Comparison of Machine Learning Techniques for Convective Morphology Classification from Radar Imagery

Jonathan E. Thielen

Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa

William A. Gallus – Mentor

Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa

Alex M. Haberlie – Mentor

Department of Geography and Anthropology, Louisiana State University, Baton Rouge, Louisiana

ABSTRACT

Mesoscale convective systems (MCSs) play a critical role in the hydroclimate and occurrence of severe weather in the central United States. In the analysis and forecasting of these convective systems, the morphology—the shape, structure, and organization—is an important system characteristic. Numerous past studies have used radar mosaic imagery to evaluate both observed and modeled systems according to categories or modes of convective morphology. However, their methods have relied on manual classification, which limits potential sample sizes and risks inconsistency of results. Recent advancements in the use of image analysis software to automatically extract convective systems from radar mosaics and of machine learning algorithms to classify said systems have shown promise in the automation of convective morphology analysis. However, no framework yet exists for automated classification according to detailed convective modes with subtypes of cellular and linear systems, and so, this study seeks to evaluate and compare machine learning techniques in addressing this problem of detailed convective classification. Results from this study show that an ensemble of decision tree ensemble classifiers which utilize a large set of input parameters designed to differentiate between the convective modes performs best at this detailed classification task. This classifier performs better than any of the decision tree ensemble classifiers relying on only the basic areal and intensity parameters used in past studies or convolutional neural networks (CNNs), even though prior studies have suggested that CNNs tend to outperform other image classification techniques. However, with an overall accuracy score of 59.37%, this best-performing decision tree ensemble technique remains insufficient for future use as an automated tool in research or operations. Therefore, additional steps are considered for how to improve the classification accuracy and obtain a more reliable method for future use.

1. Introduction

Across the central United States, mesoscale convective systems (MCSs) have a significant role in both the climatology of precipitation and the occurrence of severe hazards. In this region, MCSs are responsible for roughly 30% to 70% of the warm-season precipitation and are therefore crucial to the region's agricultural production (Fritsch et al. 1986). However, numerous severe weather risks, including hail, wind, tornadoes, and flooding, are also known to occur within MCSs (Jirak and Cotton 2007; Haberlie and Ashley 2018a, hereafter HA18a). These systems have therefore continued to be a focus of intense study in recent years (Houze 2004; Geerts et al. 2017), particularly because they remain poorly forecasted (Jirak and Cotton 2007).

When investigating MCSs, most studies use radar imagery as a critical diagnostic tool. While MCSs are often objectively defined according to physical system characteristics not tied to a particular form of remote sensing, their presence and organization are most commonly evaluated using radar (Parker and Johnson 2000). Additionally, to obtain a sense of the ongoing dynamical processes of these systems in an operational context, the shape, organization, and structure—the morphology—of the convection needs to be evaluated from radar imagery in place of in situ measurements. Numerous schemes have therefore been developed to systematically classify the morphology of convective systems into distinct categories or modes using manual analysis (Parker and Johnson 2000; Fowle and Roebber 2003; Gallus et al. 2008, hereafter G08). In addition to reflecting the dynamical processes associated with the systems, the convective mode is strongly associated with the varieties of hazards the system could produce. Specifically, cellular modes are most strongly associated with hail and tornadoes, whereas linear modes give all types of severe weather depending on the exact classification, but with wind and flooding threats being of particular note (G08). For these reasons, the classification of convective morphology and the implications of morphology have remained essential problems within the field.

In recent years, however, some concerns have been raised in regards to the traditional, manual methods for classifying systems. First, manual methods place severe practical limits on the amount of data that can be feasibly analyzed in a single study, thereby restricting the possible sample sizes for such studies (G08, Lakshmanan and Smith 2009, Thielen et al. 2018). Because of this, many promising areas of research, such as probabilistic forecasts of morphology using large ensembles or climatologies of modes, are prohibitively intensive if they are reliant upon manual procedures. Additionally, these subjective methods rely upon the investigator's pattern recognition, which is "open to judgment" and potentially inconsistent (Corfidi et al. 2016). Therefore, significant work has been undertaken to implement automated procedures for MCS classification.

While these efforts can be traced back to the work of Biggerstaff and Listemaa (2000) in developing automated techniques to discriminate between radar signatures of

convective and stratiform precipitation regions, one of the first substantial studies directly implementing an automated classification procedure for rainfall systems was Baldwin et al. (2005). This study utilized a nearest-neighbor classifier on morphological and rainfall parameters to broadly convective systems into linear, cellular, and stratiform classes with approximately 85% accuracy. Gange et al. (2009) sought to use a more detailed set of six morphological modes (split into three cellular types and three linear types) and more sophisticated machine learning techniques. They found that the more robust random forest technique, consisting of an ensemble of decision trees, attained the best performance in both the general cellular vs. linear classification problem (91.8% accuracy) and the specific-mode problem (70.1% accuracy). While some later work found that the addition of near-storm environmental data to the radar-based techniques can improve the classification procedures (Lack and Fox 2012), the accuracy of classification for detailed schemes using radar data alone, such as the nine-category scheme of G08 that demonstrated strong correlations with storm hazard types, has remained poor.

However, in the past several years, interest in the use of machine learning techniques in analyzing severe weather, especially with larger datasets and through more advanced techniques like convolutional neural networks (CNNs), has grown tremendously (McGovern et al. 2017). This includes substantial interest in the detailed analysis of convective precipitation systems (Herman and Schumacher 2018a, b) and climatological perspectives on MCSs (HA18a). This latter study sought to develop automated MCS segmentation and classification procedures using image processing software and machine learning algorithms to detect MCSs from radar mosaics over the conterminous U.S. (CONUS), all while creating this classification system to be highly configurable and publicly shared. They used three primary algorithms, all of which are based on ensembles of decision trees: random forests, gradient boosting, and XGBoost. For their classification, which was a broad classification between MCS vs. non-MCS types, they found results consistent with to slightly better than those of Gange et al. (2009) in terms of accuracy (91-96%), and they stated that their overall measures of model performance were higher than those of Gange et al. (2009) and Lack and Fox (2012). However, due to their use of a more general scheme, Haberlie and Ashley (2018a) identify that a substantial area of future work with their procedure is the improvement of subtype classification, thereby leaving the exploration of detailed morphology classification by automated procedures as an unsolved problem. They suggest that even more sophisticated techniques such as CNNs may be needed to accomplish this improvement. These CNNs have shown success in classifying MCSs between quasi-linear convective system (QLCS) and non-QLCS types (Haberlie and Ashley 2018b).

Motivated by the state of remaining work and the need for a robust and reliable automated method to classify MCSs according to detailed modes, this study seeks to extend the existing segmentation and classification procedures of HA18a to the nine-category

scheme of G08. Two main approaches will be compared. In the first, the same decision tree ensemble-based algorithms will continue to be used, but the list of morphological parameters used will be expanded to include parameters able to discriminate between the various cellular and linear subtypes of the G08 scheme. In the second, a CNN will be applied to classify extracted system slices according to the detailed modes of G08. Through the development of these two procedures, this study will seek to answer the questions of how the performance of these two techniques compares in categorizing radar signatures of convective systems according to detailed modes and whether or not either technique has sufficiently high reliability to be useful as an automated technique in research or operations.

Section 2 further elaborates on the methods used in the two procedures, as well as the data and analytical techniques employed. Analysis and results then follow in Section 3, and Section 4 presents the conclusions, summary, and directions for future work.

2. Data and Methods

a. Radar Data and Cases Under Study

This study takes its source radar data from GridRad, a 3D gridded NEXRAD radar product (Bowman and Homeyer 2017). These data exist on a regular 0.02° longitude by 0.02° latitude by 1 km altitude grid across the continental United States, with hourly reflectivity data available from 2004 to 2016. For this study, only 2D column-maximum (also known as composite) reflectivity is used, in accord with past studies of convective morphology (G08, HA18a). Additionally, because this study relies upon image analysis that is sensitive to extent, intensity, and orientation, these radar data are regridded to a U.S.-centered Lambert Conformal Conic projection with 2 km grid spacing using nearest neighbor interpolation. The domain is also restricted to the central U.S. where MCS activity is most common and GridRad data are consistently available (Fig.1).

A random sample of hours from all available hours of GridRad data during the warm season (May through September) was taken to obtain cases to use as training and testing data for the machine learning models. If a convective system was present in the study domain during that hour, and no scan quality issues were noted in the image, the timestamp was added to the list of events under consideration. If no suitable system was identified, the timestamp was excluded from

Fig. 1: The spatial domain used in this study, which ranges from 108° W to 80° W and 29° N to 49° N. Candidate convective systems extending beyond this domain were filtered out and not classified.

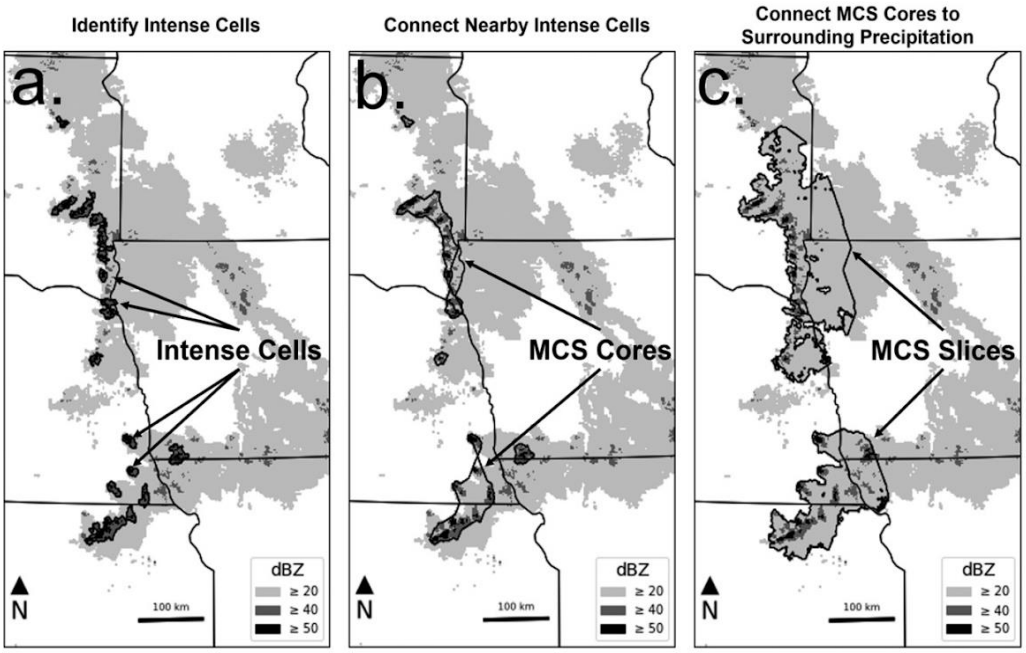

Fig. 2: Demonstration of the segmentation procedure for a candidate convective system. First, convective cells containing regions of intense convection are extracted. Then, they are connected if the cells are within the convective search radius to form "Cores." Finally, stratiform regions within the stratiform search radius of any given Core are joined to form the candidate "Slice." (Reproduced from Figure 3 of HA18.)

the study. This process was then repeated until 4,000 times with convective systems were identified.

b. Slice Extraction

From the images identified as containing convective systems, the automated procedures of HA18a were used to extract the candidate system "slices" for analysis (Fig. 2). This procedure uses the morphological operations and image processing of scikit-image (van der Walt et al. 2014) to identify and extract convective cells and their associated stratiform regions using a three-step process. First, convective cells are identified as those regions having ≥40 dBZ reflectivity (the "convective" threshold) over an area greater than 40 km^2 with at least one pixel of ≥50 dBZ reflectivity (the "intense" threshold). These convective cells were then merged into convective cores according to the convective search radius (48 km in this study) so that all cells within this distance of each other were joined into a single core. Finally, any adjacent reflectivity region meeting the stratiform threshold (≥20 dBZ) occurring within the stratiform search radius (192 km in this study) were merged with the cores to form the system slice (HA18a). If a single radar image contained multiple slices, the slices were categorized as separate systems for analysis. A total of 14,000 system slices were extracted from the sampled cases using these procedures.

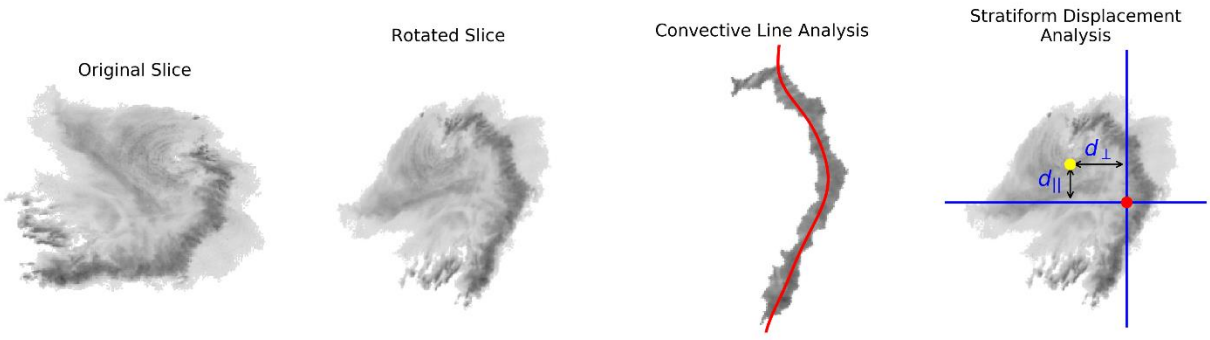

Fig. 3: Illustration of the characteristic curve and stratiform displacement determination process. The identified system slice is rotated so that the convective line is oriented in the vertical. Then, the smoothed characteristic curve is fit to the convective core as a function of distance along the line. The displacement of the stratiform region centroid from the convective core centroid is then split into line-parallel ($d_{||}$) and line-normal ($d_\perp$) components.

c. Parameter Identification

From the previously extracted system slice images, morphological and intensity parameters were derived to be able to categorize the systems. These included the 14 base areal, length, and intensity features of HA18a (Table 1a), as well as 24 additional parameters designed to potentially discriminate between the detailed convective modes of G08 (Table 1b). These additional parameters include convective length-width ratio (to which G08 assigned a minimum 3:1 threshold for the manual classification of linear systems), normal and parallel stratiform distances (based on the "mean stratiform distance" of Gange et al. (2009)), and mean and maximum signed curvature of the primary convective line (Fig. 3). Also included are several quantities relating to an analysis of the cellular structure of the system as represented by the mesh corresponding to the Delaunay triangulation of cells within the system (local maxima of reflectivity). The values of these parameters and the georeferenced system images were then saved for use by the machine learning models.

d. Manual Classification and Slice Filtering

To complete the preparation of training and testing data for the machine learning models, the system images were manually labeled according to the nine-category scheme of G08 and the extraneous non-MCS signatures identified in HA18a. The G08 scheme separates systems into three general types—cellular, linear, and non-linear—with further separation into three cellular subtypes—isolated cells (IC), clusters of cells (CC), and broken lines (BL)—and five linear subtypes—lines without a stratiform precipitation region (NS), bow echoes (BE), and lines with leading (LS), parallel (PS), and trailing (TS) stratiform regions (Fig. 4). In addition to MCS and unorganized convective complex types (which are considered in this study to be subsumed by the modes of G08), HA18a identified three other non-MCS signatures that the segmentation procedure

Table 1a: A list of the parameters extracted from the system slices and used in the decision tree-type classifiers that were taken from HA18.

Parameter	*Definition*
Total Area	Total area of system (km^2)
Intense Area	Area (km^2) of pixels meeting the intense criterion (≥50 dBZ)
Convective Area	Area (km^2) of pixels meeting the convective criterion (≥40 dBZ)
Intense-Total Area Ratio	Ratio of Intense Area to Total Area
Convective-Total Area Ratio	Ratio of Convective Area to Total Area
Intense-Convective Area Ratio	Ratio of Intense Area to Convective Area
Convex Area	Area of the convex hull of the system region (km^2)
Solidity	Ratio of Total Area to Convex Area
Major Axis Length	Length of the major axis of the ellipse best fitting the system region (km)
Minor Axis Length	Length of the minor axis of the ellipse best fitting the system region (km)
Eccentricity	Eccentricity of the ellipse best fitting the system region
Normal Stratiform Displacement*	Displacement of the stratiform region centroid from the centroid of the largest convective core measured perpendicular to the convective line orientation
Parallel Stratiform Displacement*	Displacement of the stratiform region centroid from the centroid of the largest convective core measured parallel to the convective line orientation
Mean Intensity	Mean value of reflectivity in the system region
Max Intensity	Maximum value of reflectivity in the system region
Intensity Variance	Variance of reflectivity values in the system region

may extract as a candidate system: ground clutter (CLT), synoptic systems (SYN), and tropical systems (TRP). Extracted systems fully depicted in the radar mosaic were manually classified according to these 12 labels, and systems that went outside the domain or occurred with bad radar scans were filtered out. Due to their relatively small frequency of occurrence in the sample of cases, signatures labeled as CLT were filtered out and those labeled as SYN or TRP were combined as a single SYN type. Of all the extracted system slices, a minimal sample of 3,000 cases was selected for use as training and testing data, with at least 130 slices present for each mode (Table 2).

Table 1b: Same as Table 1a, but for additional parameters added for this study.

Parameter		*Definition*
Convective Solidity*		Ratio of Convective Area to Convex Area
Normal Stratiform Displacement*		Displacement of the stratiform region centroid from the centroid of the largest convective core measured perpendicular to the convective line orientation
Parallel Stratiform Displacement*		Displacement of the stratiform region centroid from the centroid of the largest convective core measured parallel to the convective line orientation
Normalized Cell Count*		Number of convective cells divided by Total Area (count per km^2)
Mean Characteristic Curvature*		Mean signed curvature of the characteristic curve fit to the convective line
Max Characteristic Curvature*		Maximum signed curvature of the characteristic curve fit to the convective line
Convective Length*		Maximum length of the largest convective core of the system
Convective Width*		Average width of the largest convective core of the system
Convective Length-Width Ratio*		Ratio of Convective Length to Convective Width
Stratiform Width*		Average of the stratiform characteristic widths, which are the largest stratiform segments in each row of the normalized system image (segments taken in line-normal direction, averaged along line-parallel direction)
System-Convective Length Ratio*		Ratio of Major Axis Length to Convective Length
Stratiform-Convective Width Ratio*		Ratio of Stratiform Width to Convective Width
Delaunay Edges*		Number of edges in the Delaunay Mesh The Delaunay Mesh is the mesh corresponding to Delaunay triangulation of all cell centroids, which are taken as local maxima of reflectivity within a neighborhood of 15 km, given that the reflectivity meets the convective threshold
Edge Proportion with Minimum at*	None	Proportion of Delaunay Edges where minimum reflectivity along the edge is below the stratiform threshold
	Stratiform	Proportion of Delaunay Edges where minimum reflectivity along the edge is above the stratiform threshold but below the convective threshold
	Convective	Proportion of Delaunay Edges where minimum reflectivity along the edge is above the convective threshold
Edge Proportion with Average at*	None	Proportion of Delaunay Edges where average reflectivity along the edge is below the stratiform threshold
	Stratiform	Proportion of Delaunay Edges where average reflectivity along the edge is above the stratiform threshold but below the convective threshold
	Convective	Proportion of Delaunay Edges where average reflectivity along the edge is above the convective threshold but below the intense threshold
	Intense	Proportion of Delaunay Edges where average reflectivity along the edge is above the intense threshold
Edge Mean Length*		Mean length of the Delaunay Edges
Cell Centroid Spread*		Interquartile range of displacement of cell centroids (of the Delaunay Mesh) in the line-normal direction

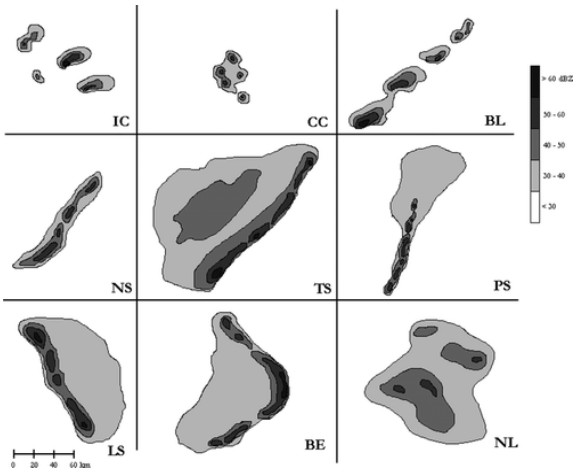

Fig. 4: The nine convective modes used in this study. (Reproduced from Figure 2 of G08.)

Table 2: Summary of the distributions of hand-labeled modes in the dataset supplied to the machine learning models, broken down by testing and training subsets.

Mode	*Testing (2004-2006)*	*Training (2007-2016)*	*Total*	*Percentage of Total*
IC	53	214	267	8.90
CC	86	333	419	13.97
BL	65	239	304	10.13
NS	39	174	213	7.10
LS	21	109	130	4.33
PS	35	189	224	7.47
TS	114	415	529	17.63
BE	80	299	379	12.63
NL	73	288	361	12.03
SYN	37	137	174	5.80
Total	603	2397	3000	

e. Machine Learning Classifiers

This study investigates two general automated techniques for classifying the extracted system slices according to the detailed modes of G08. The first uses the decision tree ensemble-based algorithms of HA18a. These classifiers are random forests (RFC; Pedregosa et al. 2011), gradient boosted trees (GBC; Pedregosa et al. 2011), and XGBoost (XGBC; Chen and Guestrin 2016), with an ensemble classifier (ENS) combining the three individual classifiers. This first approach extends these four classifiers by adding in the previously discussed new parameters (Fig. 1) in the model input and using all ten detailed labels. However, to demonstrate the effects of including the additional parameters, these four classifiers are also trained solely using the original 14 parameters of HA18a.

The second approach uses a CNN model configured in Keras (Chollet 2015). While the system slices varied greatly in extent, CNNs require fixed-size images as input and computational resource constraints limit the image sizes that are feasible. And so, three approaches were taken to obtain a characteristic image for each system. In the first (referred to as Scaled), the square region surrounding the reflectivity-weighted convective line centroid was extracted and then resized to 128 by 128 pixels by upscaling. If the system was small and no upscaling was required (i.e., both dimensions of the system were less than 256 km), the slice image was padded rather than upscaled to 128 by 128 pixels. In the second (referred to as Chopped), the 256 by 256 km area centered on the reflectivity-weighted

convective line centroid was taken directly, ignoring data outside that region, to obtain the 128 by 128 pixel image. In the third and final approach (referred to as 4km Chopped), the 512 by 512 km area centered on the reflectivity-weighted convective line centroid was taken, after which the data were upscaled from 2 km to 4 km grid spacing to obtain the 128 by 128 pixel image. Separate CNNs sharing the same structure were trained on these three image sets (Table 3) after applying data augmentation techniques that perturb the input image data by small random rotations and scalings to reduce model overfitting.

Model training data (for all model configurations tested) is taken from the 2007-2016 period and model testing data from the 2004-2006 period to assure independence of samples. While all the machine learning processes used in this study produce probabilities of classification for each label, the label with the highest probability was selected as the classifier's single result for each test case.

The standard classifier evaluation metrics of accuracy, precision, and recall are used to compare the models' performance with respect to the hand-labeled testing data. Accuracy is evaluated based on all model predictions and, in the context of this study, is simply the proportion of systems of the testing dataset the model classified correctly. Precision and recall are evaluated for each mode in each model and are computed as

$$Precision = \frac{TP_i}{TP_i + FP_i} \qquad (1)$$

$$Recall = \frac{TP_i}{TP_i + FN_i} \qquad (2),$$

where TP_i, FP_i, and FN_i refer to the count of true positives, false positives, and false negatives for a given mode, respectively. In general, precision refers to the proportion of predictions of a mode that were correct, and recall refers to the proportion of systems with a mode that were correctly predicted. Additionally, to evaluate any particular ways in which the models failed to classify systems correctly, plots of confusion matrices, which are heat maps of true vs. predicted classification, are shown.

3. Results and Analysis

a. Decision Tree Ensembles with Base Parameters

According to both overall accuracy (Table 4) and precision and recall for each mode (Table 5), all four decision tree ensemble classifiers trained using the 14 parameters of HA18a perform similarly, but with RFC performing slighter better than the others. For each classifier, the two modes of IC and SYN have relatively high precision and recall values on the order of 65-85%, signifying that the models perform relatively well in matching the manual classification. This is likely the result of these two modes having the most distinct signatures—SYN having extensive areas and low average intensity and IC having small areas and low solidity (area of cells is much less than the convex region bounding the cells). On the other hand, the two modes of LS and PS are the worst performing, with low precision and recall values ranging from 0-22%. This result is expected because discrimination of the two

Table 3: Configuration of the convolutional neural network tested in this study, which started with 128 by 128-pixel images and ended in probabilistic classifications of the ten categories. Table reproduced from Supplemental Table 1 of Haberlie and Ashley (2018b), with modification to reflect this study's 10 final categories and lack of batch normalization (which was removed due to computational resource constraints).

Layer Type	*# Features*	*Filter Size*	*Stride*	*Activation*	*Dropout*
Convolutional	64	7 x 7	--	ReLu	--
Convolutional	64	3 x 3	--	ReLu	--
Max Pooling	--	2 x 2	2 x 2	--	--
Convolutional	128	3 x 3	--	ReLu	--
Convolutional	128	3 x 3	--	ReLu	--
Max Pooling	--	2 x 2	2 x 2	--	--
Convolutional	256	3 x 3	--	ReLu	--
Convolutional	256	3 x 3	--	ReLu	--
Max Pooling	--	2 x 2	2 x 2	--	--
Convolutional	512	3 x 3	--	ReLu	--
Convolutional	512	3 x 3	--	ReLu	--
Max Pooling	--	2 x 2	2 x 2	--	--
Flatten	--	--	--	--	--
Dense	4096	--	--	ReLu	Dropout (0.3)
Dense	4096	--	--	ReLu	Dropout (0.3)
Dense	10	--	--	Softmax	--

modes require a way to determine if the position of the dominant stratiform region is ahead of or parallel to the dominant convective line, and none of the parameters of HA18a would do so.

Analysis of the confusion matrices for these classifiers again shows relative agreement (Fig. 5). In these diagrams, entries along the major diagonal represent correct classifications, and off-diagonal entries are incorrect classifications. The diagonal entries for all modes except LS, NS, and PS show

Table 4: Comparison of overall accuracy scores for each of the ten models evaluated in this study, with best performing model bolded and worst performing italicized.

Model		*Accuracy*
RFC	14 Parameters (Table 1a)	0.47761
GBC	14 Parameters (Table 1a)	0.43615
XGBC	14 Parameters (Table 1a)	0.46932
ENS	14 Parameters (Table 1a)	0.45439
RFC	38 Parameters (Table 1a,b)	0.57048
GBC	38 Parameters (Table 1a,b)	0.56551
XGBC	38 Parameters (Table 1a,b)	0.56053
ENS	38 Parameters (Table 1a,b)	0.59370
CNN (Table 3)	Scaled Images	0.44776
CNN (Table 3)	Chopped Images	0.30846
CNN (Table 3)	4km Chopped Images	0.38143

Table 5: Comparison of the precision and recall for the ten modes for each of the decision tree ensemble models utilizing the original 14 parameters of HA18a.

	RFC		***GBC***		***XGBC***		***ENS***	
Mode	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
BE	0.42	0.45	0.3	0.33	0.34	0.35	0.31	0.33
BL	0.5	0.4	0.54	0.43	0.52	0.51	0.49	0.45
CC	0.49	0.55	0.49	0.58	0.45	0.52	0.49	0.57
IC	0.7	0.85	0.68	0.72	0.69	0.77	0.68	0.74
LS	0	0	0	0	0	0	0	0
NL	0.54	0.51	0.49	0.47	0.55	0.48	0.51	0.48
NS	0.34	0.36	0.4	0.46	0.46	0.49	0.45	0.49
PS	0.15	0.09	0.13	0.11	0.22	0.17	0.17	0.11
SYN	0.82	0.84	0.81	0.68	0.73	0.73	0.78	0.76
TS	0.37	0.43	0.33	0.35	0.4	0.43	0.35	0.39

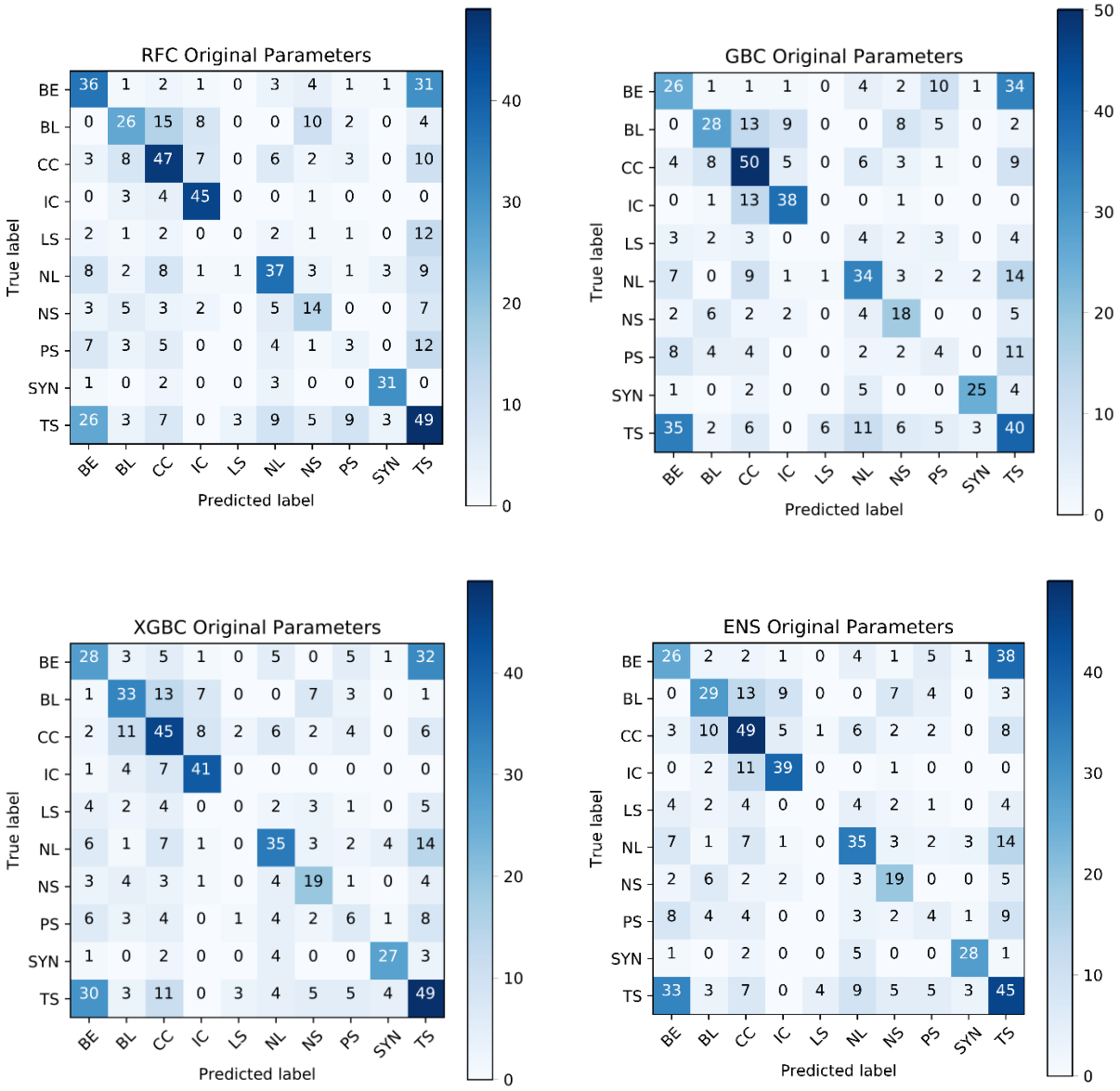

Fig. 5: Confusion matrices for the decision tree ensemble models utilizing the original 14 parameters of HA18a. True mode label is depicted along the y-axis and predicted mode label along the x-axis, with cell shading and indicated number representing the count of the number of systems in the testing dataset having that pair of true and predicted mode labels.

high counts, indicating the classifiers had at least some skill in recognizing those modes. However, some important off-diagonal entries also have high counts. First, the TS true-BE predicted and BE true-TS predicted entries have counts comparable to the correct BE and TS entries. This indicates the classifiers cannot sufficiently discriminate between these two modes, which is an expected result due to the lack of convective line bowing information in the model input to tell bow echoes apart from lines with trailing

stratiform regions. Also, several other entries along the true TS row and predicted TS column have high counts, indicating that the classifiers do relatively poorly overall in handling this mode. Finally, the high counts in the off-diagonal entries of the cellular modes (BL, CC, and IC) indicate that the classifiers also struggle with discriminating between the cellular subtypes.

b. Decision Tree Ensembles with All Parameters

Similar to the classifiers using the base parameters, according to both overall accuracy (Table 4) and precision and recall for each mode (Table 6), the four decision tree ensemble classifiers trained using the full set of 38 parameters perform similarly. In this case, however, ENS performs slighter better than the others. For each classifier, the two modes of IC and SYN remain the best performing with high precision and recall values on the order of 75-90%, which is expected because the parameters used in these classifiers is a superset of those used in the past set of classifiers. The performance in regards to all the other modes generally increased. Despite this, LS and PS continue to have problems, with precision and recall values generally below 50% and as low as 17% (RFC correctly predicts only 17% of the occurrences of PS). This indicates that the added stratiform displacement parameters are still insufficient to numerically determine the dominant position of the stratiform region on a consistent basis. Additionally, a general usefulness threshold for these precision and recall values is around 90%. If the classifier either has more than 10% of its classifications of a mode disagree with the manual labeling (<90% precision) or misses more than 10% of mode occurrence from the manual labeling (<90% recall), it is difficult to rely on the automated classifier for accurate classification. In the current configuration of these four classifiers, none of the precision or recall values meet this threshold.

Table 6: Comparison of the precision and recall for the ten modes for each of the decision tree ensemble models utilizing the full set of 36 parameters.

	RFC		***GBC***		***XGBC***		***ENS***	
Mode	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
BE	0.52	0.56	0.48	0.47	0.46	0.47	0.5	0.5
BL	0.62	0.46	0.57	0.49	0.62	0.52	0.65	0.54
CC	0.58	0.59	0.54	0.56	0.54	0.58	0.57	0.62
IC	0.77	0.87	0.78	0.81	0.75	0.77	0.79	0.83
LS	0.33	0.24	0.44	0.52	0.38	0.43	0.46	0.52
NL	0.55	0.53	0.54	0.55	0.58	0.58	0.58	0.59
NS	0.48	0.59	0.52	0.56	0.53	0.59	0.54	0.64
PS	0.33	0.17	0.38	0.37	0.4	0.34	0.41	0.37
SYN	0.86	0.81	0.85	0.78	0.82	0.84	0.88	0.81
TS	0.51	0.61	0.57	0.57	0.53	0.51	0.58	0.56

Similar to before, analysis of the confusion matrices for these classifiers again shows relative agreement (Fig. 6). Now, all modes show relatively high counts along the diagonal, indicating the classifiers had at least some skill in recognizing all the modes. Also, the counts of off-diagonal entries have decreased from the previous set of classifiers, demonstrating an overall increase in skill, which agrees with the increase in overall accuracy score (Table 4). However, some problems remain. First, while less severe, the confusion between TS and BE is still present with mismatch counts around 25 for each type. This means that the additional convective line curvature parameters have helped, but are still insufficient, to numerically discriminate between the bowing convective line of BE and the non-bowing convective line of TS. Also, some moderately high counts in the off-diagonal entries of the cellular modes (BL, CC, and IC) remain, indicating that added Delaunay mesh parameters are also insufficient to fully discriminate between the cellular modes. Finally, CC and NL remain sometimes confused by the classifiers. Given the subjective division between these two modes when it comes to manual labeling, it may be the case that this is simply an artifact of ambiguity in the manual labeling.

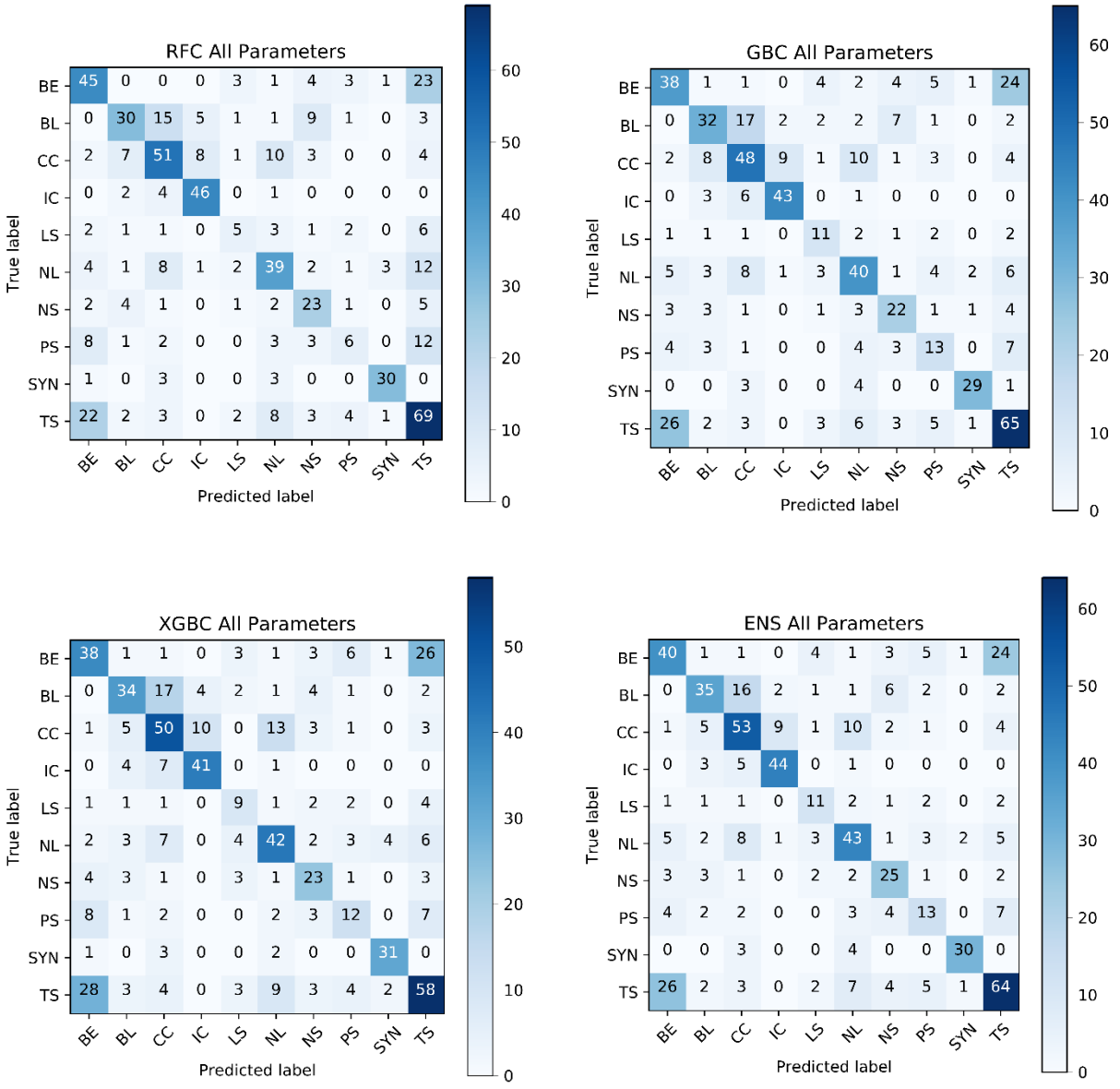

Fig. 6: Same as Fig. 5, but for the decision tree ensemble models utilizing the full set of 38 parameters.

c. Convolutional Neural Networks

Contrary to the initially expected results hypothesized on the basis of past studies, the convolutional neural networks perform generally worse than the decision tree ensemble methods, with accuracy scores near or below those of the HA18a parameter set classifiers (Table 4). The precision and recall values are correspondingly poor, except for a few cases such as the values for IC (Table 7). The Chopped Image CNN was especially poor performing, with the lowest accuracy score (Table 4) and a low number of counts along the diagonal of the confusion matrix (Figure 7). In particular, for the Chopped

Table 7: Comparison of the precision and recall for the ten modes for convolutional neural network classifier.

	Scaled Image CNN		***Chopped Image CNN***		***4km Chopped Image CNN***	
Mode	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
BE	0.46	0.53	0.3	0.61	0.4	0.68
BL	0.4	0.15	0.24	0.09	0	0
CC	0.43	0.62	0.38	0.35	0.5	0.15
IC	0.64	0.77	1	0.23	0.84	0.6
LS	0.33	0.29	0.25	0.29	0.32	0.48
NL	0.38	0.68	0.43	0.14	0.4	0.36
NS	0.38	0.38	0.09	0.08	0.16	0.62
PS	0.33	0.14	0.18	0.23	0.47	0.4
SYN	1	0.03	0.29	0.3	0.53	0.24
TS	0.49	0.41	0.31	0.45	0.45	0.42

configuration, TS and BE have become "default" classifications where many systems are incorrectly labeled as TS or BE, as demonstrated by high counts in the predicted columns for those modes in the confusion matrix. While the Scaled Image CNN improves upon this mode depiction and attains the highest accuracy score of the CNNs (44.78%) with several high counts along the diagonal of the confusion matrix (BE, CC, IC, NL, and TS), significant problems remain. BE and TS remain commonly confused, BL systems are often misclassified as CC or NS, and NL is predicted in many cases where it should not have been (when the true mode was CC, SYN, or TS). Similar to the Scaled configuration, the 4km Chopped Image CNN demonstrates some skill through several high counts along the confusion matrix diagonal (in modes such as BE, IC, NL, NS, and TS). However, this configuration has a substantial problem with over-prediction of NS as many high counts exist in the NS-predicted column (especially for true BL and CC). This, combined with the common result of CC being often confused for NL and TS likewise for BE, made the overall accuracy relatively low.

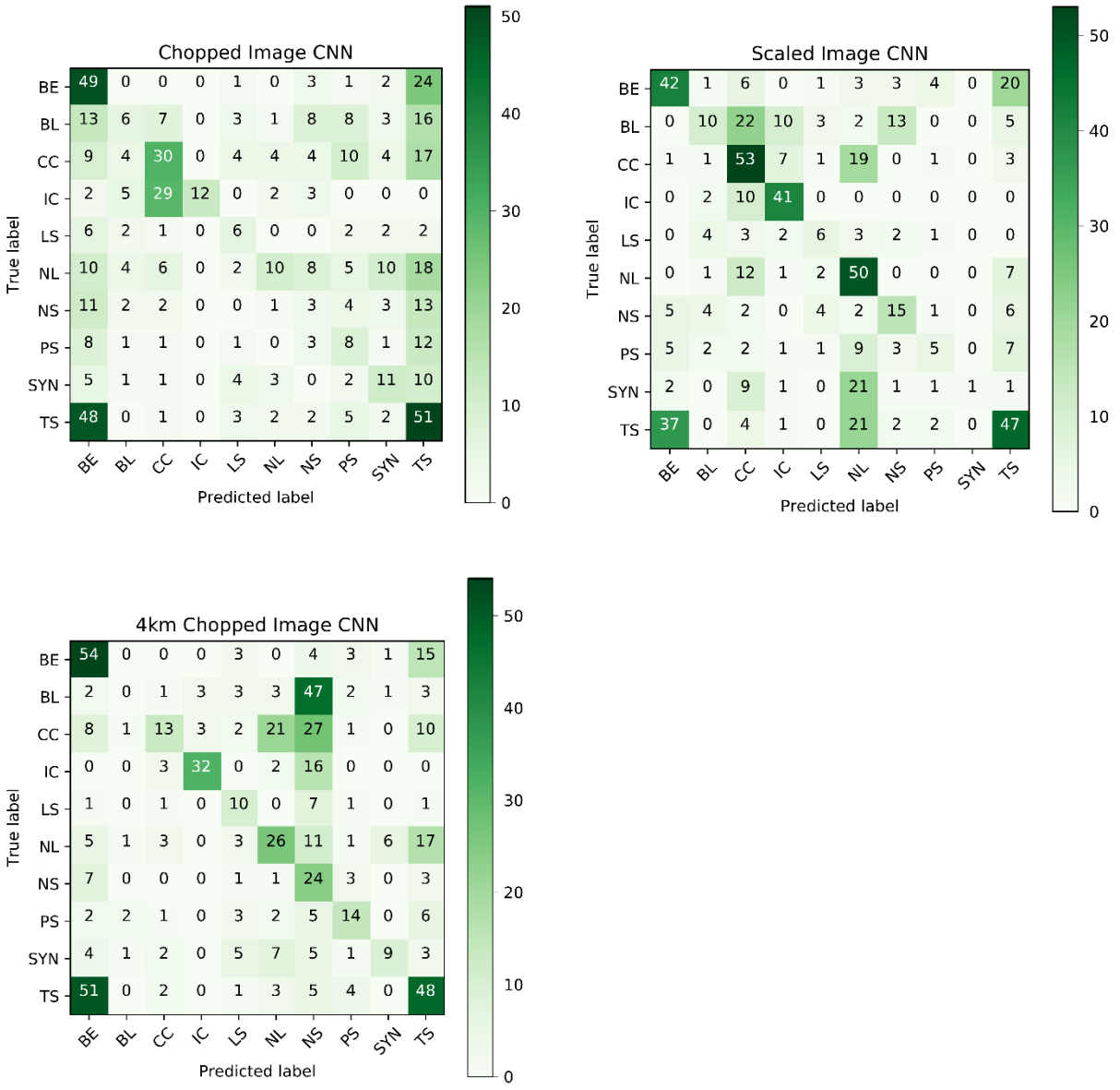

Fig. 7: Same as Fig. 5, but for the three convolutional neural network models.

4. Discussion and Conclusion

This study sought to evaluate and compare two general machine learning techniques—decision tree ensembles and convolutional neural networks—in their ability to classify convective morphology according to the detailed scheme of G08. A secondary aim was determining if either would be reliable enough for use as an automated procedure for future studies of convective morphology. While the current results indicate that the decision tree ensembles utilizing a large set of morphological parameters outperform the CNNs, and that neither attain sufficiently high accuracy, precision, and recall, these findings are highly contingent on the present configurations of the classifiers.

As demonstrated by the increase in accuracy, precision, and recall, and better subjective appearance of the confusion matrices, adding the additional morphology parameters to the decision tree ensembles aided their performance, thereby showing promise for this technique. However, some critical issues such as CC/NL and BE/TS confusion, LS and PS inaccuracies, and insufficiently high metrics remain. These issues are critical to resolve because the differences in mode result in different implications about convective system properties and potential hazards.

As previously discussed, one reason for these problems may be inadequacies in the more complex added parameters. And so, improvements to the convective line curvature, stratiform displacement, and cell connectivity parameters should be investigated. Additionally, given the relatively large number of output classes (ten), the input sample may have been too small. This is especially clear in the LS and PS modes, which are rare relative to the other eight and therefore had poor detection and reliability in the classifiers. Future work should expand the sample of labeled systems for model input and investigate if this results in model improvements. Finally, an important caveat to all supervised machine learning is demonstrated in these decision tree ensemble results. Since the model is trained on manually labeled input data, any biases or ambiguities in subjective classification are likely to show up as model errors. For instance, given the subjectivity involved in determining between CC (clusters of cells) and NL (non-linear) in many borderline cases, the classifier confusion between those two modes is not surprising. And so, future work may need to more carefully consider what separates convective modes and more clearly specify what is and what is not a particular mode.

While past studies have indicated that CNNs tend to perform better than decision tree methods in many subjective image classification tasks, this research unexpectedly found the opposite with its tested CNN configurations. However, this result does not yet stand in contradiction to those past findings for several reasons.

First and foremost, only variations on a single neural network structure were evaluated, and since many other possible structures exist, additional structures must be evaluated before a general conclusion can be drawn. Alongside this, the computational resources available for this study also placed limits on the robustness and complexity of the CNN structure, and so, increased computational power will be required for future work in this area.

Additionally, within the context of CNN studies, this work's 3,000 input sample size for ten output classes is rather small, even with the data augmentation procedures used. Along with the performance increase that occurred as the input sample size increased in the process of model development, this suggests that the current input data may have been insufficient to train the neural networks properly.

Finally, each of the image preparation methods may have eliminated important information about system characteristics. Both scaling process likely eliminated details

of larger systems and blurred out important gradients of reflectivity, and the full scaling removed information about the true scale of the system. The process of chopping the data to the 256 by 256 km box in the Chopped method also eliminated system data outside that region, which resulted in poor accuracy given that many systems were larger than 256 by 256 km. Future work will need to investigate if larger image sizes or other image preparation techniques (such as those that take into account the thresholds specified in slice extraction) could make the fixed-size input images supply sufficient morphology depiction to the model. It remains to be seen if improvements in these areas of model structure, sample size, and image preparation will be sufficient to make CNNs a reliable method for this particular application or if other machine learning methods are more appropriate.

Acknowledgments. The author would like to sincerely thank Dr. William Gallus for his mentorship and input on convective morphology throughout this project and Dr. Alex Haberlie for providing the original slice extraction code and for his guidance on the use of machine learning procedures. Portions of this work were also inspired by the machine learning sessions at the 2018 Unidata Users Workshop. The author would also to like to thank Melissa Piper for assistance in filtering cases for the occurrence of convective systems. Computational resources for the convolutional neural networks used in this study were provided by Google Colaboratory.

REFERENCES

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.

Biggerstaff, M. I., and S. A. Listemaa, 2000: An Improved Scheme for Convective/Stratiform Echo Classification Using Radar Reflectivity. *J. Appl. Meteor.*, **39**, 2129–2150, https://doi.org/10.1175/1520-0450(2001)040%3C2129:AISFCS%3E2.0.CO;2.

Bowman, K. P., and C. R. Homeyer, 2017: GridRad - Three-Dimensional Gridded NEXRAD WSR-88D Radar Data. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO. [Available online at https://doi.org/10.5065/D6NK3CR7.] Accessed 2 September 2018.

Chollet, F., 2015: Keras. [Available online at https://keras.io.] Accessed 2 September 2018.

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.

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.

Fowle, M. A. and P. J. Roebber, 2003: Short-Range (0–48 h) Numerical Prediction of Convective Occurrence, Mode, and Location. *Wea. Forecasting*, **18**, 782–794, https://doi.org/10.1175/1520-0434(2003)018<0782:SHNPOC>2.0.CO;2.

Fritsch, J. M., R. J. Kane, and C. R. Chelius, 1986: The Contribution of Mesoscale Convective Weather Systems to the Warm-Season Precipitation in the United States. *J. Climate Appl. Meteor.*, **25**, 1333–1345, https://doi.org/10.1175/1520-0450(1986)025<1333:TCOMCW>2.0.CO;2.

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.

Gallus, W. A., 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., 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, 2018: A Method for Identifying Midlatitude Mesoscale Convective Systems in Radar Mosaics. Part I: Segmentation and Classification. *J. Appl. Meteor. Climatol.*, **57**, 1575–1598, https://doi.org/10.1175/JAMC-D-17-0293.1.

Haberlie, A. M. and W. S. Ashley, 2018: Climatological Representation of Mesoscale Convective Systems in a Dynamically Downscaled Climate Simulation. *Int. J. Climatol.*, In Press, https://doi.org/10.1002/joc.5880.

Herman, G. R., and R. S. Schumacher, 2018: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev.*, **146**, 1571-1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Herman, G. R., and R. S. Schumacher, 2018: "Dendrology" in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation. *Mon. Wea. Rev.*, **146**, 1785-1812, https://doi.org/10.1175/MWR-D-17-

0307.1.

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

Jirak, I.L. and W.R. Cotton, 2007: Observational Analysis of the Predictability of Mesoscale Convective Systems. *Wea. Forecasting*, **22**, 813–838, https://doi.org/10.1175/WAF1012.1.

Kamani, M. M., F. Farhat, S. Wistar, and J. Z. Wang, 2017: Skeleton matching with applications in severe weather detection. *Appl. Soft Comput.*, **70**, 1154-1166, https://doi.org/10.1016/j.asoc.2017.05.037.

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. Research*, **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.

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

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

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine Learning in Python. *J. Machine Learning Research*, **12**, 2825–2830.

Thielen, J. E., W. A. Gallus, and B. J. Squitieri, 2018: Microphysical and Horizontal Grid Spacing Influences on WRF Forecasts of Stratiform Rain Regions and General Convective Morphology Evolution in Nocturnal MCSs. *25th Conf. on Numer. Wea. Predict.*, Denver, CO, Amer. Meteor. Soc., 10B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper344877.html.

van der Walt, S., J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, 2014: scikit-image: image processing in Python. *PeerJ*, **2:e453,** https://doi.org/10.7717/peerj.453.