explore how this modifies the sensitivity and specificity of a model (Zheng 2015). If this threshold is set to a relatively low MCS-label prediction probability (e.g., 0.05), the model would be very sensitive—that is, any sample with an MCS prediction probability exceeding 0.05 would be labeled as an MCS and thus very few actual MCSs would be labeled as non-MCS. The downside to this is that many non-MCS samples may be incorrectly labeled as MCS if they have some features that are similar to typical MCS features. Conversely, a relatively high prediction probability threshold (e.g., 0.95) would result in a very specific model—that is, only samples with a classification probability exceeding 0.95 would be labeled as an MCS. For a well-calibrated model, one would be confident that the most of the hard MCS-label predictions using this strict threshold would be correct. The downside to this extreme is that many actual MCS samples may be labeled as non-MCS. The ROC curve in Fig. 8a shows that the distributions of MCS and non-MCS samples along the prediction probability domain have little overlap. This result suggests that the model probabilities sufficiently separate the two types of samples. The ensemble model also performs better than a few select scikit-learn (0.18) algorithms, and much better than a "model" that predicts only MCS labels with 100% confidence. In addition, model performance is consistent for each of the three years in the testing data (Fig. 8c), and probabilistic classifications are reliable, in particular for probabilities above 0.90 and probabilities below 0.20 (Fig. 8d). Inconsistent results inside that range may be caused by a low number of samples. These results suggest that no further model calibration is needed and that the probabilistic predictions output by the classifiers and ensemble are representative of the expected model confidence.

6. 2015 and 2016 warm-season slice occurrence

The utility of the segmentation and classification procedure is explored by processing composite reflectivity

images every 15 min for May–September of 2015. This period was chosen because 1) May–September is when the frequency of MCSs is the highest (Ashley et al. 2003; Haberlie and Ashley 2016), 2) this period does not overlap with the testing or training data from 2003 to 2013, and 3) the results can be compared with a "climatology" generated for the PECAN project (see Fig. 1 in Geerts et al. 2017). In addition, the same period for 2016 is included to assess year-to-year consistency in the spatial frequency of slices. For each composite reflectivity image, the segmentation procedure extracts qualifying MCS slice candidates and calculates their features as described in Table 3. The values of these features, as well as spatiotemporal intensity information, are then stored in a database. This process is repeated using the thresholds defined in Table 2, where the values of convective-region search radius (CRSR; 6, 12, 24, and 48 km) and stratiform search radius (SSR; 48, 96, and 192 km) are modified between each iteration. This approach results in a total of 12 perturbations. This process is similar to the methods used by studies such as Clark et al. (2014) that explore the impact of varying thresholds on track density (cf. their Figs. 3 and 4). Further, various thresholds of classification prediction probability are used to gauge how modifying these values changes the spatial frequency of MCS slice candidates and MCS swaths. Feature information from qualifying MCS slice candidates is passed into the ensemble classifier, and the probability of an MCS label as output by the classifier is then associated with each slice. At this point, the database can then be queried to retrieve those slices that meet or exceed various probability values. These results are reported in Table 11.

In total, there were 1 518 093 slices generated from the 12 perturbations for both warm seasons, including 743 212 from 2015 and 774 881 from 2016. We removed 970 slices, representing 290 different time stamps from 2015, because they did not have a major axis length that was greater than or equal to 100 km, resulting in a total of 742 242 slices used in the following analysis. For 2016, 1179 slices, representing 341 time stamps, were similarly removed, resulting in a total of 773 702 slices used. When not using classifier predictions to stratify the dataset into MCS and non-MCS samples, the count of qualifying slices ranges from 43 443 to 84 198 for the 5-month period
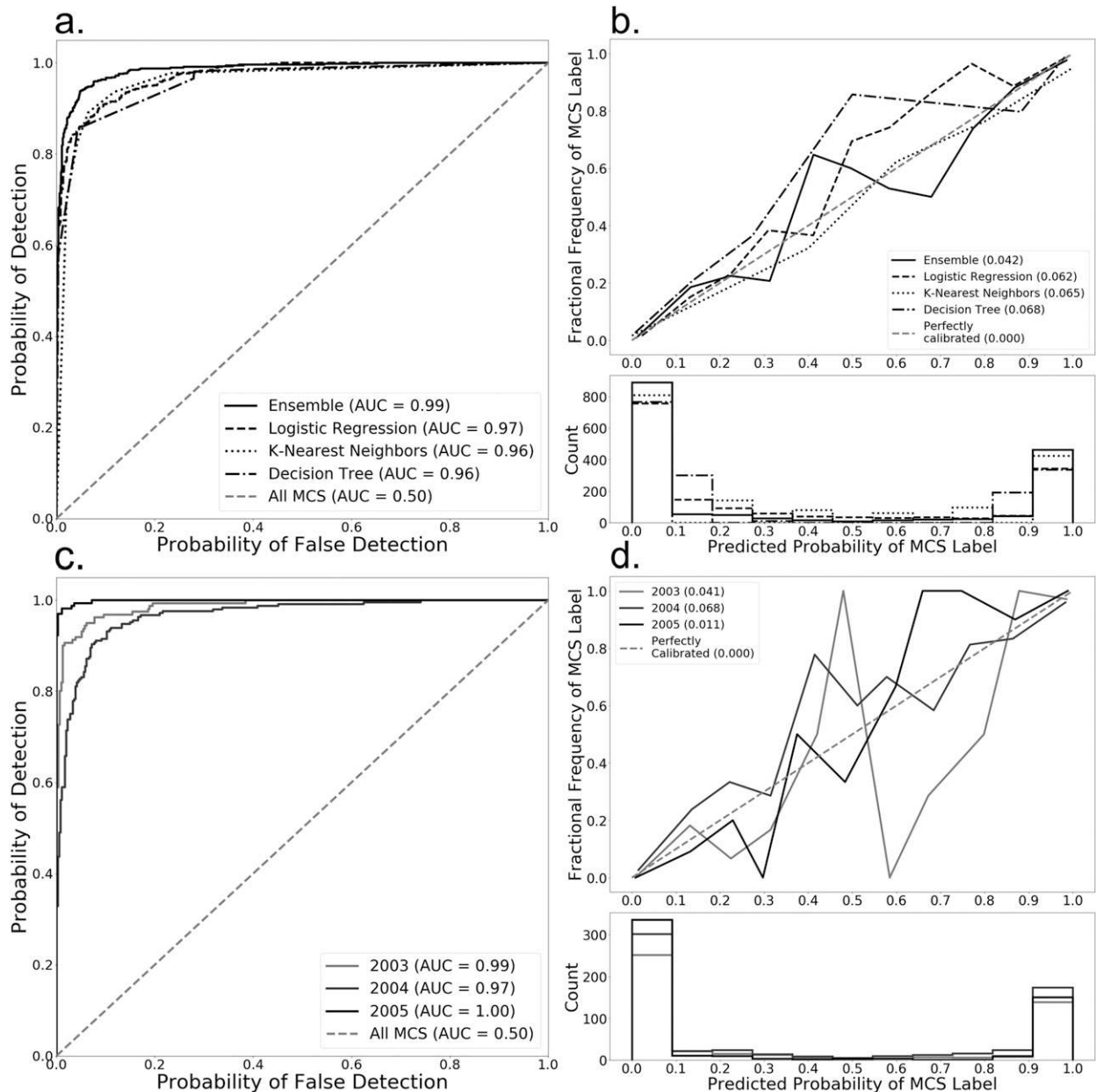
FIG. 8. (a) ROC curves and AUC values from the testing dataset comparing the ensemble classifier (solid black line) with three scikit-learn models with default parameters: logistic-regression (dashed lines), *k*-nearest-neighbor (dotted lines), and decision-tree (dot–dashed lines) classifiers. Also included is an "all MCS" classifier (dashed gray line) that predicts MCS as the label for every sample. (b) Results for the model calibration test, showing the relationship between the predicted probabilities of non-MCS (0) and MCS (1) samples, and the distribution of probabilistic classifications per 0.1 bin from 0.0 (non-MCS) to 1.0 (MCS). Included are Brier loss scores for each model. (c),(d) As in (a) and (b), respectively, but only for the ensemble model and broken down for each of the years in the testing dataset.

in 2015 and from 44 997 to 88 119 in 2016 (Table 11). After using the probabilistic threshold employed by the machine-learning algorithms to perform binary classification (e.g., MCS probability ≥ 0.5), this range is reduced to a minimum of 20 830 and to a maximum of 34 998, depending on the choices of search radius. The same range for 2016 is 21 914–37 069. When using

stricter MCS probability thresholds of 0.9 and 0.95, the slice count is further reduced (12 471–24 243 and 10 677–19 852, respectively), and the range of slice counts as a percentage of total slice counts is also reduced. For 2016, this count is reduced to 13 277–25 970 for a threshold of 0.9 and 11 392–20 901 for a threshold of 0.95. This result suggests that using a higher probability threshold is

TABLE 11. The effect of varying CRSR (km), SSR (km), and MCS probability thresholds (0.0, 0.5, 0.9, and 0.95) on slice count and total slice area (10^9 km^2). Probabilities for each slice are predicted by an ensemble classifier for slices gathered from May to September in 2015 and 2016.

CRSR (km)	SSR (km)	≥0.0		≥0.5		≥0.9		≥0.95	
		Count	Area	Count	Area	Count	Area	Count	Area
				2015					
6	48	43 443	0.79	20 830	0.61	12 471	0.47	10 677	0.42
6	96	41 120	1.10	25 177	0.96	18 361	0.81	15 984	0.74
6	192	39 666	1.44	25 379	1.23	19 339	1.02	16 685	0.90
12	48	54 008	0.93	23 447	0.70	14 519	0.55	12 539	0.50
12	96	51 318	1.27	28 153	1.08	20 333	0.91	17 578	0.82
12	192	49 827	1.62	28 393	1.35	21 112	1.09	17 830	0.95
24	48	70 443	1.15	26 825	0.84	17 231	0.67	14 923	0.61
24	96	67 312	1.52	31 642	1.24	22 705	1.04	19 188	0.92
24	192	65 756	1.86	31 940	1.49	23 014	1.17	18 838	1.00
48	48	89 371	1.46	30 204	1.03	20 417	0.84	16 914	0.74
48	96	85 780	1.85	34 998	1.46	24 725	1.17	19 852	1.00
48	192	84 198	2.20	34 975	1.66	24 243	1.25	18 711	1.02
				2016					
6	48	44 997	0.83	21 914	0.65	13 277	0.50	11 392	0.45
6	96	42 489	1.16	26 348	1.01	19 442	0.86	16 963	0.79
6	192	41 115	1.49	26 664	1.29	20 582	1.09	17 904	0.98
12	48	56 113	0.97	24 657	0.74	15 270	0.58	13 264	0.53
12	96	53 298	1.33	29 354	1.13	21 467	0.96	18 608	0.87
12	192	51 854	1.67	29 765	1.40	22 322	1.16	19 133	1.03
24	48	73 595	1.20	27 967	0.87	17 996	0.70	15 621	0.64
24	96	70 355	1.58	33 077	1.30	23 718	1.08	20 219	0.97
24	192	68 894	1.93	33 569	1.55	24 378	1.25	20 182	1.09
48	48	93 199	1.51	31 493	1.07	21 181	0.87	17 613	0.76
48	96	89 674	1.92	36 772	1.51	25 952	1.22	20 901	1.05
48	192	88 119	2.27	37 069	1.74	25 970	1.34	20 252	1.11

capturing events that are less sensitive to changes in search-radius values.

Spatial patterns of slice frequency using reflectivity and probability thresholds for 2015 are illustrated in Fig. 9. The shaded regions show the number of hours that locations in the CONUS were under the stratiform shield of a slice candidate with a 0.5 or higher probability of being an MCS as output by the ensemble classifier. Increasing the MCS predicted probability threshold to 0.95 generally limits the 40-h isopleth to the central United States. Despite the general agreement with the PECAN climatology (Geerts et al. 2017), there are regional maxima that appear to be spurious from previous work (Anderson and Arritt 1998; Ashley et al. 2003). For example, the maximum centered on Florida is related to the diurnal cycle of land and sea breezes that can generate bands of DMC (Lericos et al. 2002; Rickenbach et al. 2015). The maximum centered on the North Carolina coast is due to an unorganized convection maximum associated with the Gulf Stream (Rickenbach et al. 2015) and the incorrect labeling of Tropical Storm Ana (Stewart 2016). Although the maximum along the southwestern coast of Texas appears to be climatologically unusual, this region experienced several MCS passages in May of 2015, resulting in extreme rainfall and flooding (Wang et al. 2015). An encouraging result is that the Texas regional maximum is retained in all of the perturbations when the MCS slice probability threshold is increased. In contrast, the Florida and Carolina maxima are reduced when this threshold is increased, suggesting that the slices in these regions only have marginal similarities to actual MCS slices that were used to train the classifiers. For 2016, many of the same patterns are evident (Fig. 10). For example, increasing the probability threshold to 0.95 causes the 40-h isopleth to retreat north and west away from the East and Gulf Coasts. Of interest is that the Texas maximum shows up in 2016 as well, even with the higher probability thresholds. Upon further examination, this region experienced multiple MCS passages again in 2016, particularly during late May and early June. To be specific, two relatively slow moving MCSs produced almost constant precipitation in southeastern Texas from 26 to 28 May and three MCSs produced similar conditions in this area from 2 to 4 June. Relative to 2015, the Midwest and plains maximum shifted to the north and west, showing that the segmentation and classification procedure is capturing interannual variability in the occurrence of slices with high MCS probabilities.
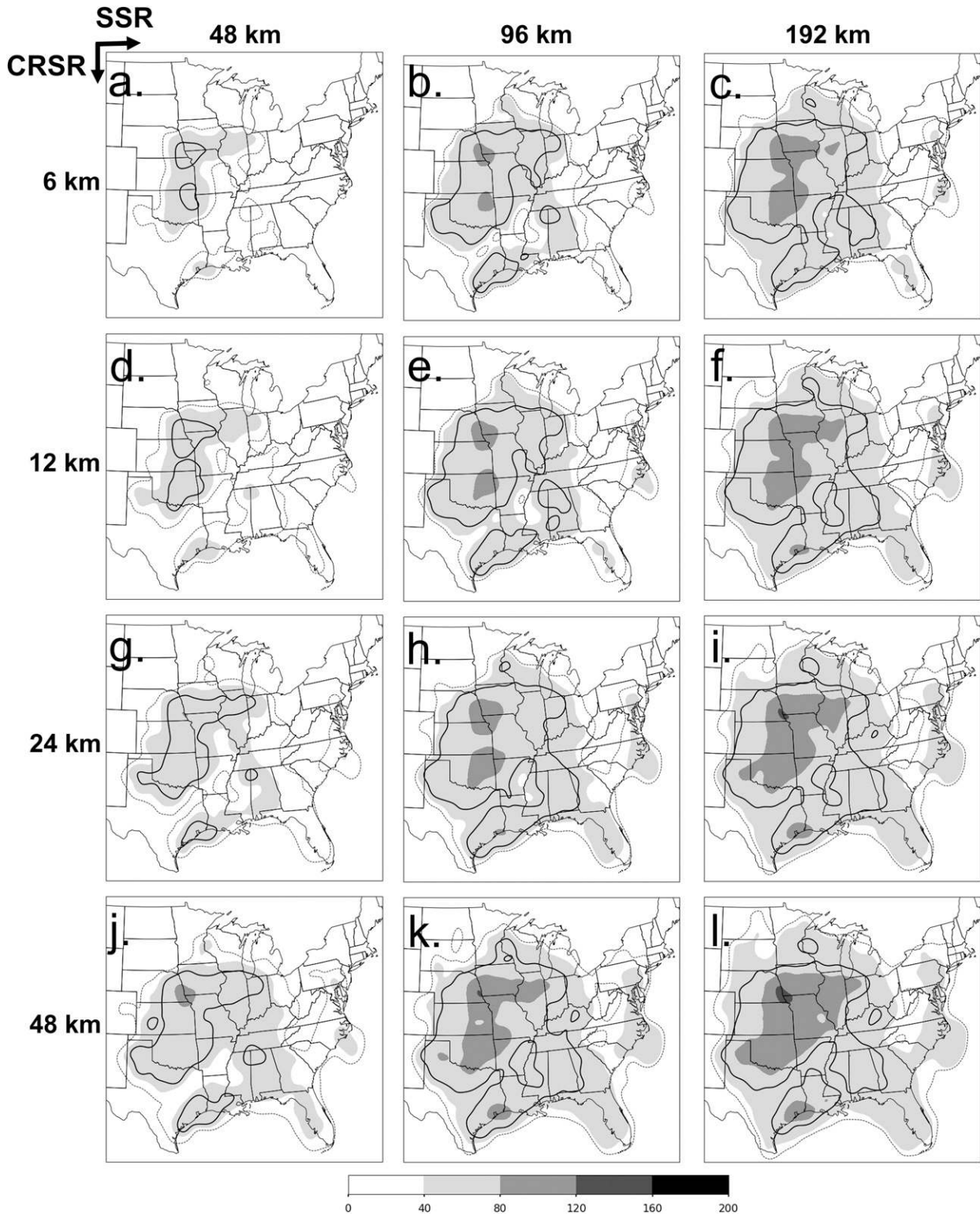
FIG. 9. Spatial occurrence (h; shaded) of slices with an MCS probability of 0.5 or higher in 2015 during May–September for varying convective-region and stratiform search radii. The solid line denotes the 40-h isopleth for slices with an MCS probability of 0.95 or higher, and the dotted line denotes the 40-h isopleth for all qualifying slices. The CRSR values are (a)–(c) 6, (d)–(f) 12, (g)–(i) 24, and (j)–(l) 48 km. The SSR values are (left) 48, (center) 96, and (right) 192 km.
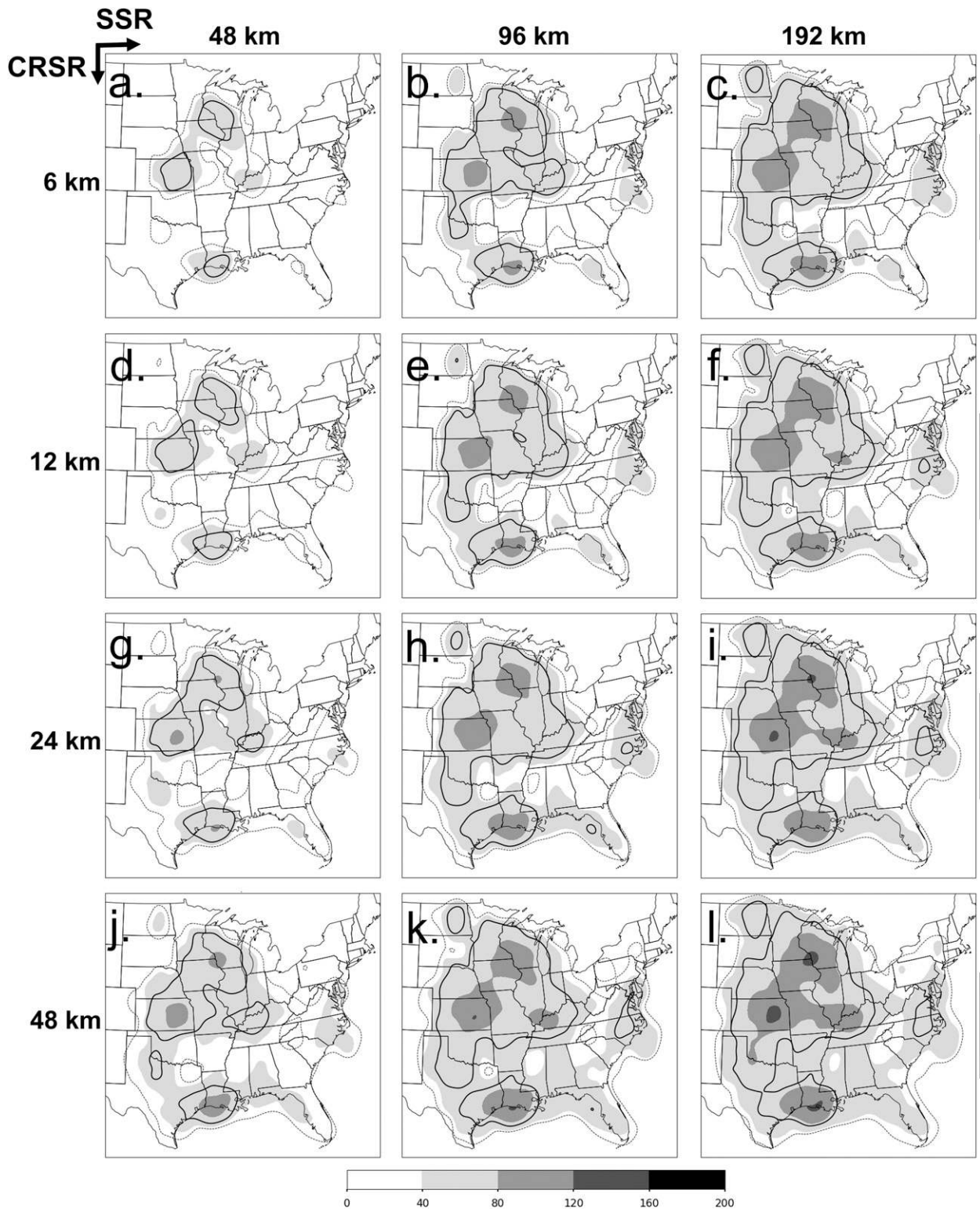
FIG. 10. As in Fig. 9, but for 2016.

7. Discussion and conclusions

This study examined the utility of three machine-learning algorithms for the problem of classifying MCSs in composite reflectivity images. A segmentation approach specific to identifying MCSs was outlined. It is composed of many steps, including 1) identifying regions of convection with intense precipitation, 2) connecting nearby regions to generate cores of convection, 3) measuring the major axis length of the cores to determine if they meet PJ00 criteria, and 4) associating stratiform precipitation regions within a specified distance of qualifying core regions. Select machine-learning algorithms were trained and tested using feature information from over 3000 manually labeled precipitation clusters for the purpose of classifying these clusters as 1) a midlatitude MCS, 2) an unorganized convective cluster, 3) a tropical system, 4) a synoptic system, or 5) ground clutter or noise. Application-specific classification was also explored by assigning the training and testing data binary labels of MCS or non-MCS. Next, precipitation clusters (slices) were automatically extracted from composite reflectivity data from May to June of 2015 using varying association radii for connecting 1) convective regions and 2) stratiform to qualifying core regions. Features from these clusters were extracted and used to generate predictions from an ensemble classifier containing the three trained machine-learning algorithms. Probabilistic predictions from the ensemble classifier, along with the varying association radii, were used to stratify the dataset. Statistics for these stratifications were generated, and frequency maps showed the spatial occurrence of precipitation clusters that met the various reflectivity and probabilistic thresholds.

This work affirmed the ability of machine-learning algorithms to accurately classify regions of convection on the basis of the subjective identification of events (e.g., Baldwin et al. 2005; Gagne et al. 2009; Lack and Fox 2012). Assessment of model accuracy was based on metrics (Table 6) designed to examine the ability of the trained algorithms to classify previously unseen labeled testing samples. On a class-by-class basis for the testing data, probability of detection ranged from 0.62 to 0.95 and probability of false detection ranged from less than 0.01 to 0.09, depending on the algorithm employed (Table 8). A binary-classification approach (MCS or non-MCS) produced better overall performance by reducing the probability of false detection for MCS samples. An ensemble of the three classifiers produced better probabilistic classifications than did output from other machine-learning algorithms alone (Fig. 8), and this performance was consistent across the three testing-data years. These values of model performance exceed

those presented in previous work (Gagne et al. 2009; Lack and Fox 2012). One of the reasons for this improvement could be the more sophisticated algorithms that were employed in this study. For example, Lack and Fox (2012) used a decision-tree algorithm to generate classifications. Although decision trees allow a relatively easy assessment of model behavior, they prioritize local split decisions, which can result in suboptimal exploration of the decision space (Breiman 2001). Further, both Gagne et al. (2009) and Lack and Fox (2012) used classifications with samples that were potentially very similar to one another. For example, Gagne et al. (2009) separated linear-system samples into leading, trailing, and parallel stratiform subtypes. As a result, separating samples on the basis of their features becomes more difficult, which can lead to a reduction in model performance. A similar issue occurred in this study, because tropical systems commonly share similar attributes with other classifications. To improve subtype classification, even more sophisticated image classification approaches, like convolutional neural networks (LeCun and Bengio 1995; Krizhevsky et al. 2012; Dieleman et al. 2015), may be needed to improve performance. Future related work will explore convolutional neural networks and similar approaches for the purposes of increasing the accuracy of precipitation cluster classification.

Application of the segmentation and classification procedure to the 2015 and 2016 warm seasons produces reasonable results that are comparable to existing climatologies (Anderson and Arritt 1998; Ashley et al. 2003; Geerts et al. 2017). Although tracking of the identified slices was not demonstrated in this paper (tracking will be discussed in Part II), the spatial occurrence of MCS-like regions of precipitation should still be clustered in climatologically favored regions of MCS activity. Increasing the MCS probability threshold (and thus increasing the specificity of the models) resulted in a reduction of MCS activity along the eastern Gulf and Atlantic coasts (Figs. 9 and 10). Stricter MCS probability thresholds appear to result in more agreement between the automated segmentation and classification output (Figs. 9 and 10) and location of the training and testing samples of MCS (Fig. 7a), despite not including features that encode spatial information. This result illustrates a potential limitation of this study: the machine-learning predictions are based on the authors' interpretation of the subjective definition of an MCS. The results suggest that this may be particularly important when working with banded convection in the southeastern United States. Users of this framework could modify the balance between the probability of detection and the probability of false detection or the hand-labeling approach to be more appropriate for their

needs. In that theme, this study should motivate future work to combine the objective and subjective definitions of MCS in a way that emphasizes computational approaches over manual identification. One physical mechanism common to MCSs that may inform this exploration is the perceived interaction between a cold pool and warm inflow, caused in part by the decay of relatively short-lived DMC and its outflow generating new DMC (PJ00). This critical process is not encoded into the objective definition, despite its widespread ubiquity in MCS studies (Table 1). We do not claim that the methods described in this study fully discriminate between events that exhibit this process and those that do not. Indeed, to specifically identify proxies of spatiotemporal processes such as interactions between cold pools and warm inflow, more complex machine-learning models such as long short-term memory recurrent neural networks (Hochreiter and Schmidhuber 1997) may be required. Also, removing non-MCS slices from consideration during the tracking procedure reduces the complexity of spatiotemporally associating candidate MCS slices. This topic is discussed in in Part II of this paper. The process described here (in Part I), combined with the tracking detailed in Part II, will ultimately allow for an objective, automated, spatiotemporal assessment of CONUS MCS activity.

REFERENCES

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecasting*, **31**, 581–599, https://doi.org/10.1175/WAF-D-15-0113.1.

Anderson, C. J., and R. W. Arritt, 1998: Mesoscale convective complexes and persistent elongated convective systems over the United States during 1992 and 1993. *Mon. Wea. Rev.*, **126**, 578–599, https://doi.org/10.1175/1520-0493(1998)126<0578:MCCAPE>2.0.CO;2.

Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.*, **47**, 805–818, https://doi.org/10.1175/2007JAMC1611.1.

Ashley, W. S., and T. L. Mote, 2005: Derecho hazards in the United States. *Bull. Amer. Meteor. Soc.*, **86**, 1577–1592, https://doi.org/10.1175/BAMS-86-11-1577.

——, ——, P. G. Dixon, S. L. Trotter, E. J. Powell, J. D. Durkee, and A. J. Grundstein, 2003: Distribution of mesoscale convective complex rainfall in the United States. *Mon. Wea. Rev.*, **131**, 3003–3017, https://doi.org/10.1175/1520-0493(2003)131<3003:DOMCCR>2.0.CO;2.

——, M. L. Bentley, and J. A. Stallins, 2012: Urban-induced thunderstorm modification in the southeast United States. *Climatic Change*, **113**, 481–498, https://doi.org/10.1007/s10584-011-0324-1.

Baldwin, M. E., J. S. Kain, and S. Lakshmivarahan, 2005: Development of an automated classification procedure for rainfall systems. *Mon. Wea. Rev.*, **133**, 844–862, https://doi.org/10.1175/MWR2892.1.

Bluestein, H. B., and M. H. Jain, 1985: Formation of mesoscale lines of precipitation: Severe squall lines in Oklahoma during the spring. *J. Atmos. Sci.*, **42**, 1711–1732, https://doi.org/10.1175/1520-0469(1985)042<1711:FOMLOP>2.0.CO;2.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brooks, H. E., and D. J. Stensrud, 2000: Climatology of heavy rain events in the United States from hourly precipitation observations. *Mon. Wea. Rev.*, **128**, 1194–1201, https://doi.org/10.1175/1520-0493(2000)128<1194:COHREI>2.0.CO;2.

Byers, H. R., and R. R. Braham Jr., 1948: Thunderstorm structure and circulation. *J. Atmos. Sci.*, **5**, 71–86, https://doi.org/10.1175/1520-0469(1948)005<0071:TSAC>2.0.CO;2.

Carbone, R. E., and J. D. Tuttle, 2008: Rainfall occurrence in the U.S. warm season: The diurnal cycle. *J. Climate*, **21**, 4132–4146, https://doi.org/10.1175/2008JCLI2275.1.

——, ——, D. A. Ahijevych, and S. B. Trier, 2002: Inferences of predictability associated with warm season precipitation episodes. *J. Atmos. Sci.*, **59**, 2033–2056, https://doi.org/10.1175/1520-0469(2002)059<2033:IOPAWW>2.0.CO;2.

Chang, W., M. L. Stein, J. Wang, V. R. Kotamarthi, and E. J. Moyer, 2016: Changes in spatiotemporal precipitation patterns in changing climate conditions. *J. Climate*, **29**, 8355–8376, https://doi.org/10.1175/JCLI-D-15-0844.1.

Chen, T., and C. Guestrin, 2016: XGBoost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, Association for Computing Machinery, 785–794, https://dl.acm.org/citation.cfm?id=2939785.

Clark, A. J., R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517–542, https://doi.org/10.1175/WAF-D-13-00098.1.

Cohen, A. E., M. C. Coniglio, S. F. Corfidi, and S. J. Corfidi, 2007: Discrimination of mesoscale convective system environments using sounding observations. *Wea. Forecasting*, **22**, 1045–1062, https://doi.org/10.1175/WAF1040.1.

Comstock, I. J., 2011: A classification scheme for landfalling tropical cyclones based on precipitation variables derived from GIS and ground radar analysis. Ph.D. dissertation, University of Alabama, 93 pp., https://ir.ua.edu/handle/123456789/1025.

Coniglio, M. C., J. Y. Hwang, and D. J. Stensrud, 2010: Environmental factors in the upscale growth and longevity of MCSs derived from Rapid Update Cycle analyses. *Mon. Wea. Rev.*, **138**, 3514–3539, https://doi.org/10.1175/2010MWR3233.1.

Corfidi, S. F., M. C. Coniglio, A. E. Cohen, and C. M. Mead, 2016: A proposed revision to the definition of "derecho." *Bull. Amer. Meteor. Soc.*, **97**, 935–949, https://doi.org/10.1175/BAMS-D-14-00254.1.

Crum, T. D., R. L. Alberty, and D. W. Burgess, 1993: Recording, archiving, and using WSR-88D data. *Bull. Amer. Meteor. Soc.*, **74**, 645–653, https://doi.org/10.1175/1520-0477(1993)074<0645:RAAUWD>2.0.CO;2.

Cunning, J. B., 1986: The Oklahoma-Kansas Preliminary Regional Experiment for STORM-Central. *Bull. Amer. Meteor. Soc.*, **67**, 1478–1486, https://doi.org/10.1175/1520-0477(1986)067<1478:TOKPRE>2.0.CO;2.

Davis, C., and Coauthors, 2004: The Bow Echo and MCV Experiment: Observations and opportunities. *Bull. Amer. Meteor. Soc.*, **85**, 1075–1093, https://doi.org/10.1175/BAMS-85-8-1075.

Dieleman, S., K. W. Willett, and J. Dambre, 2015: Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Mon. Not. Roy. Astron. Soc.*, **450**, 1441–1459, https://doi.org/10.1093/mnras/stv632.

Doswell, C. A., III, 2001: Severe convective storms—An overview. *Severe Convective Storms*, *Meteor. Monogr.*, No. 50, Amer. Meteor. Soc., 1–26.

——, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Elith, J., J. R. Leathwick, and T. Hastie, 2008: A working guide to boosted regression trees. *J. Anim. Ecol.*, **77**, 802–813, https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Fabry, F., V. Meunier, B. P. Treserras, A. Cournoyer, and B. Nelson, 2017: On the climatological use of radar data mosaics: Possibilities and challenges. *Bull. Amer. Meteor. Soc.*, **98**, 2135–2148, https://doi.org/10.1175/BAMS-D-15-00256.1.

Fairman, J. G., Jr., D. M. Schultz, D. J. Kirshbaum, S. L. Gray, and A. I. Barrett, 2016: Climatology of banded precipitation over the contiguous United States. *Mon. Wea. Rev.*, **144**, 4553–4568, https://doi.org/10.1175/MWR-D-16-0015.1.

——, D. M. Schultz, D. J. Kirshbaum, S. L. Gray, and A. I. Barrett, 2017: Climatology of size, shape, and intensity of precipitation features over Great Britain and Ireland. *J. Hydrometeor.*, **18**, 1595–1615, https://doi.org/10.1175/JHM-D-16-0222.1.

Feng, Z., L. R. Leung, S. Hagos, R. A. Houze, C. D. Burleyson, and K. Balaguru, 2016: More frequent intense and long-lived storms dominate the springtime trend in central US rainfall. *Nat. Commun.*, **7**, 13429, https://doi.org/10.1038/ncomms13429.

Fiolleau, T., and R. Roca, 2013: An algorithm for the detection and tracking of tropical mesoscale convective systems using infrared images from geostationary satellite. *IEEE Trans. Geosci. Remote Sens.*, **51**, 4302–4315, https://doi.org/10.1109/TGRS.2012.2227762.

Fritsch, J. M., and G. S. Forbes, 2001: Mesoscale convective systems. *Severe Convective Storms*, *Meteor. Monogr.*, No. 50, Amer. Meteor. Soc., 323–358.

Gagne, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *J. Atmos. Oceanic Technol.*, **26**, 1341–1353, https://doi.org/10.1175/2008JTECHA1205.1.

——, ——, J. B. Basara, and R. A. Brown, 2012: Tornadic supercell environments analyzed using surface and reanalysis data: A spatiotemporal relational data-mining approach. *J. Appl. Meteor. Climatol.*, **51**, 2203–2217, https://doi.org/10.1175/JAMC-D-11-060.1.

——, ——, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, ——, S. E. Haupt, and J. K. Williams, 2017: Evaluation of statistical learning configurations for gridded solar irradiance forecasting. *Sol. Energy*, **150**, 383–393, https://doi.org/10.1016/j.solener.2017.04.031.

Gallus, W. A., Jr., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, https://doi.org/10.1175/2007WAF2006120.1.

Geerts, B., 1998: Mesoscale convective systems in the southeast United States during 1994–95: A survey. *Wea. Forecasting*, **13**, 860–869, https://doi.org/10.1175/1520-0434(1998)013<0860:MCSITS>2.0.CO;2.

——, and Coauthors, 2017: The 2015 Plains Elevated Convection at Night field project. *Bull. Amer. Meteor. Soc.*, **98**, 767–786, https://doi.org/10.1175/BAMS-D-15-00257.1.

Haberlie, A. M., and W. S. Ashley, 2016: A U.S. climatology of mesoscale convective systems. *15th Annual Student Conf.*, New Orleans, LA, Amer. Meteor. Soc., S81, https://ams.confex.com/ams/96Annual/webprogram/Paper292206.html.

——, and ——, 2018: Identifying mesoscale convective systems in radar mosaics. Part II: Tracking. *J. Appl. Meteor. Climatol.*, **57**, 1599–1621, https://doi.org/10.1175/JAMC-D-17-294.1.

——, ——, and T. Pingel, 2015: The effect of urbanization on the climatology of thunderstorm initiation. *Quart. J. Roy. Meteor. Soc.*, **141**, 663–675, https://doi.org/10.1002/qj.2499.

——, ——, A. J. Fultz, and S. M. Eagan, 2016: The effect of reservoirs on the climatology of warm-season thunderstorms in southeast Texas, USA. *Int. J. Climatol.*, **36**, 1808–1820, https://doi.org/10.1002/joc.4461.

Halverson, J. B., 2014: A mighty wind: The derecho of June 29, 2012. *Weatherwise*, **67**, 24–31, https://doi.org/10.1080/00431672.2014.918788.

Hane, C. E., J. A. Haynes, D. L. Andra, and F. H. Carr, 2008: The evolution of morning convective systems over the U.S. Great Plains during the warm season. Part II: A climatology and the influence of environmental factors. *Mon. Wea. Rev.*, **136**, 929–944, https://doi.org/10.1175/2007MWR2016.1.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: Unsupervised learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., T. Hastie, R. Tibshirani, and J. Friedman, Eds., Springer, 485–585.

Hilgendorf, E. R., and R. H. Johnson, 1998: A study of the evolution of mesoscale convective systems using WSR-88D data. *Wea. Forecasting*, **13**, 437–452, https://doi.org/10.1175/1520-0434(1998)013<0437:ASOTEO>2.0.CO;2.

Hitchens, N. M., M. E. Baldwin, and R. J. Trapp, 2012: An object-oriented characterization of extreme precipitation-producing convective systems in the midwestern United States. *Mon. Wea. Rev.*, **140**, 1356–1366, https://doi.org/10.1175/MWR-D-11-00153.1.

Hochreiter, S., and J. Schmidhuber, 1997: Long short-term memory. *Neural Comput.*, **9**, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

Hocker, J. E., and J. B. Basara, 2008: A 10-year spatial climatology of squall line storms across Oklahoma. *Int. J. Climatol.*, **28**, 765–775, https://doi.org/10.1002/joc.1579.

Houze, R. A., Jr., 2004: Mesoscale convective systems. *Rev. Geophys.*, **42**, RG4003, https://doi.org/10.1029/2004RG000150.

Jirak, I. L., W. R. Cotton, and R. L. McAnelly, 2003: Satellite and radar survey of mesoscale convective system development. *Mon. Wea. Rev.*, **131**, 2428–2449, https://doi.org/10.1175/1520-0493(2003)131<2428:SARSOM>2.0.CO;2.

Kolodziej Hobson, A. G., V. Lakshmanan, T. M. Smith, and M. Richman, 2012: An automated technique to categorize storm type from radar and near-storm environment data. *Atmos. Res.*, **111**, 104–113, https://doi.org/10.1016/j.atmosres.2012.03.004.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: Imagenet classification with deep convolutional neural networks. *Proc. 25th Conf. on Advances in Neural Information Processing Systems*, Lake Tahoe, NV, Neural Information Processing Systems Foundation, 1097–1105, https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

Kunkel, K. E., D. R. Easterling, D. Kristovich, B. Gleason, L. Stoecker, and R. Smith, 2012: Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. *J. Hydrometeor.*, **13**, 1131–1141, https://doi.org/10.1175/JHM-D-11-0108.1.

Lack, S. A., and N. I. Fox, 2012: Development of an automated approach for identifying convective storm type using reflectivity-derived and near-storm environment data. *Atmos. Res.*, **116**, 67–81, https://doi.org/10.1016/j.atmosres.2012.02.009.

Lakshmanan, V., and T. Smith, 2009: Data mining storm attributes from spatial grids. *J. Atmos. Oceanic Technol.*, **26**, 2353–2365, https://doi.org/10.1175/2009JTECHA1257.1.

——, and ——, 2010: An objective method of evaluating and devising storm-tracking algorithms. *Wea. Forecasting*, **25**, 701–709, https://doi.org/10.1175/2009WAF2222330.1.

——, M. Miller, and T. Smith, 2013: Quality control of accumulated fields by applying spatial and temporal constraints. *J. Atmos. Oceanic Technol.*, **30**, 745–758, https://doi.org/10.1175/JTECH-D-12-00128.1.

LeCun, Y., and Y. Bengio, 1995: Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., MIT Press, 255–258.

Lericos, T. P., H. E. Fuelberg, A. I. Watson, and R. L. Holle, 2002: Warm season lightning distributions over the Florida peninsula as related to synoptic patterns. *Wea. Forecasting*, **17**, 83–98, https://doi.org/10.1175/1520-0434(2002)017<0083:WSLDOT>2.0.CO;2.

Lombardo, K. A., and B. A. Colle, 2010: The spatial and temporal distribution of organized convective structures over the Northeast and their ambient conditions. *Mon. Wea. Rev.*, **138**, 4456–4474, https://doi.org/10.1175/2010MWR3463.1.

Markowski, P., and Y. Richardson, 2011: *Mesoscale Meteorology in Midlatitudes.* Wiley-Blackwell, 424 pp.

Matyas, C. J., 2009: A spatial analysis of radar reflectivity regions within Hurricane Charley (2004). *J. Appl. Meteor. Climatol.*, **48**, 130–142, https://doi.org/10.1175/2008JAMC1910.1.

——, 2010: Use of ground-based radar for climate-scale studies of weather and rainfall: Radar and climatology. *Geogr. Compass*, **4**, 1218–1237, https://doi.org/10.1111/j.1749-8198.2010.00370.x.

——, 2014: Conditions associated with large rain-field areas for tropical cyclones landfalling over Florida. *Phys. Geogr.*, **35**, 93–106, https://doi.org/10.1080/02723646.2014.893476.

McEnery, J., J. Ingram, Q. Duan, T. Adams, and L. Anderson, 2005: NOAA's Advanced Hydrologic Prediction Service: Building pathways for better science in water forecasting. *Bull. Amer. Meteor. Soc.*, **86**, 375–385, https://doi.org/10.1175/BAMS-86-3-375.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Miller, P. W., and T. L. Mote, 2017: Standardizing the definition of a "pulse" thunderstorm. *Bull. Amer. Meteor. Soc.*, **98**, 905–913, https://doi.org/10.1175/BAMS-D-16-0064.1.

Mulder, K. J., and D. M. Schultz, 2015: Climatology, storm morphologies, and environments of tornadoes in the British Isles: 1980–2012. *Mon. Wea. Rev.*, **143**, 2224–2240, https://doi.org/10.1175/MWR-D-14-00299.1.

Parker, M. D., and R. H. Johnson, 2000: Organizational modes of midlatitude mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 3413–3436, https://doi.org/10.1175/1520-0493(2001)129<3413:OMOMMC>2.0.CO;2.

——, and J. C. Knievel, 2005: Do meteorologists suppress thunderstorms?: Radar-derived statistics and the behavior of moist convection. *Bull. Amer. Meteor. Soc.*, **86**, 341–358, https://doi.org/10.1175/BAMS-86-3-341.

——, and D. A. Ahijevych, 2007: Convective episodes in the east-central United States. *Mon. Wea. Rev.*, **135**, 3707–3727, https://doi.org/10.1175/2007MWR2098.1.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830, http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh model's ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30**, 892–913, https://doi.org/10.1175/WAF-D-14-00118.1.

Rickenbach, T. M., R. Nieto-Ferreira, C. Zarzar, and B. Nelson, 2015: A seasonal and diurnal climatology of precipitation organization in the southeastern United States. *Quart. J. Roy. Meteor. Soc.*, **141**, 1938–1956, https://doi.org/10.1002/qj.2500.

Schumacher, R. S., and R. H. Johnson, 2005: Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **133**, 961–976, https://doi.org/10.1175/MWR2899.1.

Smith, J. A., D. J. Seo, M. L. Baeck, and M. D. Hudlow, 1996: An intercomparison study of NEXRAD precipitation estimates. *Water Resour. Res.*, **32**, 2035–2045, https://doi.org/10.1029/96WR00270.

Stewart, S. R., 2016: The 2015 Atlantic hurricane season: Fewer storms, with some highlights. *Weatherwise*, **69** (3), 28–35, https://doi.org/10.1080/00431672.2016.1159488.

Theodoridis, S., and K. Koutroumbas, 2003: *Pattern Recognition.* Academic Press, 689 pp.

Trapp, R. J., S. A. Tessendorf, E. S. Godfrey, and H. E. Brooks, 2005: Tornadoes from squall lines and bow echoes. Part I: Climatological distribution. *Wea. Forecasting*, **20**, 23–34, https://doi.org/10.1175/WAF-835.1.

Tuttle, J. D., and R. E. Carbone, 2011: Inferences of weekly cycles in summertime rainfall. *J. Geophys. Res.*, **116**, D20213, https://doi.org/10.1029/2011JD015819.

van der Walt, S., S. C. Colbert, and G. Varoquaux, 2011: The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30, https://doi.org/10.1109/MCSE.2011.37.

——, and Coauthors, 2014: Scikit-image: Image processing in Python. *PeerJ*, **2**, e453, https://doi.org/10.7717/peerj.453.

Wang, S.-Y., W. Huang, H. Hsu, and R. R. Gillies, 2015: Role of the strengthened El Niño teleconnection in the May 2015 floods over the southern Great Plains. *Geophys. Res. Lett.*, **42**, 8140–8146, https://doi.org/10.1002/2015GL065211.

Weckwerth, T. M., and Coauthors, 2004: An overview of the International H$_2$O Project (IHOP_2002) and some preliminary highlights. *Bull. Amer. Meteor. Soc.*, **85**, 253–277, https://doi.org/10.1175/BAMS-85-2-253.

Weisman, M. L., and Coauthors, 2015: The Mesoscale Predictability Experiment (MPEX). *Bull. Amer. Meteor. Soc.*, **96**, 2127–2149, https://doi.org/10.1175/BAMS-D-13-00281.1.

Westerling, A. L., A. Gershunov, D. R. Cayan, and T. P. Barnett, 2002: Long lead statistical forecasts of area burned in western US wildfires by ecosystem province. *Int. J. Wildland Fire*, **11**, 257–266, https://doi.org/10.1071/WF02009.

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

Zheng, A., 2015: *Evaluating Machine Learning Models*. O'Reilly Media, https://www.oreilly.com/data/free/evaluating-machine-learning-models.csp.

Zipser, E. J., 1982: Use of a conceptual model of the life-cycle of mesoscale convective systems to improve very-short-range forecasts. *Nowcasting*, K. A. Browning, Ed., Academic Press, 191–204.