Diagnosing Storm Mode with Deep Learning in Convection-Allowing Models

Ryan A. Sobash, David John Gagne II, Charlie L. Becker, David Ahijevych, Gabrielle N. Gantos, Craig S. Schwartz

National Center for Atmospheric Research, Boulder, CO

Corresponding author: Ryan A. Sobash, sobash@ucar.edu

Early Online Release: This preliminary version has been accepted for publication in *Monthly Weather Review*, may be fully cited, and has been assigned DOI 10.1175/MWR-D-22-0342.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: While convective storm mode is explicitly depicted in convection-allowing model (CAM) output, subjectively diagnosing mode in large volumes of CAM forecasts can be burdensome. In this work, four machine learning (ML) models were trained to probabilistically classify CAM storms into one of three modes: supercells, quasi-linear convective systems, and disorganized convection. The four ML models included a dense neural network (DNN), logistic regression (LR), a convolutional neural network (CNN) and semi-supervised CNN-Gaussian mixture model (GMM). The DNN, CNN, and LR were trained with a set of hand-labeled CAM storms, while the semi-supervised GMM used updraft helicity and storm size to generate clusters which were then hand labeled. When evaluated using storms withheld from training, the four classifiers had similar ability to discriminate between modes, but the GMM had worse calibration. The DNN and LR had similar objective performance to the CNN, suggesting that CNN-based methods may not be needed for mode classification tasks. The mode classifications from all four classifiers successfully approximated the known climatology of modes in the U.S., including a maximum in supercell occurrence in the U.S. Central Plains. Further, the modes also occurred in environments recognized to support the three different storm morphologies. Finally, storm mode provided useful information about hazard type, e.g., storm reports were most likely with supercells, further supporting the efficacy of the classifiers. Future applications, including the use of objective CAM mode classifications as a novel predictor in ML systems, could potentially lead to improved forecasts of convective hazards.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

SIGNIFICANCE STATEMENT: Whether a thunderstorm produces hazards such as tornadoes, hail, or intense wind gusts is in part determined by whether the storm takes the form of single cell or a line. Numerical forecasting models can now provide forecasts that depict this structure. We tested several automated algorithms to extract this information from forecast output using machine learning. All of the automated methods were able to distinguish between a set of three convective types, with the simple techniques providing similarly skilled classifications compared to the complex approaches. The automated classifications also successfully discriminated between thunderstorm hazards, potentially leading to new forecast tools and better forecasts of high-impact convective hazards.

1. Introduction

The environmental and dynamical factors that determine storm mode (e.g., whether convection is organized into cells or lines) also influence the likelihood that storms will produce tornadoes, hail, and straight-line winds (Smith et al. 2012; Thompson et al. 2012). While environmental properties such as deep-layer wind shear and instability have been traditionally used to diagnose convective mode and hazards (Thompson et al. 2003), more recent NWP model configurations that partially resolve convection (i.e., convection-allowing models, or CAMs) can explicitly depict convective mode and in turn the potential for hazards, giving forecasters a macro-scale signal of the potential for unresolved, micro-scale hazards at high spatial and temporal resolution. For example, the presence of supercell storms in CAMs may suggest a higher chance of tornadoes and severe hail, while bowing line segments indicate a higher potential of severe winds (Smith et al. 2012).

Subjectively diagnosing the convective mode simulated by CAMs presents multiple challenges stemming from the amount of information to process and the subjective nature of mode classification. A forecaster can often quickly assess the mode of storms in a deterministic forecast, but doing so becomes burdensome when interrogating CAM ensembles (Jergensen et al. 2020), whose size, integration length, and update frequency are likely to increase in the future (Stensrud et al. 2009). Further, convective mode exists along a spectrum, resulting in storms that do not fit cleanly into existing classification systems (Gallus et al. 2008), necessitating the use of probabilities to represent uncertainty. Automated convective mode diagnosis systems propose to solve both of these issues by applying a consistent identification process across datasets as large as computing and

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

latency requirements allow. Objective mode classification could also benefit other problems, such as the identification of different convective modes in convection-allowing climate model output (e.g., Prein et al. 2015).

Multiple automated convective mode identification systems have been developed over the past 20 years utilizing a mixture of heuristic and machine learning (ML) approaches. The vast majority of this work has focused on extracting mode from observed radar observations. For example, Gagne et al. (2009) evaluated the performance of decision trees and other ML methods for discriminating between cellular and linear storm modes, while other decision-tree-based and ML-based systems have also been developed (Lakshmanan et al. 2010; Lack and Fox 2012; Kolodziej-Hobson et al. 2012; Jergensen et al. 2020). The output from the storm mode classification algorithms have been used in various ways, including examining mode in relation to National Weather Service warning performance (Guillot et al. 2008; Brotzge et al. 2013). While these studies have identified methods that excel at classifying storms in radar imagery, few studies have applied ML techniques to classify storms in CAMs (e.g., Potvin et al. 2022).

One of the main limiting factors for building ML classification systems is the need for large amounts of labeled training data. Gallus et al. (2008) hand-labeled nearly 1000 storms by reviewing radar images of the same storm at multiple times and subjectively determined the primary storm morphology for that period. Trapp et al. (2005b) and Smith et al. (2012) focused their hand-labeling efforts to storms that produced severe weather reports. Doing so increases the likelihood of labeling more mature storms of different modes, but biases the sampled datasets toward supercell modes. If manual labeling is performed as part of existing forecaster duties, then those labels can be used directly for ML training, as in the case of frontal analysis (Biard and Kunkel 2019; Lagerquist et al. 2019). While some large scale ML morphology projects have crowd-sourced the labeling of their images (Lintott et al. 2011), distinguishing between convective modes requires some degree of expert knowledge. Additional factors, such as inconsistencies between human labels of the same storm, and potential changes in the labeled training data (e.g., due to frequent NWP model upgrades) make creating robust labeled datasets challenging.

With the limitations of hand-labeling in mind, scientists have developed different classes of approaches to minimize the burden of creating a training dataset. One solution is to build a non-ML heuristic or rule-based automated algorithm to identify mode (Lakshmanan and Smith

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

2009; Potvin et al. 2022). For instance, the size, shape, and intensity properties of storm objects can be used to infer storm mode given a set of rules. In Potvin et al. (2022), a rule-based system provided useful classifications of mode in observed radar and CAM output using only gridded reflectivity and rotation fields. While these techniques may be more transparent than fully trained ML models, they often can only use a small number of input fields and typically do not provide uncertainty estimates. Other approaches aim to reduce the dimensionality of the dataset to be labeled prior to hand labeling, for example with techniques such as principle component analysis or using convolutional neural networks (CNNs) to extract relevant features from input imagery (e.g., Gagne et al. 2019). In this "representational learning" paradigm, hand-labels would only be needed to distinguish between modes in the low-dimensional space, rather than across thousands of individual storms.

Here, we evaluate four different ML and deep learning paradigms to determine how well they can diagnose the mode of simulated storms in a dataset of approximately 500 CAM forecasts over the contiguous United States, covering the next-day time period (i.e., forecast lead-times between 12 - 36 hours). First, a convective mode training dataset was constructed by hand-labelling thousands of storms present within a subset of the CAM forecasts. Then, a set of four ML algorithms were trained: a logistic regression, a dense neural network, a CNN, and a semi-supervised CNN/clustering-based algorithm. These four algorithms differ in the type of input fields required (e.g., scalar object-based properties such as size vs. two-dimensional imagery such as reflectivity), as well as in their need for training data. The CNN/clustering-based approach uses representational learning to learn features and cluster with less human input compared to the three fully supervised methods. To evaluate the output from the classification systems, we use multiple evaluation approaches, including comparisons between the predicted convective mode and the "ground-truth" hand-labels in withheld training data, as well as using severe storm reports to determine if the mode predictions are useful in anticipating convective hazards. Together, these inter-comparisons and evaluations provide a set of best practices for using ML to classify storm mode in CAM output.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

2. Data and Methods

a. Convection-allowing model forecast dataset

All of the ML algorithms examined in this work use a set of convective storm objects extracted from deterministic forecasts within a Weather Research and Forecasting (WRF)-based convection-allowing model reforecast dataset generated at the National Center for Atmospheric Research (NCAR). This forecast dataset (henceforth referred to as the NCAR-WRF) has been used in several prior studies (Sobash et al. 2019, 2020; Schwartz and Sobash 2019) to investigate the predictability and prediction of high-impact convective weather events over the CONUS, including the application of ML algorithms for convective hazard prediction (Sobash et al. 2020). The original NCAR-WRF dataset in Sobash et al. (2019) included 497 severe weather events occurring between 15 October 2010 and 15 July 2017 that were selected based on their inclusion in the Storm Prediction Center (SPC) severe weather event archive. Configurations for all NCAR-WRF forecasts, including physics choices and a list of the original set of simulated 2010 – 2017 severe weather events, is described in Sobash et al. (2019). Some notable WRF model configuration choices include the use of WRF version 3.6.1, 0000 UTC operational GFS analyses and forecasts as initial and boundary conditions, a computational domain spanning the entire CONUS with 3-km horizontal grid spacing, and a 36-hour forecast integration length. In this work, we only use the 2010 – 2016 NCAR-WRF forecasts, as described below.

b. Object segmentation and labeling

1) Object segmentation and patch extraction

Storm objects were identified in the NCAR-WRF forecasts using the hysteresis segmentation method (Lakshmanan et al. 2009) in the hagelslag Python package (Gagne II et al. 2017). In hysteresis, storm centers are first identified as contiguous areas that exceed a simulated composite radar reflectivity (CREF) threshold of 50 dBZ. The hysteresis algorithm then expands these storm centers in order from most intense to least intense by incorporating contiguous grid cells with CREF > 35 dBZ. The hysteresis approach was preferred over other segmentation techniques (e.g., enhanced watershed) because it keeps organized systems, such as mesoscale convective systems and squall lines, together rather than segmenting them into many smaller objects.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

Storm object attributes (e.g., object size, shape, within-storm and near-storm environment properties) and two-dimensional storm patches were extracted from the NCAR-WRF dataset using CREF output between forecast hours 12 and 35 (the first 12 hours of the forecast were not used to reduce issues with model spin-up). The majority of storm objects were disorganized storms or regions of a storm not including the convective cores (e.g., the stratiform area within a mesoscale convective system). While storm objects were extracted for all 2010–2016 NCAR-WRF forecasts, only objects within the 2013 subset of NCAR-WRF forecasts that contained at least one grid point with the magnitude of the hourly-maximum 2km – 5km above ground level (AGL) updraft helicity (UH) $\geq$ 25 m^2/s^2were considered for hand labeling (including storms with both cyclonic and anticyclonic rotation). Applying these criteria resulted in a sample of ~11,000 storms to be hand labeled. This UH threshold was applied to focus on the most intense convective storms, although this threshold is low enough to include both disorganized and QLCS modes as well that may not contain appreciable UH.

2) Hand labeling interface

A web interface was developed to assist with the hand labeling process (Fig. 1). The interface provided users with three different types of plots and buttons to select both the storm mode and their confidence in the classification. The images included simulated CREF with swaths of high magnitude UH over the previous hour ($|UH| \geq 50m^2/s^2$) shaded in gray, 2-m temperature with 10-m wind barbs, and most-unstable CAPE with 0 – 6 km AGL wind shear barbs which together provide enough information to make a determination of the convective mode. Each image was centered on the centroid of the storm object and the object to be labeled was contoured in black to distinguish between the storm of interest and other storms within the image domain. The user was free to toggle between plots valid for the forecast hour of the storm object, as well as the one and two previous and following forecast hours.

Possible labels for storm mode fell into three primary categories: (D)isorganized, (S)upercell, and (Q)uasi-linear. These categories were further divided into seven subtypes: disorganized cell (D1), disorganized cluster (D2), isolated supercell (S1), supercell(s) within a line (S2), supercell(s) in a cluster (S3), well-developed bow echo (Q1), or squall line or smaller-scale convective line (Q2). These seven categories are similar to the primary modes in Smith et al. (2012). Experts also

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

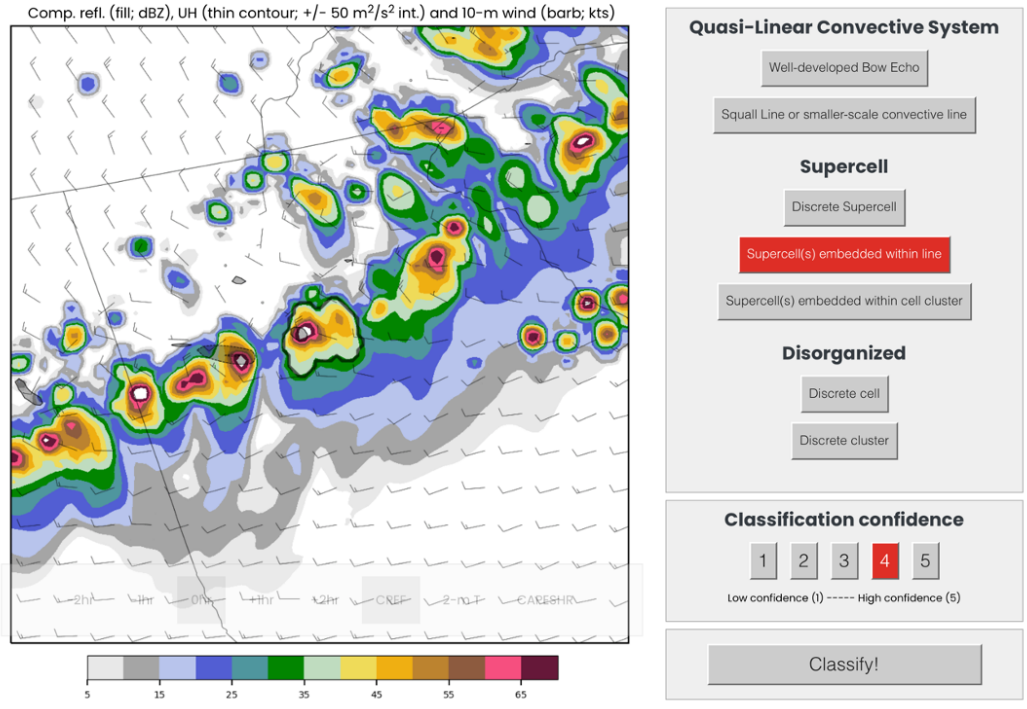

FIG. 1. Web interface for mode classification. The storm to be classified is contoured in black in the center of the CREF plot. An overlay provided the human classifier the option to toggle between imagery valid at 0, 1, and 2 hours before and after the central time for three fields: CREF/10-m wind, 2-m temperature/10-m wind, and CAPE/deep layer (0 – 6km AGL) shear. Once the images were interrogated, the human classifier selected one of the 7 modes and a confidence rating.

quantified how confident they were in each storm label by providing an integer between one and five, with one being the least confident and five being the most. Through the web interface, each storm's label and confidence rating were stored in a database, although the confidence ratings are not used in this work.

Several labelling campaigns were completed with 10 different experts contributing to the labeling, including the authors, NCAR scientists, and two severe weather forecasters from the SPC. The interface was designed such that a user would login and see a random storm from a batch of 500 storms. Each batch consisted of a random collection of storms among the reduced set of ~11,000 storms in the 2013 NCAR-WRF forecasts. Initially, each storm was labeled five times to provide an indication of uncertainty in the classification. Once all 500 storms received five labels, storms from the next batch would be shown to the human classifiers. This constraint was later relaxed so

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

each storm was only labeled by one human in order to obtain a larger diversity of storms. To be consistent, only the first hand label was used for all storms. Among the subset of storms with 5 labels, 90% had a consensus label for the primary category (i.e., three experts in agreement); there was more disagreement when considering the subtypes.

The final hand labeled storm mode dataset consisted of 2,627 storms. Roughly 58% of these storms were labeled as disorganized, while 26% and 16% were labeled as supercells and quasi-linear convective systems (QLCSs), respectively. The full labeled dataset of storms from the 2013 NCAR-WRF forecasts was split such that storms in forecasts initialized between 1 January 2013 – 24 June 2013 were used for training the supervised ML models (1,871 storms), while forecasts initialized between 25 June 2013 – 31 December 2013 were used for testing (756 storms). All ML models were trained to predict the three primary categories of modes, given that the differences between the subtypes were more subtle and some subtypes only contained a small sample of storms (e.g., the Q1 category contained only 21 storms). Finally, given the similarities between the S2 category and the Q1 and Q2 categories, we chose to include the S2 storms in the QLCS category, this produced more skillful predictions of QLCSs without impacting the skill of the other two primary categories.

3) Properties of hand labeled storms

The distributions of object and storm intensity properties were examined in the set of 2,627 hand-labeled storms to assess how the human classifications matched the intuition of how various properties associated with each mode should behave (Fig. 2). For example, the storms identified as supercells tended to have larger object maximum UH and updraft speeds than QLCSs and disorganized storms (Fig. 2a, b). The median object maximum UH value for supercells was 88 m^2/s^2, while for the disorganized cells the median object maximum UH was 33 m^2/s^2. The storms labeled as QLCSs had larger object maximum surface wind speeds, areas, and major axis lengths than the supercells and disorganized storms, and tended to occur slightly later in the day (Fig. 2c-f). For example, the median object forecast hours were 25, 24, and 23 for QLCSs, supercells, and disorganized storms, respectively, and the QLCS distribution had a larger fraction of storms occurring at forecast hours > 26 compared to the supercells and disorganized storm distributions.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

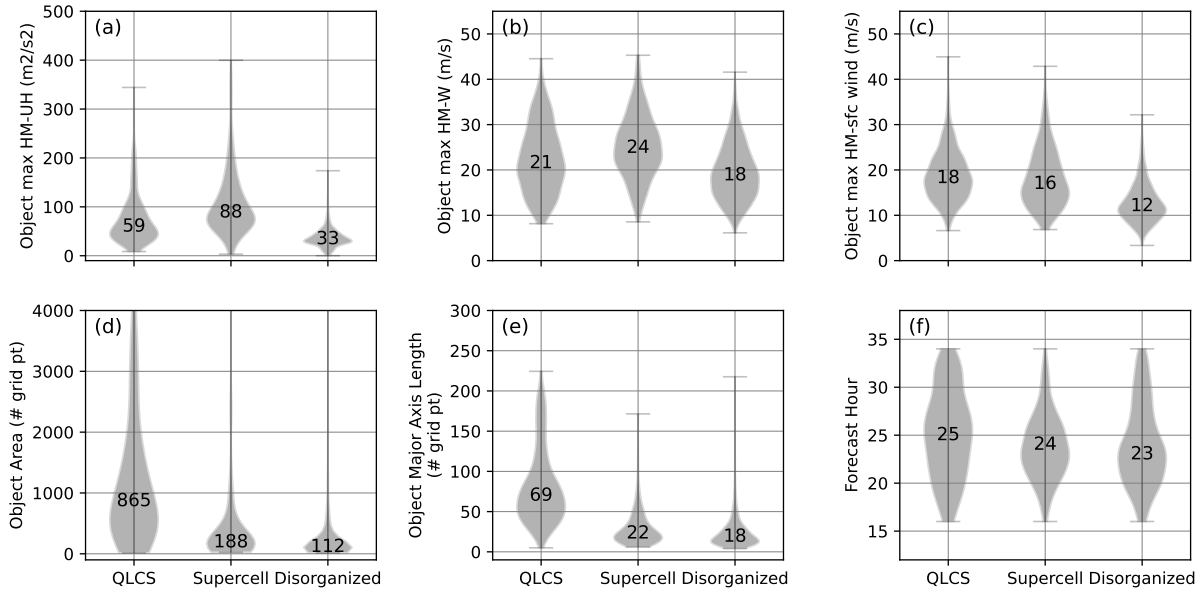

FIG. 2. Violin plots consisting of kernel density estimates of the distributions of (a) object-maximum UH, (b) object-maximum hourly-maximum updraft speed, (c) object-maximum hourly-maximum 10-m wind speed, (d) object area, (e) object major axis length, and (f) forecast hour, relative to 00 UTC, for all 2,627 storms labeled as supercells, QLCSs, or disorganized by the human classifiers. The median values of each distribution are provided within each violin.

Together, these statistics provide evidence that the labeling process was able to effectively distinguish between supercells, QLCSs, and disorganized storms. While differences in these distributions among the three modes exist, they often exhibit substantial overlap. For example, based on the present set of human mode classifications, solely using a UH threshold of 75 m^2/s^2 to identify supercells (e.g., Sobash et al. 2016, 2019) would inevitably lead to a large number of QLCSs being incorrectly classified as supercells (Fig. 2a). Further, producing a binary classification does not provide information related to the certainty of the classifications. Thus, we explore the use of four different statistical and ML classifiers, ranging in complexity, to provide probabilistic classifications of CAM storm mode using the hand labeled storm dataset as truth.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

c. ML Models

1) Fully Supervised Convolutional Neural Network

The hand labeled storm dataset was used to train a CNN to predict a simulated storm's convective mode. A CNN is able to identify spatial features, such as gradients in reflectivity, within the storm patches and relate those to convective mode, guided by a hand-labeled training dataset. The design of the CNN was similar to that used in Gagne et al. (2019), with slightly different hyperparameters due to the larger size of the input fields; the high-level architecture is provided in Fig. 3. Specifically, the patches had a radius of 32 grid points, which is twice the size of the storm patches in Gagne et al. (2019), and an initial set of 16 filters. The larger patch size necessitated having twice as many convolutional and pooling layers (conv+pool), since each conv+pool layer reduces the radius of the patch by one half. Each storm patch was centered on the storm centroid and only two fields were input to the CNNs: two-dimensional CREF and UH. Thus, the input field was 64x64x2 in size. Then, 4 conv+pool layers reduced the output to 4x4x128, with the final dense and output layers consisting of 9 and 3 neurons respectively. The output consisted of the probability that a storm was a supercell, QLCS, or disorganized. As in Gagne et al. (2019), dropout was used throughout the CNN during training, with a 20% dropout rate. Batch normalization was not used. Further details of the model configuration can be found in Gagne et al. (2019).

Given the relatively small size of the training dataset (i.e., 1,871 storms), image augmentation was used to increase the training dataset size. Using augmentation was possible since the labeled convective mode does not change with image rotation (i.e., a rotated storm should be classified identically to a non-rotated storm). Each storm was rotated by 2.5 degrees, between -90 and 90 degrees, producing 72 augmented storms for each labeled storm. Thus, the final training dataset for the CNN consisted of 136,583 images. The image augmentation approach substantially improved model performance, especially since the labels were unevenly divided between the three storm modes (e.g., only 16% of the 1,871 labeled storms were classified as QLCSs).

2) Fully Supervised Dense Neural Network

A dense neural network (DNN) was trained to compare to the CNN. The DNN uses scalar storm object properties as input predictors, rather than the 2D storm patches in the CNN (Fig. 3). The same training dataset was used to train the DNN, consisting of 1,871 storms, and did not use

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

1

1. Centroid Latitude
2. Centroid Longitude
3. Forecast Hour
4. Major Axis Length
5. Object eccentricity
6. Hourly-maximum object-maximum 2km - 5km AGL Updraft Helicity
7. Hourly-maximum object-mean column-max updraft speed
8. Hourly-maximum object-mean column integrated graupel
9. Environmental maximum surface-based lifted condensation level
10. Environmental minimum zonal component of Bunkers motion

TABLE 1. Input fields used in DNN classification system.

the augmented set of images that was used to train the CNN. While a large set of storm object properties was considered for training, a small subset of 10 predictors were used, guided by a correlation clustering analysis. These 10 predictors are provided in Table 1 and were chosen to include predictors that were not strongly correlated, providing independent information. The DNN architecture was designed similarly to Sobash et al. (2020), although here we used two hidden layers with 16 neurons, rather than one hidden layer and 1024 neurons in Sobash et al. (2020), given the smaller number of input fields. The output was identical to the CNN, including three probabilities for the supercell, QLCS, and disorganized storm modes.

3) Semi-supervised CNN-Gaussian Mixture Model

Creating a hand-labeled training dataset is often labor-intensive and imposes recurring costs on fully-supervised ML models. In addition to issues with inconsistent labels, if the initial segmentation approach is altered or the underlying forecast system is upgraded, then a new round of hand-labeling may be necessary and old labels may need to be discarded. Creating new ways to reduce the amount of hand-labeling would make ML emulation of expert analysis more sustainable.

Here, we tested a semi-supervised learning approach to determine if a hand-labeled training dataset was necessary to produce useful mode classifications. A CNN was trained to predict the probability of occurrence of three storm features related to mode (i.e., a proxy task): UH in the hour following the valid time of the storm exceeding 75 m^2s^{-2}, storm object eccentricity exceeding 0.95, and the storm major axis length exceeding 75 km. The CNN architecture in the semi-supervised approach was similar to the architecture of the fully-supervised CNN, consisting of four conv+pool

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

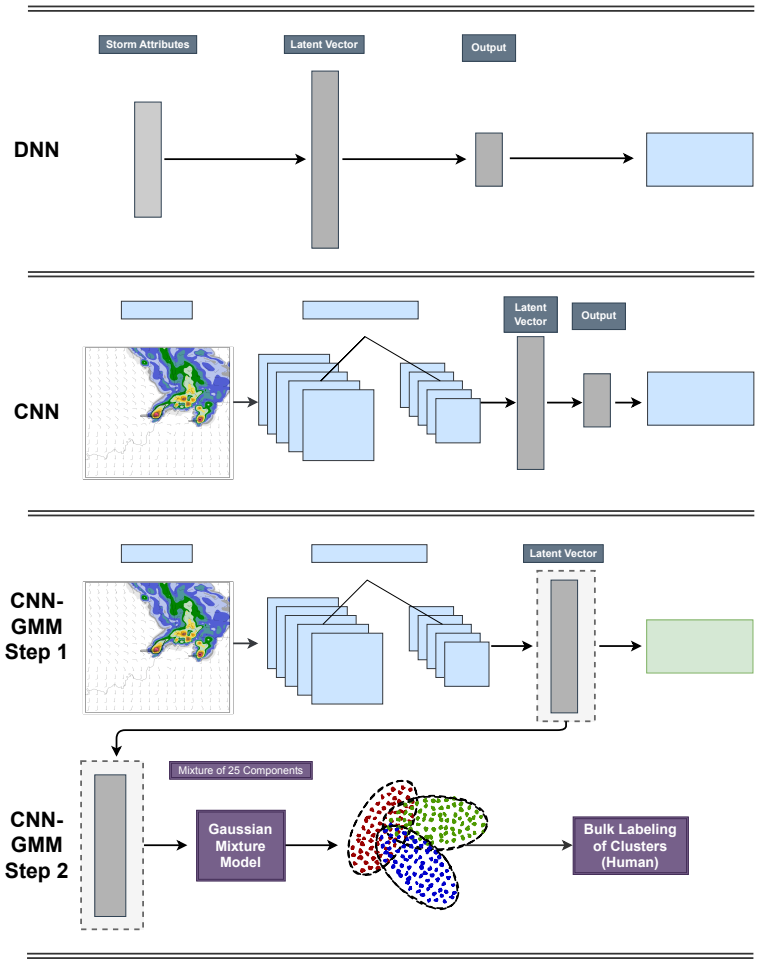

FIG. 3. Basic architecture for the DNN, CNN, and GMM machine learning models. The LR is not shown. Mode label and proxy label boxes indicate the dataset used for training each model. All classifiers output three probabilities, one for each convective mode.

layers with increasing filter widths, followed by a bottleneck dense latent vector of 128 neurons before the classification output (Fig. 3). The semi-supervised CNN was trained with a significantly larger set of storms than the supervised methods, including all storm objects in the 2010 – 2015 NCAR-WRF forecasts, but excluding the set of 2013 NCAR-WRF forecasts used as a withheld testing dataset for the supervised ML models.

After training, inference was run on the training data and the activation values from the dense latent layer were separated into 25 clusters using a Gaussian mixture model (GMM; Fig. 3). A number of clustering techniques could be used; however, we found a GMM a good choice for a variety of reasons: (a) its soft clustering attribute which provides probabilistic output, (b) its ability

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

to account for sample covariance amongst a potential cluster which can identify hyper-ellipsoids, (c) its ability to provide out-of-sample predictions, and (d) its relatively low computational cost. In a perfect scenario, only three clusters would be needed, one for each representative mode. However, as the model is likely to pick up on many other features not perfectly correlated with convective mode, we found that by increasing the number of clusters we could more effectively tease apart mode by individually assigning each of the 25 clusters to one of the three convective modes.

After clustering, each cluster was assigned to one of the three primary modes by a human by examining a set of representative storms. The representative storms were chosen by selecting examples from among the storms that had the highest probability of being in the cluster as well as those that achieved a plurality of probability for that cluster. An interactive python interface was developed to quickly perform the bulk labeling and shown in Figure 4. These new labels and probabilities were then assigned to unseen storm patches according to their cluster assignment. With this method, bulk labeling of an arbitrary sized dataset was performed in a small amount of time (i.e., less than an hour).

4) Logistic Regression

Multinomial logistic regression (LR), i.e., LR that permits multi-class output, was used as a simple supervised ML model for comparison to the three more complex ML methods. LR is desirable since it can output probabilities and has a limited set of parameters to tune. To keep the LR model as simple as possible, only three predictors were used: object maximum UH, object area, and object major axis length. These three predictors were selected due to their ability to discriminate between the distributions in Fig. 2 (e.g., there is limited overlap between the supercell and disorganized modes for object maximum UH, while area can discriminate between the cells and linear modes). L2 regularization was used with the regularization strength set to 1.0 (the python scikit-learn defaults). The LR model was fit using data from the training dataset only, as in the CNN and DNN systems, and produced three probabilities, one for each of the three primary modes.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

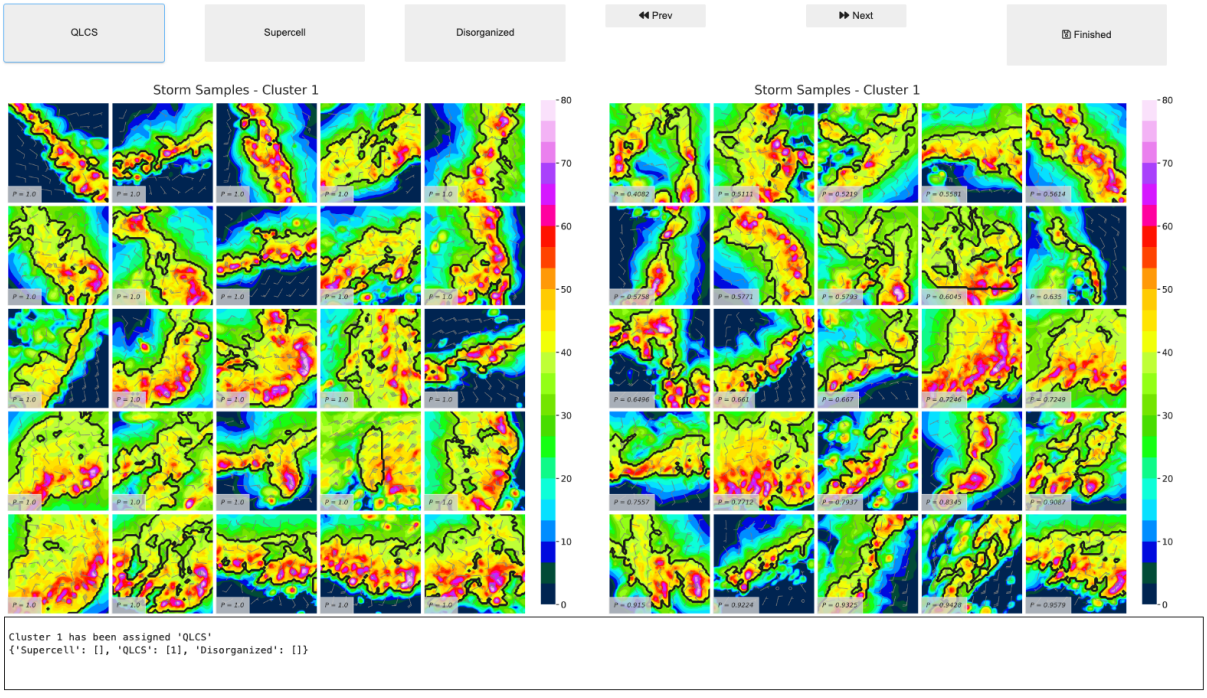

FIG. 4. Python Jupyter widget interface for bulk labeling of storms based on GMM cluster assignment. The left panel shows storms with the highest probability of belonging to that cluster and the right panel shows the lowest probability storms that were still identified as that cluster.

3. Results

Given that a ground-truth dataset for convective mode is not easily obtained, we rely on a suite of analysis methods to validate the automated storm mode classifications. First, we examine the output of the classifiers for several cases in the 2016 NCAR-WRF set of forecasts, an independent dataset not used in the training of any of the ML models. Next, we utilize the withheld hand labeled data from the storms within the 2013 NCAR-WRF forecasts and compute objective verification statistics. Finally, the set of storm objects extracted from the 2016 NCAR-WRF forecasts was used to infer indirectly the robustness of the mode classifications, including using storm reports as a proxy for mode, as well as examining the spatial and temporal climatologies of mode occurrence.

a. Example classifications within 2016 NCAR-WRF forecasts

To summarize the characteristics of storms that generate similar predictions among the four different classification algorithms, the CREF presentations of the top five storms with the highest

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

average probability across the 4 classifiers for each of the three modes are shown in Fig. 5. Each panel is centered on the storm, with the thick black line indicating the boundary of the storm object identified by the object-finding algorithm described in section 2b, and used in the DNN and LR classifiers. All 15 storms were chosen from different initialization times to document classifications from different events.

For all 15 storms, the average probability was ≥ 97%, indicating agreement among the four classification algorithms. The five supercells possessed a CREF structure characteristic of supercells, including a forward flank precipitation region and strong updraft rotation as inferred by the swaths of hourly-maximum UH (Fig. 5a–e). In fact, all five of these storms possessed hourly-maximum UH > 300 m^2/s^2, well above the typical threshold (75 m^2/s^2) used to identify severe convection within CAMs (Sobash et al. 2011, 2016). These five storms were also discrete, with limited interactions with nearby storms. In contrast to the supercells, the five storms classified as QLCSs all had objects that were larger than the supercells, composed of objects with large major axis

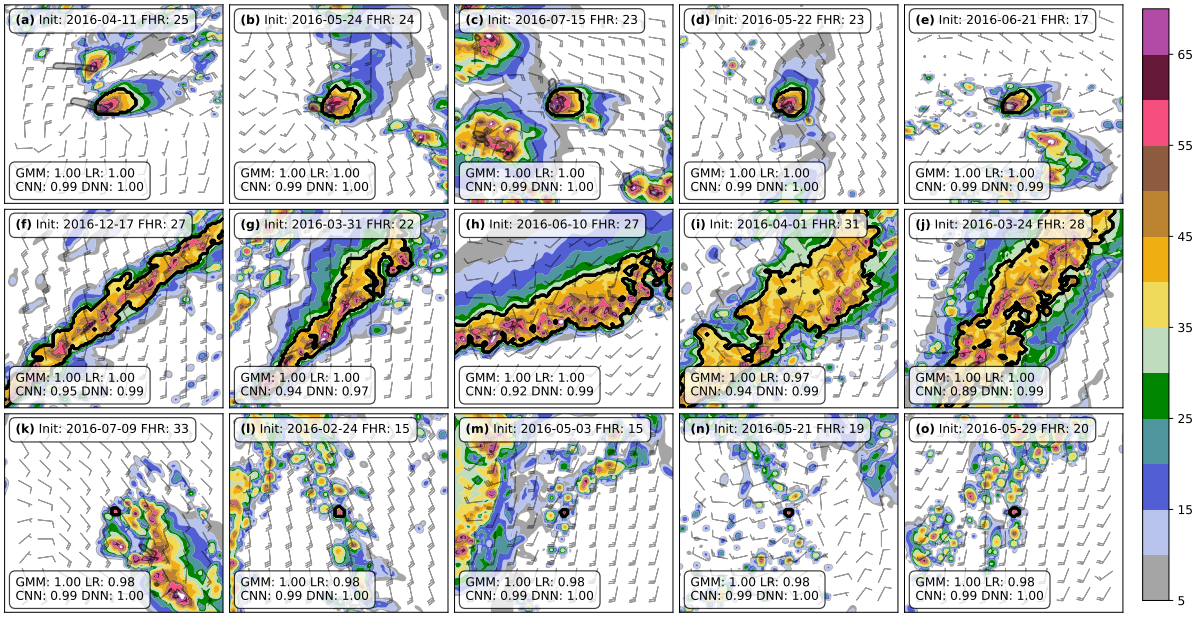

FIG. 5. Five storms with the highest average classification probabilities for the (a-e) supercell, (f-j) QLCS, and (k-o) disorganized classes. CREF (dBZ; color fill), UH > 75 m^2/s^2 (thin black contour and grey fill), 10-m wind speed and direction (kts; barbs), and storm object boundary (thick black contour) are shown in each panel, along with the probability from each classifier. Forecast initialization time and valid hour are annotated in the upper left of each panel

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

lengths relative to their minor axes (Fig. 5f–j). For the five QLCSs, the maximum within-object UH was between 76 – 137 m^2/s^2, which was lower than the supercells but still high enough to be identified as severe convection using typical UH thresholds, underscoring the difficulty of identifying only supercells with UH alone. Finally, the five disorganized storms with the highest probabilities were very small objects and often embedded within regions of other disorganized convection. The maximum within-object UH for these storms was $< 3\ m^2/s^2$. Overall, it appears that the four algorithms can provide objective classifications that match subjective impressions for a set of archetypal examples of each mode.

In contrast to the set of 15 storms where the four classifiers agreed, in many instances the classifiers disagreed on the mode, reflecting either uncertainty in the storm mode or a deficiency in a particular classification algorithm at predicting convective mode. To investigate these situations, we identified 15 random storms, from different 2016 NCAR-WRF initializations, where three different modes were predicted among the four classifiers (Fig. 6). All 15 of these storms were more challenging to subjectively classify than the 15 storms in Fig. 5. The CREF structure in many of these examples was more disorganized, although in several instances (e.g., storms in Fig. 6a, c, k, n) linear structures were apparent in the CREF field, although on smaller scales than the storms in Fig. 5f–j. The four classifiers depicted this uncertainty in two ways: 1) by producing three different estimations of the mode, as determined for each classifier by the mode with the highest probability, and 2) by producing lower probability values for the determined mode. For example, for the storm in Fig. 6g, the highest probability among the three modes for the LR classifier is only 39%, indicating that the three mode probabilities were similar. Given the likely subjective uncertainties in how a human would classify these 15 storms, it is promising that the ML algorithms can objectively reflect this uncertainty in the probability magnitudes.

Another way to assess the storm morphologies that produce spread amongst the classifications can be provided by looking at the variance among the probabilities for a particular storm across the 4 classifiers. The five storms with the highest variance among the mode probabilities were selected for further interrogation (Fig. 7). For nearly all of these storms, the high variance arises due to two classifiers producing probabilities near 1 and the other two producing probabilities near 0. In most cases, the LR probabilities were strongly correlated with the DNN probabilities, since both used storm object properties as predictors. Correlation coefficients computed with the set of ~40,000

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

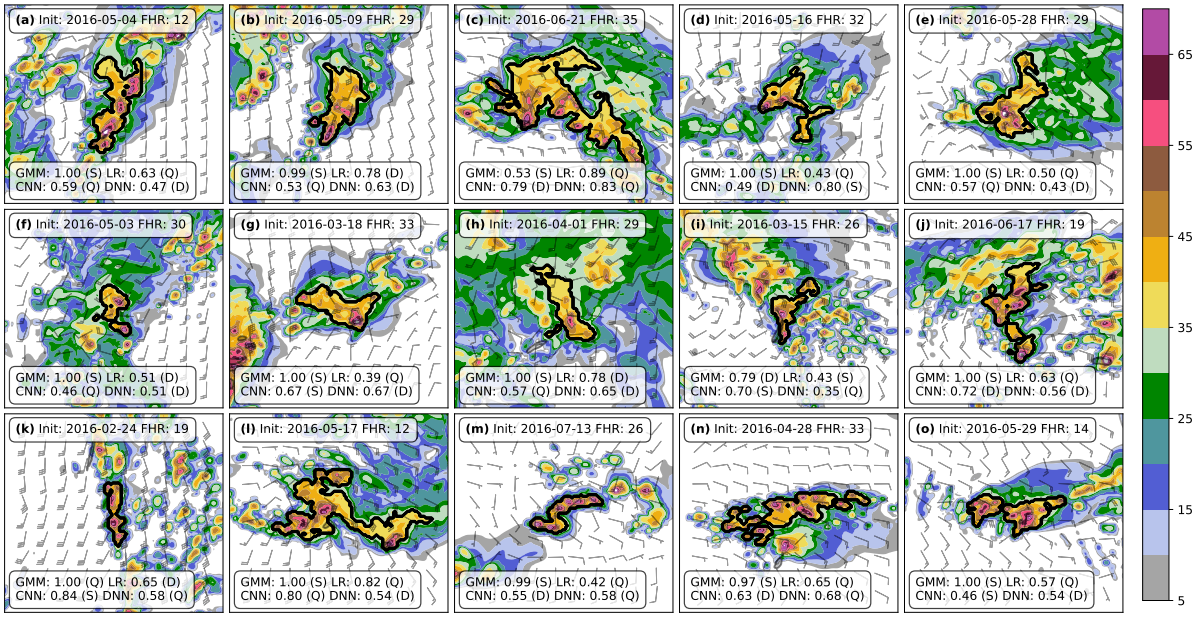

FIG. 6. As in Fig. 5, but for 15 randomly chosen storms where the four classifiers produced three different predictions of the convective mode. The highest probability and mode from each classifier is provided in each panel (S, Q, D, indicate supercell, QLCS, and disorganized, respectively).

storms in the 2016 NCAR-WRF dataset demonstrate these relationships, with the DNN and LR predictions having a correlation coefficient > 0.9 for all three modes (Fig. 8). Other correlations, e.g., between the DNN and CNN, or the DNN and GMM, were lower, ranging between 0.72 - 0.83 for the different modes. The CNN and GMM, although both using a CNN to identify features, were less correlated than the DNN and LR (Fig. 8), although in the examples in Fig. 7, the CNN and GMM generated similar probabilities.

Among the five storms with the highest variance in the supercell probabilities (Fig. 7a-e), the LR and DNN classifiers produced high probabilities for the three storms in Fig. 7a–c, while the GMM and CNN approaches produce probabilities ≤ 6%. For these three storms, the high UH magnitudes (all three storms had object-maximum UH values > 200 m^2/s^2) likely played a role in the generation of high superell probabilities for the LR and DNN classifiers, while the CNN and GMM may have relied more on the CREF structures and less on the underlying UH magnitudes, leading to lower supercell probabilities. Even though the object-maximum UH was large, the storm object in Fig. 7a was very small, and a larger intense storm was located just north of the object to be classified. While the input for the DNN and LR only has information related to the

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

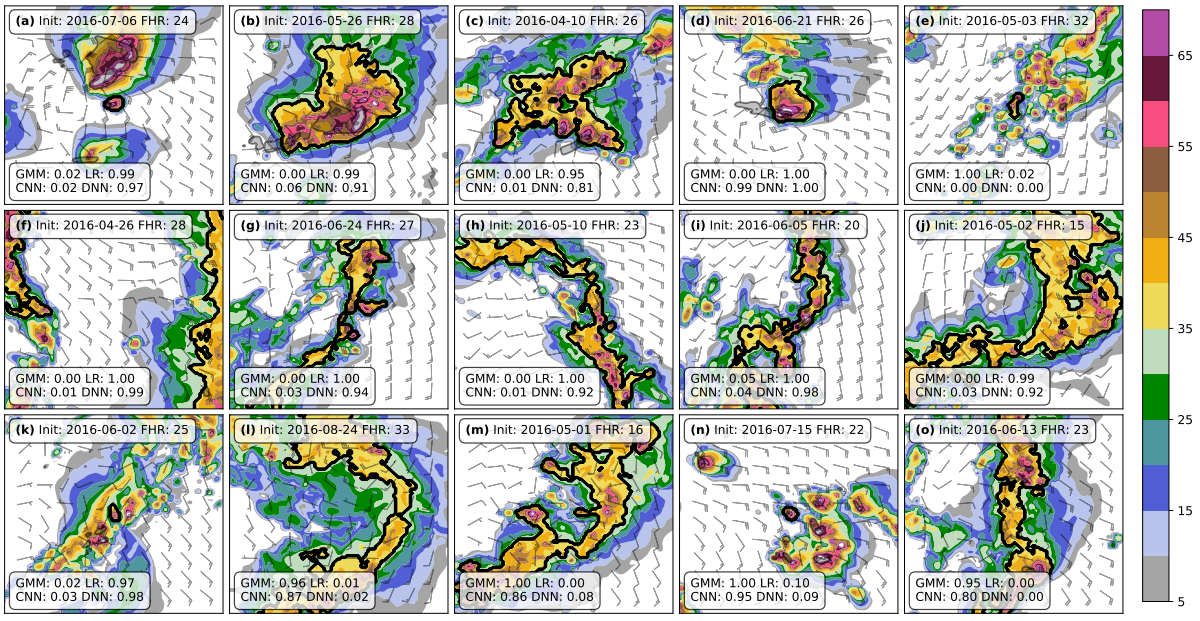

FIG. 7. As in Fig. 5, but for storms with the largest variance among the four (a-e) supercell, (f-j) QLCS, and (k-o) disorganized predictions.

storm object, the CNN and GMM are influenced by the entire scene surrounding the central storm. Interestingly, both the CNN and GMM classified this storm as disorganized, with probabilities greater than 95%. The cases in Fig. 7d,e were both situations where the GMM was an outlier. While clearly a supercell, the GMM produced a prediction of 0 for the storm in Fig. 7d, while the other three classifiers had supercell probabilities near 100%, and vice versa for the disorganized storm in Fig. 7e. It is unclear why the GMM was not able to correctly classify these storms, but the tendency for the GMM to produce predictions near 0 and 1 will be discussed in Section 3b.

For the five storms with the largest QLCS probability variance, the LR and DNN produced high probabilities of QLCSs, while the CNN and GMM probabilities were near zero (Fig. 7f-j), similar to the storms in Fig. 7a-d. All of these cases consisted of disorganized linear convective modes. The large area of the objects likely influenced the LR and DNN predictions, while the CNN and GMM weighted more heavily the disorganized CREF structure. In the case of the storm in Fig. 7f, the centroid of the storm was in an area with CREF = 0, thus the CNN and GMM were not well suited to provide accurate classifications. Finally, three of the storms with large variance in the disorganized probabilities were similar to the QLCS examples, in that the LR and DNN techniques produced low disorganized probabilities due to the larger object size (Fig. 7l, m, o). While these

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

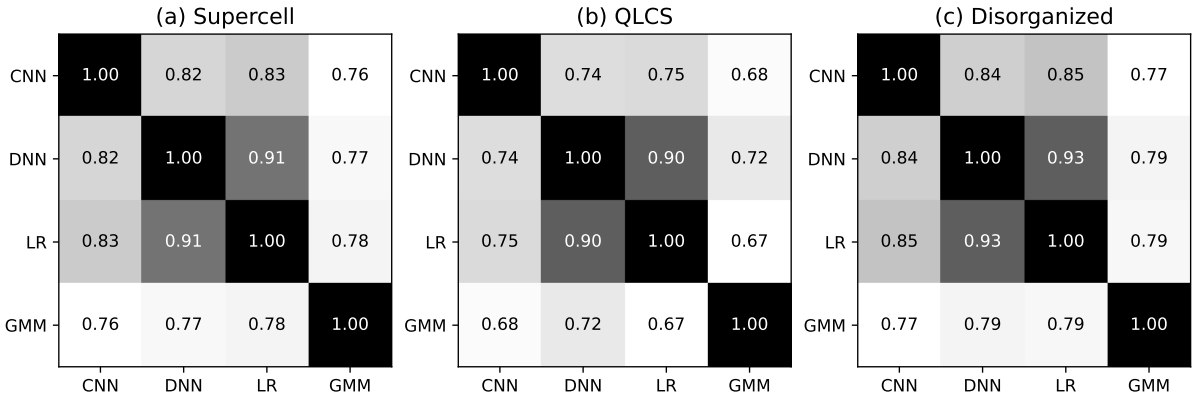

FIG. 8. Correlation matrices for the (a) supercell, (b) QLCS, and (c) disorganized predictions from the 4 classifiers for all ~40,000 storms in the 2016 NCAR-WRF dataset. Darker shading indicates higher correlation.

subjective examinations revealed important distinctions about the robustness of the classification algorithms across a wide spectrum of convective morphologies, objective validation is necessary to examine the performance over a large collection of classified storms.

b. Objective evaluation of classifiers using withheld 2013 NCAR-WRF forecasts

To objectively compare the performance of the four probabilistic classifiers, we first provide verification metrics using the withheld validation dataset of 756 hand labeled storms. Two probabilistic forecast verification metrics were used to summarize the probabilistic classification performance for each of the three modes: the area under the receiver operating characteristic curve (ROCA; Mason 1982), a measure of forecast discrimination, and the Brier Skill Score (BSS; Murphy 1973), a measure of the skill of the forecasts relative to the climatology. In this case, climatology is the fraction of storms of a particular mode within the validation dataset (i.e., the base rate). Since the 756 storm testing dataset was not evenly divided among the three modes, the supercell, QLCS, and disorganized modes had base rates of 0.15, 0.15, and 0.7, respectively. Reliability diagrams and the reliability component of the Brier Score were used to assess the calibration of the probabilistic classifications, while confusion matrices (i.e., contingency tables) were produced to determine how well the four classifiers could place each storm into the correct mode category using the maximum of the three probability values.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

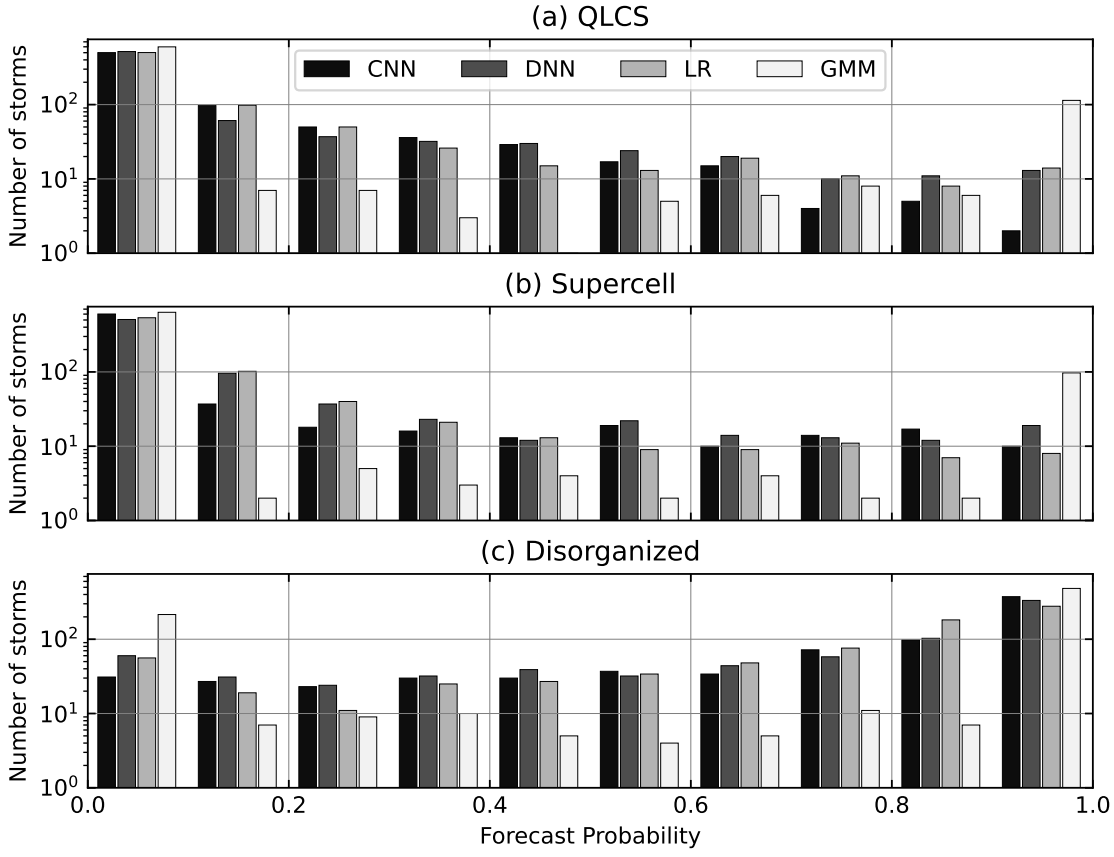

FIG. 9. Histograms of (a) QLCS, (b) supercell, and (c) disorganized probabilities from the 4 classifiers for all storms within the 2013 NCAR-WRF test dataset. Probabilities aggregated into 10% bins (0-9%, 10%-19%, etc.)

The distributions of probabilities for the three modes reveal differences in how the four classifiers generated probabilities for the testing sample storms (Fig. 9). Probabilities were assigned to storms in similar ways for both the DNN and LR across all three modes. The CNN also generated similar predictions to the DNN and LR, although the number of QLCS probabilities $\geq$ 70% is smaller than the DNN and LR (Fig. 9a). This may be due to the limited size of the patches, compared to the DNN and LR which used features directly related to storm size and shape. On the other hand, the GMM algorithm produced substantially smaller numbers of mid-range probability values for the three modes, tending to prefer to classify storms within the extreme bins (i.e., either 0–10% or 90–100%).

In terms of classification accuracy, the four classifiers performed differently for each of the three modes (Fig. 10). The LR was the most accurate at classifying the true disorganized storms

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

(i.e., storms labeled as disorganized by human classifiers) as disorganized, followed by the CNN, DNN, and the GMM. Given that 70% of the storms in the test dataset were disorganized, this should be the easiest classification task, and all four classifiers were > 85% accurate for the true disorganized storms. For the other two classes, the GMM performed the best, classifying 65% of the true supercells and QLCSs into the correct category, while the CNN performed the worst, only correctly classifying 31% and 51% of the QLCSs and supercells correctly (Fig. 10). In fact, the CNN classified most of the true QLCSs as disorganized (68%). This behavior was likely due to the fact that the GMM produced sharper probabilities (Fig. 9), while the CNN and DNN produced probabilities with more uncertainty. In many cases, two or three of the modes may have similar probabilities, yet the confusion matrices reduce this information into a single choice. Probabilistic metrics such as the ROCA and BSS provide a better view of the probabilistic distributions from each of the classifiers.

In general, the CNN and DNN both produced predictions with high values of ROCA (0.88 – 0.92) and BSSs (0.29 – 0.45; Fig. 11). The DNN had the highest BSS and ROCA for all three modes, especially for the QLCS and disorganized modes. The verification metrics among the four classifiers were most similar for the supercellular mode, potentially due to the unique properties of supercells, such as large UH and small object sizes (whereas disorganized and QLCS storm modes took on a much larger variety of shapes, sizes, and intensities). While the GMM predictions had similar ROCA values to the CNN and DNN predictions, the BSSs were much lower, indicating poor calibration. While the GMM BSSs were highest for supercells (~0.3), the GMM BSSs for the QLCSs were near zero, indicating the forecasts were no better than using a QLCS base rate forecast of 15% for every storm. Finally, the LR was competitive with the other three classifiers, producing large ROCA and high BSS, although slightly smaller than the DNN metrics.

Of the four classifiers, the DNN produced the best calibrated probabilities for the storms in the test dataset, with the smallest reliability component of the Brier Score for all three modes (Fig. 12). The CNN and LR both underpredicted the supercell probabilities and overpredicted the disorganized probabilities for most probability bins, while the QLCS probabilities were better calibrated (Fig. 12a,d). The GMM probabilities were poorly calibrated, with much larger Brier score reliability component values than the other three methods (Fig. 12c). Given the larger number of storms with probabilities in the smallest (0–10%) and largest (90–100%) bins compared to the other methods,

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

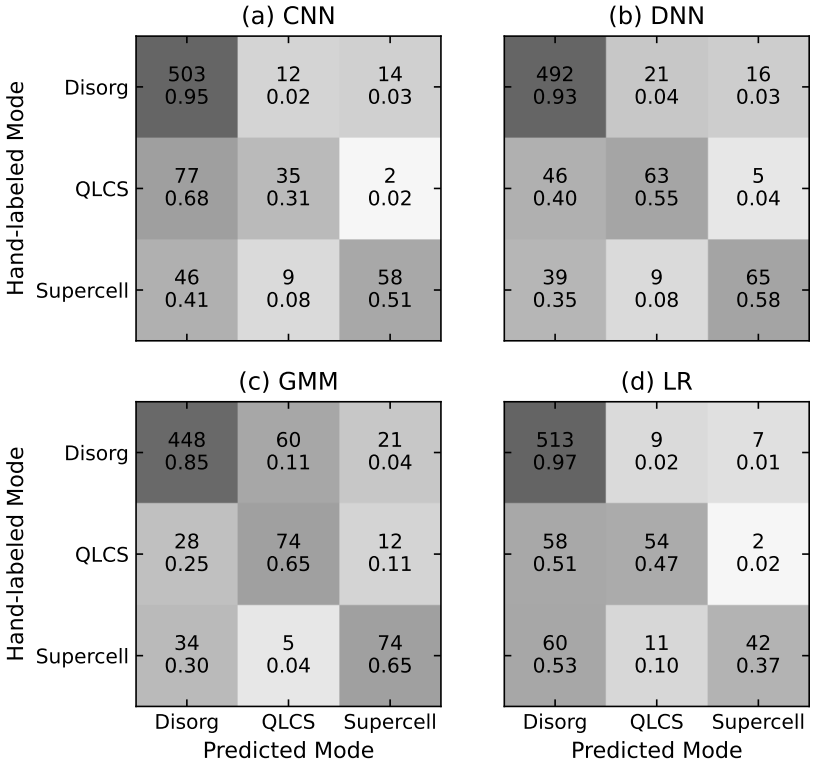

FIG. 10. Confusion matrices for the (a) CNN, (b) DNN, (c) GMM, and (d) LR predictions using the set of 756 hand-labeled storms in the test dataset. Predicted mode determined by the mode possessing the largest probability value for each storm. Number of storms and the fraction of storms relative to the total number of hand-labeled storms for each mode are provided in each quadrant. Darker shading indicates higher fraction of storms.

the sample size for the intermediate bins is less (Fig. 9), hence the large variations in the observed relative frequency. Even though the GMM probabilities were miscalibrated, the GMM was still useful as a tool to classify the mode of each storm (e.g., the GMM classifications in Fig. 10c for the QLCS and Supercell storms were slightly better than the other three techniques). Given the small size of the test dataset, further evaluations are needed. Thus, we extend the evaluation to examine the climatologies of the different mode classifications in the following sub-section.

c. Climatology of storm modes within 2016 NCAR-WRF forecasts

To examine differences in the spatial climatology of storm modes among the four classification methods, the density of storms for the three modes was obtained using the centroid locations for

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

QLCS 114 storms	ROCA	BSS
CNN	0.879	0.286
GMM	0.895	0.005
DNN	0.918	0.412
LR	0.895	0.359

Supercell 113 storms	ROCA	BSS
CNN	0.891	0.427
GMM	0.900	0.302
DNN	0.901	0.448
LR	0.866	0.373

Disorganized 529 storms	ROCA	BSS
CNN	0.855	0.335
GMM	0.869	0.186
DNN	0.891	0.452
LR	0.869	0.409

FIG. 11. Area under the receiver operating characteristic curve (ROCA) and the Brier skill score (BSS) for each mode, using the probabilities from the four classifiers for the 756 hand-labeled storms in the testing dataset. Also shown is the ROCA and BSS for the average CNN and DNN predictions.

the storm objects in the 2016 NCAR-WRF forecast dataset. The centroid locations for all storms assigned a particular mode across all forecasts were aggregated to construct a density estimate, using kernel density estimation (Fig. 13). The differences in the mode climatologies for the four different classification methods were relatively small, with all depicting distributions transitioning from Supercells, to QLCS, to Disorganized with west-to-east extent across the CONUS. All four methods produced the highest density of supercells within the central Plains, with the density decreasing to the east. The GMM method classified more storms as supercells over the eastern CONUS, as indicated by closed density contours in the mid-Atlantic and Florida (Fig. 13b). The CNN also generated relative maxima in these areas (Fig. 13a), while the DNN and LR methods did not (Fig. 13c, d). The disorganized storm mode was most common across the eastern CONUS, with the QLCS mode in between the supercell and disorganized mode maxima, with very similar depictions of storm density for all four methods for these two modes. In general, the climatologies matched subjective intuition of where and when these three modes most often occur within the CONUS (Smith et al. 2012; Ashley et al. 2019).

While the spatial climatologies are similar, the diurnal distributions for each mode vary among the four classifiers (Fig. 14). The peaks among the four classifiers for each mode are similar (early afternoon for disorganized, late afternoon and early evening for supercells, and early evening and overnight for QLCSs), but the number of storms per model run varies such that the GMM produces more QLCS and supercell classifications than the other three methods. The CNN was the least

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

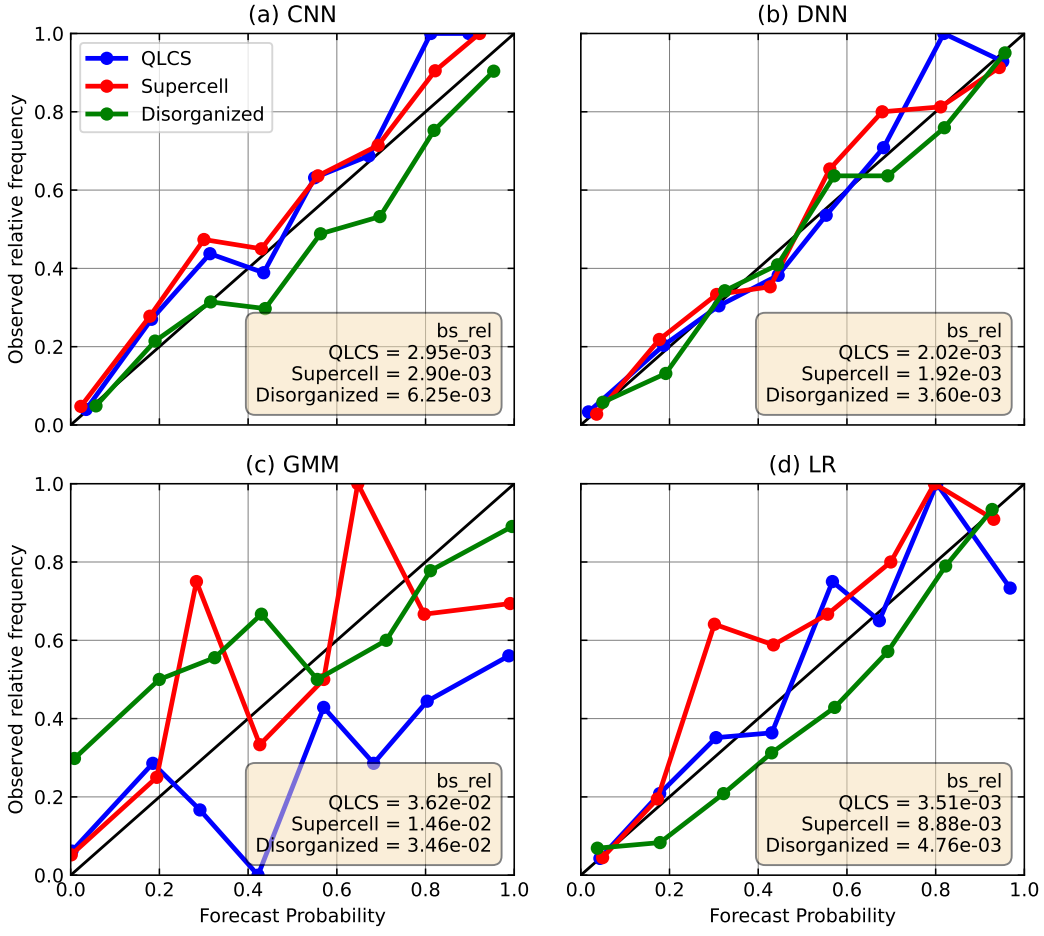

FIG. 12. Reliability diagrams for the (a) CNN, (b) DNN, (c) GMM, and (d) LR predictions for the 756 hand-labeled storms in the testing dataset and for each of the three modes. The reliability component of the Brier score is provided in each panel. Probabilities aggregated into 10% bins (0-9%, 10%-19%, etc.)

likely model to produce QLCSs. These relative differences were fairly consistent throughout the forecast period.

Given that CAPE and 0 – 6 km AGL bulk shear (Shear) are the primary environmental parameters used to distinguish between convective modes, these two parameters were examined for each storm for the three modes, with two-dimensional distributions computed for these parameters in CAPE-Shear space (Fig. 15). A clear distinction in the environmental parameters exists among the three modes, with storms classified as supercells containing larger CAPE and Shear, while storms classified as disorganized occurred in environments containing less CAPE and Shear. Storms classified as QLCS occurred in smaller CAPE environments compared to much of the distribution

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

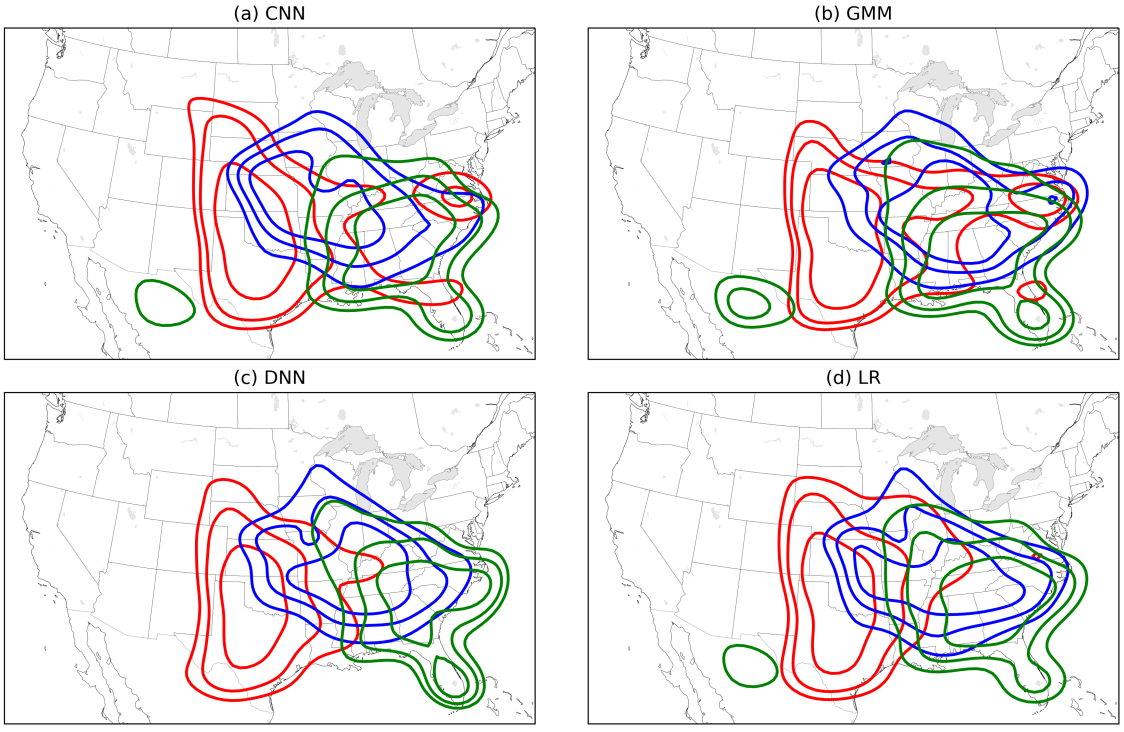

FIG. 13. Kernel density estimate of the centroids of all 2016 NCAR-WRF storms classified as (red) supercells, (blue) QLCSs, and (green) disorganized for the (a) CNN, (b) GMM, (c) DNN, and (d) LR classifiers. Density contours of 0.5, 0.66, and 0.83 are shown for each mode.

of supercells and disorganized storms, while the distribution of Shear magnitudes straddled these two other distributions. One possible explanation for the lower CAPE magnitudes is that the environmental characteristics are computed using the mean properties within the storm boundaries at time t using fields valid at forecast hour $t-1$. Since QLCS objects are larger, sometimes substantially so, compared to supercell and disorganized modes, the object mean properties are likely reduced.

d. Evaluation of 2016 NCAR-WRF mode probabilities with storm reports

Another method to evaluate the convective mode classification output is to compare the mode probabilities to the occurrence of the three storm report types (i.e., hail, wind, or tornado). A relationship should exist between mode and storm report frequency, assuming that the ML techniques are successful at classifying modes in CAM output and that the CAM mode forecasts are accurate. For example, tornadoes should be most common with supercells and least common

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

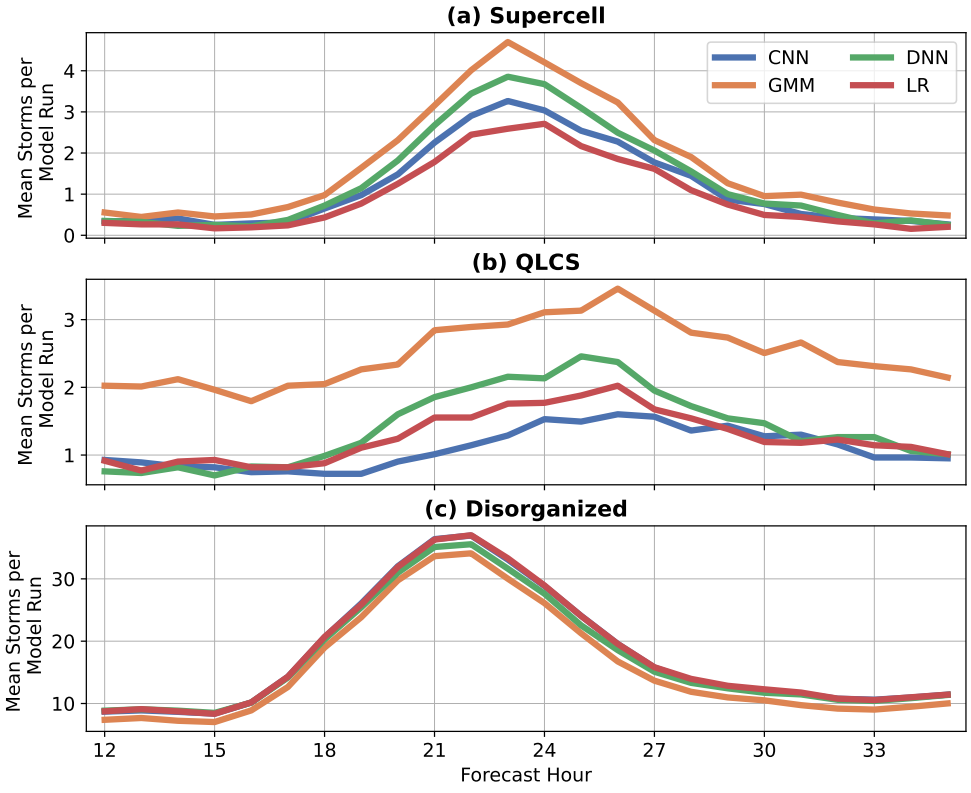

FIG. 14. Average number of 2016 NCAR-WRF storms per model run classified as (a) supercells, (b) QLCSs, and (c) disorganized for each forecast hour for the (blue) CNN, (orange) GMM, (green) DNN, and (red) LR classifiers.

within disorganized convection. To do so, the object-based mode classifications were placed on a binary 80-km grid. Each storm's mode was determined by the highest of the three mode probabilities, then, each 3-km NCAR-WRF grid point within the storm object was mapped to the closest 80-km grid box. This procedure produced a binary field for each mode indicating the 80-km grid boxes where a particular mode occurred at each forecast hour. Conditional probabilities of a storm report occurring in association with each mode were then computed by evaluating the fraction of 80-km grid boxes where at least one storm report occurred within 2-hr and 1 grid box (i.e., approximately 80-km) of the central grid box. These probabilities were computed for each mode and storm report type, as well as for the occurrence of any storm report (i.e., a total of 12 conditional probabilities for each classifier).

Similar to the results in the previous section, simply using the binary output from the classifiers produced similar results (Fig. 16). For example, at least one storm report of any type occurred

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

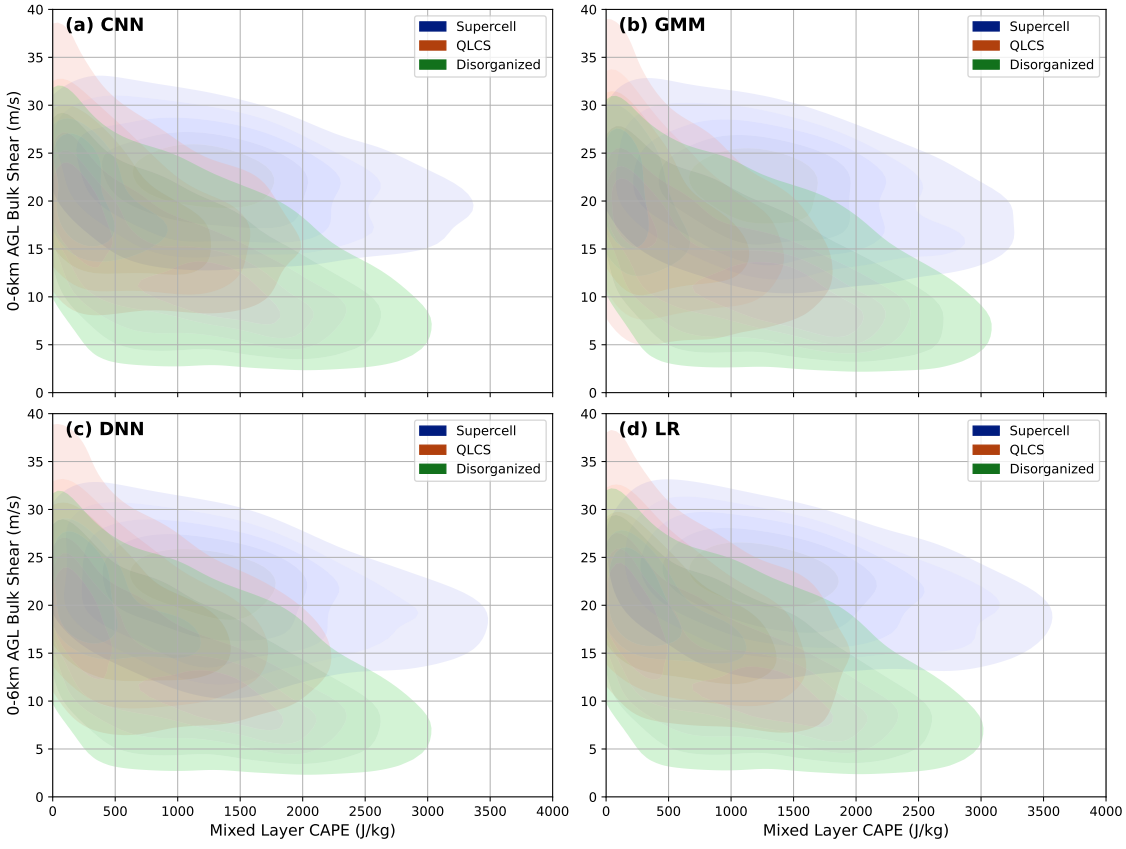

FIG. 15. Kernel density estimate of the CAPE-Shear parameters for all 2016 NCAR-WRF storms classified as (red) supercells, (blue) QLCSs, and (green) disorganized for the (a) CNN, (b) GMM, (c) DNN, and (d) LR classifiers. Density contours of 0.25, 0.375, 0.5, 0.625, 0.85 are shown for each mode.

near 55% of the grid boxes where a supercell was present in the model output, the highest of the three modes, while the probability of any report near QLCSs and disorganized storms was approximately 30-35%. This was true for all four classifiers. Among the three report types, the conditional probabilities were highest for supercells, with hail and wind reports occurring across 40-45% of grid boxes and tornadoes across 10-15% of grid boxes (Fig. 16). Again, this was similar for the four classifiers. In contrast to the supercells, wind reports (30-35%) were more likely than hail reports (10–15%) for both the QLCSs and disorganized modes, while wind reports were slightly more common in QLCSs (30-35%) than in disorganized modes (25-30%).

In addition to using the gridded binary mode fields, we also computed conditional probabilities using subsets of storms within different mode probability bins. Specifically, we computed conditional probabilities for 10 different mode probability bins, in 10% increments using the storm

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

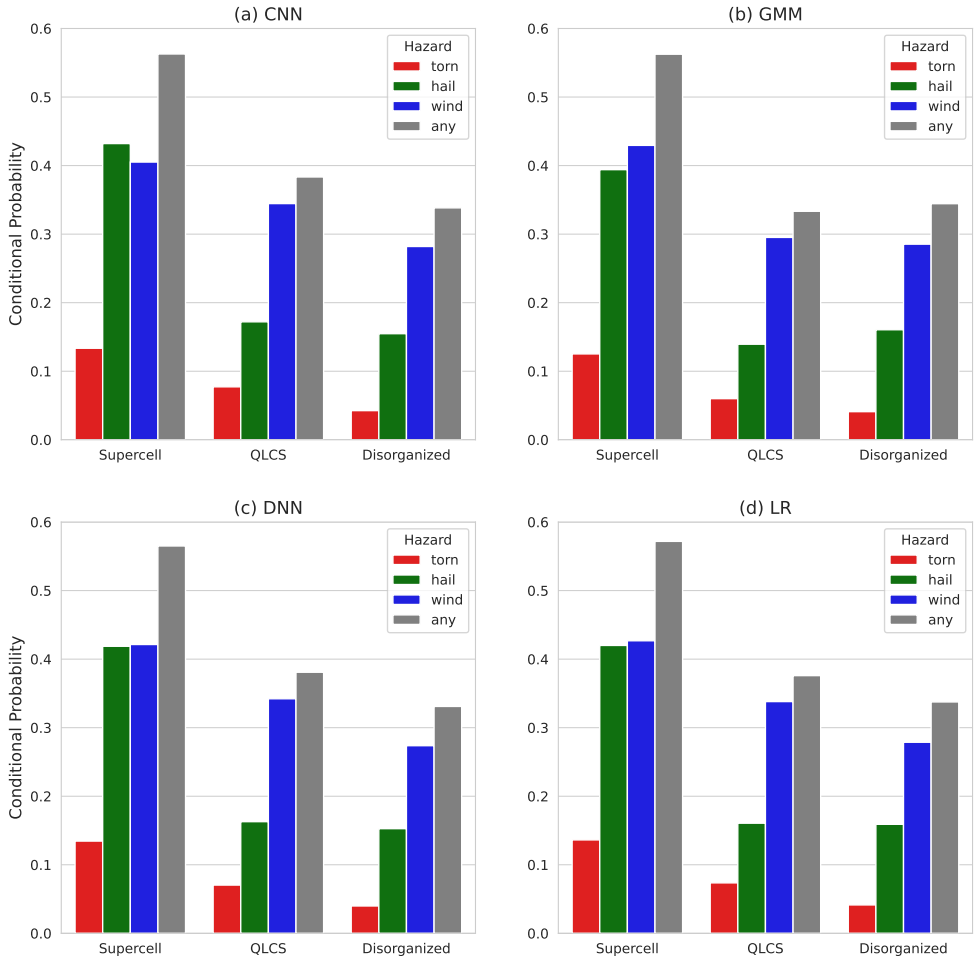

FIG. 16. Conditional probabilities of a (red) tornado, (blue) wind, (green) hail, or any (grey) report occurring within one 80-km grid box and 2-hr of three modes classified using the (a) CNN, (b) GMM, (c) DNN, and (d) LR classifiers. Results aggregated across forecast hours 12–35 for all 2016 NCAR-WRF forecasts.

centroid locations and the occurrence of a storm report within 80-km and 2-hr of the centroid location and time (Fig. 17). Using the LR classifications, the probability of occurrence of all severe report types increases sharply as the probability of supercell increases (Fig. 17a). The storms having supercell probabilities > 80% have the highest conditional probabilities of a storm report, indicating that the presence of supercells in CAM forecasts is the strongest mode-related predictor of the occurrence of severe weather reports. Specifically, for storms with a supercell probability > 90%, the probability of any report occurring is nearly 50%, while the hail and wind probabilities are both near 35%. The tornado probabilities for these storms is the highest of all three modes, at

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

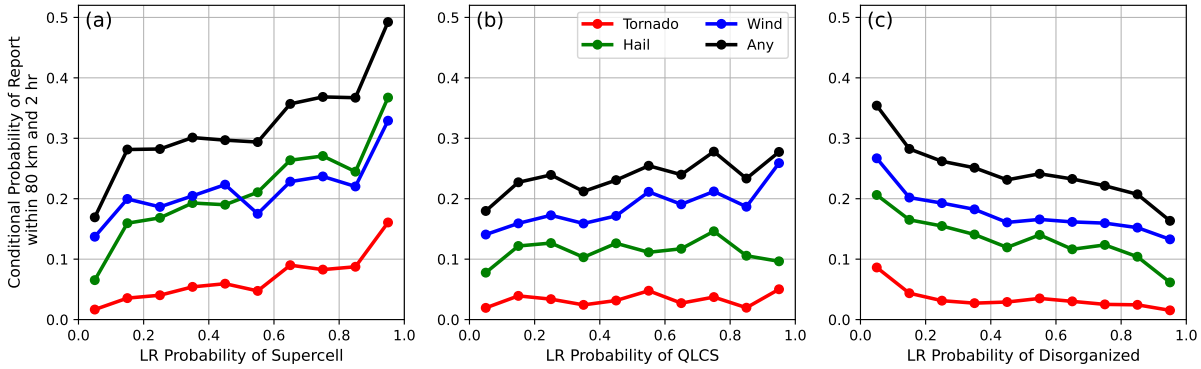

FIG. 17. Conditional probability of a (green) hail, (blue) wind, (red) tornado, or (black) any storm report occurring within 80 km and 2hr of each object centroid in the 2016 NCAR-WRF dataset, given a particular (a) supercell, (b) QLCS, and (c) disorganized probability value. Probabilities are divided into 10% bins and use only the output from the LR classifier.

15%, which is slightly higher than the results in Fig. 16. The conditional probability of a report is less dependent on the QLCS probability, although the wind probabilities do increase from 15% to 25% as the probability of a QLCS increases (Fig. 17b). In part, this may be due to the fact that the centroid of a QLCS may only encompass a small part of the overall convective system, thus being less representative of whether the storm will produce severe reports. Finally, the probability of a storm report occurring within 80 km and 2 hr decreases as the probability that the storm is disorganized increases (Fig. 17c), with wind reports being slightly more likely than hail reports for these storms.

The other three classifiers (CNN, DNN, and GMM) had slightly different depictions for these conditional report probabilities (not shown). In general, the relationship between likely mode and reports was much weaker for the GMM classifications, suggesting that the GMM had less skill at distinguishing between modes and their likely hazards due to poor calibration of the probabilities.

Overall, these results mimic prior work that has studied the likelihood of severe report types from different observed modes, but here we have applied such a method to CAM forecasts. For instance, Gallus et al. (2008) found that supercells produce a disproportionate share of severe weather reports amongst different convective modes and that disorganized modes often do not produce severe weather. The 10%-15% conditional probability of a tornado given a supercell (Fig. 16) is nearly identical to the 15% of observed mid-level mesocyclones that were associated with

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

tornadoes in Trapp et al. (2005a). Together, these results provide evidence that the ML-based mode classifications are providing useful information regarding the actual modes present within the CAM forecasts. Further, the conditional report probabilities also provide evidence that a relationship exists between the occurrence of different convective modes in CAM output and different types of severe weather hazards, which has not been documented in prior work.

4. Summary and discussion

Four different ML algorithms were used to objectively classify the modes present in a large set of 3-km horizontal grid spacing CAM forecasts, i.e., the NCAR-WRF. Each algorithm assigned three probabilities to each CAM storm object, one for each of three modes: supercell, QLCS, and disorganized. The four ML models ranged in their complexity, and included simpler techniques such as LR and a DNN, and more complex algorithms such as the CNN and GMM. Three were trained with a custom-generated hand-labeled CAM mode dataset (i.e., supervised ML), while one used CNNs with proxy targets, e.g., UH and object size, to infer mode, followed by clustering using a GMM (i.e., semi-supervised ML). Two of the ML approaches used scalar storm object attributes and environmental properties, while the other two used two-dimensional patches of CAM output fields such as CREF. The various flavors of ML techniques allowed for a thorough examination of the strengths and weaknesses of each approach.

Validation of the mode classifications included a mix of objective verification metrics such as the ROCA, BSS, and reliability using a withheld testing dataset, examinations of mode climatologies, and a comparison between ML determined modes and storm reports. All four techniques produced predictions with large ROCA (> 0.87) when evaluating with the withheld hand-labeled testing data, indicating the ability to discriminate between modes. The DNN, CNN, and LR produced large BSSs (> 0.3) for all three modes, indicating good discrimination and calibration. The GMM calibration was poor compared to the other 3 methods, especially for QLCSs, and tended to produce more high ($> 90\%$) and low ($< 10\%$) probabilities than the other methods. In spite of the calibration issues, the GMM produced similar storm mode climatologies compared to the CNN, LR, and DNN; the mis-calibration did not impact the assignment of a particular mode to a storm. Spatial and temporal climatologies of mode classifications of storms within 2016 NCAR-WRF forecasts revealed that all methods captured roughly the same spatial extent and diurnal cycle for each model,

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

but with differences in absolute frequency depending on the model used. The GMM labeled the fewest storms as disorganized and captured more QLCS events compared with other methods. Finally, a strong relationship existed between the three modes and the likelihood of different storm report types. This relationship not only bolsters the credibility of the mode classifications, but also provides an indirect assessment of the CAM mode predictions (if the CAM forecasts of mode were erroneous, then such a relationship would likely not be as robust).

The similarities among the evaluations for the DNN, LR, and CNN suggest that simple ML approaches can provide reliable depictions of the convective mode within CAM forecasts, and that the added complexity of using more complex algorithms, while providing some small benefit in the objective metrics for our small test dataset, may not be warranted. It is intriguing that the LR and DNN techniques were able to be successfully trained with ~1,500 storms and did not require the usage of data augmentation, as was used to train the CNN. Thus, for this particular problem, an LR or DNN, using a small set of predictors related to object size and intensity, may provide sufficient performance. The size of the patch used as input in the CNN, as well as the presence of multiple storms in the patch, may also have led to suboptimal performance compared to the LR and DNN, especially for QLCSs and disorganized convection (Fig. 11).

A benefit of the GMM approach is that a fully hand labeled training dataset is not needed, and the only human input is to assign the individual clusters within the GMM, which can be done rather simply by an individual looking through a small subset of storms in each cluster. While the classifications in the withheld testing dataset possessed large ROCA, similar to the other three classifiers, its predictions were poorly calibrated, and tended to produce probabilities near 0% or 100% and have lower reliability scores. While for some classification problems sharpness (the tendency for probabilities to cluster near the extremes) is desirable, there is often considerable uncertainty in assigning a convective mode for a particular storm. The output of the probabilistic classifiers should reflect this uncertainty as well as possible. Future work should explore how to better calibrate the GMM predictions without significantly increasing the labeling burden.

In addition to improving the calibration of the GMM, further optimization of the CNN, DNN, and GMM hyperparameters may result in better performance for the specific task of convective mode classification. Exploring the multi-dimensional space of hyperparameters, including identifying the optimal selection of input features and neural network architectures, is computationally intensive

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

and beyond the scope of this work. Nevertheless, given the similarity of the results among the multiple different ML algorithms, and the uncertainty inherent in storm mode classification, there is likely diminishing returns to further optimization. The current set of ML algorithms should be used as a baseline in future work, with hyperparameter optimization informing how the algorithms can be fine-tuned for use with a different training dataset.

Moreso than the ML architectures, an open-ended question is how the current results extend to other NWP systems, including operational CAMs such as the High Resolution Rapid Refresh (HRRR; Dowell et al. 2022). To examine this question, the authors have tested the use of the CNN, DNN, and GMM classifiers within the 2021 and 2022 NOAA Hazardous Weather Testbed Spring Experiments (Clark et al. 2022), using NCAR-WRF and HRRR datasets as input to produce mode classifications in real-time. Participants in these real-time experiments were provided with probabilistic convective mode output and asked to evaluate it compared to their subjective impressions of the modes present within a particular day's forecast. In general, the subjective feedback of the various classifiers was positive in that the cases where the mode was obvious to the participants, the automated classifiers performed well, and that in cases where the mode was ambiguous to the participants, the classifiers represented this uncertainty by producing low probabilities. The subjective feedback also indicated that the GMM technique was poorly calibrated, matching the objective verification metrics provided in this work. The classifiers were somewhat robust to changes to the input model, in that when a classifier trained with NCAR-WRF data was used to classify storms in the HRRR, the output was deemed to still be valuable, although the average subjective ratings for the classifiers using the HRRR input were slightly lower than the classifications for the NCAR-WRF output. Applying these trained models to other CAMs may not work as readily, given that the NCAR-WRF and HRRR have similar model components (i.e., WRF-based with similar physics choices). Future work should explore how these classifiers can be optimized across a diverse set of CAM systems.

Acknowledgments. We thank Steve Weiss, Morris Weisman, Rich Thompson, and Glen Romine, who all assisted with the labeling of CAM convective storms. We also thank two anonymous reviewers whose suggestions improved the manuscript. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. This research was

Accepted for publication in *Monthly Weather Review.* DOI 10.1175/MWR-D-22-0342.1.

supported by NOAA OAR grants NA17OAR4590114, NA19OAR4590128, NSF Grant No. ICER-2019758, and the NCAR Short-term Explicit Prediction Program. Supercomputing support was provided by NCAR Cheyenne and Casper (Computational and Information Systems Laboratory, CISL 2020).

Data availability statement. The training and analysis code are in the HWT Mode repository archived at `https://doi.org/10.5281/zenodo.7730773`. Storm objects were analyzed with the Hagelslag package `https://doi.org/10.5281/zenodo.6862433`. Data describing storms and predictions are available from Globus at `https://app.globus.org/file-manager?origin_id=b2a321ab-92ee-471b-9eac-64419d091661&origin_path=%2F`.

References

Ashley, W. S., A. M. Haberlie, and J. Strohm, 2019: A climatology of quasi-linear convective systems and their hazards in the united states. *Wea. Forecasting*, **34 (6)**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0014.1.

Biard, J. C., and K. E. Kunkel, 2019: Automated detection of weather fronts using a deep learning neural network. *Adv. Stat. Clim. Meteorol. Oceanogr.*, **5 (2)**, 147–160, https://doi.org/10.5194/ascmo-5-147-2019, URL https://ascmo.copernicus.org/articles/5/147/2019/.

Brotzge, J. A., S. E. Nelson, R. L. Thompson, and B. T. Smith, 2013: Tornado probability of detection and lead time as a function of convective mode and environmental parameters. *Wea. Forecasting*, **28 (5)**, 1261–1276, https://doi.org/10.1175/WAF-D-12-00119.1, URL https://journals.ametsoc.org/view/journals/wefo/28/5/waf-d-12-00119_1.xml.

Clark, A. J., I. L. Jirak, and Coauthors, 2022: The 3rd real-time, virtual spring forecasting experiment to advance severe weather prediction capabilities. *Bull. of the Amer. Meteor. Soc*, in press, https://doi.org/10.1175/BAMS-D-22-0213.1.

Computational and Information Systems Laboratory, CISL, 2020: Cheyenne: HPE/SGI ICE XA System (NCAR Community Computing). Tech. rep., National Center for Atmospheric Research. URL https://doi.org/10.5065/D6RX99HX.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

Dowell, D. C., and Coauthors, 2022: The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part i: Motivation and system description. *Wea. and Forecasting.*, **37 (8)**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Gagne, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *J. Atmos. Ocean. Technol.*, **26 (7)**, 1341–1353, https://doi.org/10.1175/2008JTECHA1205.1, URL https://journals.ametsoc.org/view/journals/atot/26/7/2008jtecha1205_1.xml.

Gagne, D. J., II, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Weather Rev.*, **147 (8)**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1, URL https://journals.ametsoc.org/view/journals/mwre/147/8/mwr-d-18-0316.1.xml.

Gagne II, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

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 (1)**, 101–113, https://doi.org/10.1175/2007WAF2006120.1, URL https://journals.ametsoc.org/view/journals/wefo/23/1/2007waf2006120_1.xml.

Guillot, E. M., T. M. Smith, V. Lakshmanan, K. L. Elmore, D. W. Burgess, and G. J. Stumpf, 2008: Tornado and severe thunderstorm warning forecast skill and its relationship to storm type. *Preprints, 24th Conf. on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, New Orleans, LA, Amer. Meteor. Soc. A*, Vol. 4, URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.484.3505&rep=rep1&type=pdf.

Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2020: Classifying convective storms using machine learning. *Wea. Forecasting*, **35 (2)**, 537–559, https://doi.org/10.1175/WAF-D-19-0170.1, URL https://journals.ametsoc.org/view/journals/wefo/35/2/waf-d-19-0170.1.xml.

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.*,

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

111, 104–113, https://doi.org/10.1016/j.atmosres.2012.03.004, URL https://www.sciencedirect.com/science/article/pii/S0169809512000725.

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, URL https://www.sciencedirect.com/science/article/pii/S016980951200049X.

Lagerquist, R., A. McGovern, and D. J. Gagne, II, 2019: Deep learning for spatially explicit prediction of Synoptic-Scale fronts. *Wea. Forecasting*, **34 (4)**, 1137–1160, https://doi.org/10.1175/WAF-D-18-0183.1, URL https://journals.ametsoc.org/view/journals/wefo/34/4/waf-d-18-0183_1.xml.

Lakshmanan, V., K. L. Elmore, and M. B. Richman, 2010: REACHING SCIENTIFIC CONSENSUS THROUGH a COMPETITION. *Bull. Am. Meteorol. Soc.*, **91 (10)**, 1423–1429, URL http://www.jstor.org/stable/26233036.

Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *Journal of Atmospheric and Oceanic Technology*, **26 (3)**, 523 – 537, https://doi.org/10.1175/2008JTECHA1153.1.

Lakshmanan, V., and T. Smith, 2009: Data mining storm attributes from spatial grids. *J. Atmos. Ocean. Technol.*, **26 (11)**, 2353–2365, https://doi.org/10.1175/2009JTECHA1257.1, URL https://journals.ametsoc.org/view/journals/atot/26/11/2009jtecha1257_1.xml?tab_body=abstract-display.

Lintott, C., and Coauthors, 2011: Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Mon. Not. R. Astron. Soc.*, **410 (1)**, 166–178, https://doi.org/10.1111/j.1365-2966.2010.17432.x, URL http://dx.doi.org/10.1111/j.1365-2966.2010.17432.x.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291–303.

Murphy, A. H., 1973: Hedging and skill scores for probability forecasts. *J. Appl. Meteorol. Climatol.*, **12 (1)**, 215–223, https://doi.org/10.1175/1520-0450(1973)012⟨0215:HASSFP⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/apme/12/1/1520-0450_1973_012_0215_hassfp_2_0_co_2.xml?tab_body=pdf.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

Potvin, C. K., and Coauthors, 2022: An iterative storm segmentation and classification algorithm for convection-allowing models and gridded radar analyses. *J. Atmos. Ocean. Technol.*, **39**, 999–1013, https://doi.org/10.1175/JTECH-D-21-0141.1.

Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of Geophysics*, **53 (2)**, 323–361, https://doi.org/10.1002/2014RG000475.

Schwartz, C. S., and R. A. Sobash, 2019: Revisiting sensitivity to horizontal grid spacing in convection-allowing models over the central and eastern united states. *Mon. Weather Rev.*, URL https://journals.ametsoc.org/view/journals/mwre/147/12/mwr-d-19-0115.1.xml.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous united states. part i: Storm classification and climatology. *Wea. Forecasting*, **27 (5)**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1, URL https://journals.ametsoc.org/view/journals/wefo/27/5/waf-d-11-00115_1.xml.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in Convection-Allowing model forecasts. *Wea. Forecasting*, **26 (5)**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1, URL https://journals.ametsoc.org/view/journals/wefo/26/5/waf-d-10-05046_1.xml.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of Neural-Network and Surrogate-Severe probabilistic convective hazard guidance derived from a Convection-Allowing model. *Wea. Forecasting*, **35 (5)**, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1, URL https://journals.ametsoc.org/view/journals/wefo/35/5/wafD200036.xml.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31 (1)**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1, URL https://journals.ametsoc.org/view/journals/wefo/31/1/waf-d-15-0138_1.xml.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, and M. L. Weisman, 2019: Next-Day prediction of tornadoes using Convection-Allowing models with 1-km horizontal grid spacing. *Wea. Forecasting*, **34 (4)**, 1117–1135, https://doi.org/10.1175/WAF-D-19-0044.1, URL https://journals.ametsoc.org/view/journals/wefo/34/4/waf-d-19-0044_1.xml.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. of the Amer. Meteor. Soc.*, **39 (10)**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the rapid update cycle. *Wea. Forecasting*, **18 (6)**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018⟨1243:CPSWSE⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/wefo/18/6/1520-0434_2003_018_1243_cpswse_2_0_co_2.xml.

Thompson, R. L., B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous united states. part II: Supercell and QLCS tornado environments. *Wea. Forecasting*, **27 (5)**, 1136–1154, https://doi.org/10.1175/WAF-D-11-00116.1, URL https://journals.ametsoc.org/view/journals/wefo/27/5/waf-d-11-00116_1.xml.

Trapp, R. J., G. J. Stumpf, and K. L. Manross, 2005a: A reassessment of the percentage of tornadic mesocyclones. *Wea. Forecasting*, **20 (4)**, 680–687, https://doi.org/10.1175/WAF864.1.

Trapp, R. J., S. A. Tessendorf, E. S. Godfrey, and H. E. Brooks, 2005b: Tornadoes from squall lines and bow echoes. part i: Climatological distribution. *Wea. Forecasting*, **20 (1)**, 23–34, https://doi.org/10.1175/WAF-835.1, URL https://journals.ametsoc.org/view/journals/wefo/20/1/waf-835_1.xml.

Accepted for publication in *Monthly Weather Review*. DOI 10.1175/MWR-D-22-0342.1.