A Deep-Learning Model for Automated Detection of Intense Midlatitude Convection Using Geostationary Satellite Images

JOHN L. CINTINEO,[a] MICHAEL J. PAVOLONIS,[b] JUSTIN M. SIEGLAFF,[a] ANTHONY WIMMERS,[a] JASON BRUNNER,[a] AND WILLARD BELLON[a]

[a] *Cooperative Institute of Meteorological Satellite Studies, University of Wisconsin—Madison, Madison, Wisconsin*
[b] *NOAA/NESDIS/Center for Satellite Applications and Research/Advanced Satellite Products Branch, Madison, Wisconsin*

(Manuscript received 19 February 2020, in final form 7 October 2020)

ABSTRACT: Intense thunderstorms threaten life and property, impact aviation, and are a challenging forecast problem, particularly without precipitation-sensing radar data. Trained forecasters often look for features in geostationary satellite images such as rapid cloud growth, strong and persistent overshooting tops, U- or V-shaped patterns in storm-top temperature (and associated above-anvil cirrus plumes), thermal couplets, intricate texturing in cloud albedo (e.g., ''bubbling'' cloud tops), cloud-top divergence, spatial and temporal trends in lightning, and other nuances to identify intense thunderstorms. In this paper, a machine-learning algorithm was employed to automatically learn and extract salient features and patterns in geostationary satellite data for the prediction of intense convection. Namely, a convolutional neural network (CNN) was trained on 0.64-μm reflectance and 10.35-μm brightness temperature from the Advanced Baseline Imager (ABI) and flash-extent density (FED) from the Geostationary Lightning Mapper (GLM) on board *GOES-16*. Using a training dataset consisting of over 220 000 human-labeled satellite images, the CNN learned pertinent features that are known to be associated with intense convection and skillfully discriminated between intense and ordinary convection. The CNN also learned a more nuanced feature associated with intense convection—strong infrared brightness temperature gradients near cloud edges in the vicinity of the main updraft. A successive-permutation test ranked the most important predictors as follows: 1) ABI 10.35-μm brightness temperature, 2) ABI GLM flash-extent density, and 3) ABI 0.64-μm reflectance. The CNN model can provide forecasters with quantitative information that often foreshadows the occurrence of severe weather, day or night, over the full range of instrument-scan modes.

SIGNIFICANCE STATEMENT: Trained human forecasters are particularly adept at picking out indicators of intense thunderstorms in weather satellite imagery. While previous algorithms have been developed to detect certain aspects of intense thunderstorms, this research is unique as it uses deep learning to incorporate the detection of all satellite-based features of intense thunderstorms, mimicking human pattern recognition. The model described in this research can provide forecasters rapid guidance on evolving severe weather threats day or night, even in the absence of precipitation-sensing weather radar.

KEYWORDS: Deep convection; Satellite observations; Neural networks; Machine learning

1. Introduction

Since the advent of weather satellites, researchers have been investigating signatures of intense convection from satellite images (e.g., Purdom 1976; Adler and Fenn 1979; Menzel and Purdom 1994; Schmit et al. 2005, 2015). Forecasters frequently scrutinize satellite imagery to help infer storm dynamics and diagnose and forecast the intensity of thunderstorms, which can generate a variety of hazards. Intense convective updrafts frequently penetrate the tropopause, resulting in overshooting cloud tops. These features may block strong upper-level wind flow, which is diverted around the overshooting tops, carrying cloud debris from the updraft summit, resulting in U- or V-shaped thermal couplets in infrared brightness temperature imagery (e.g., Setvák et al. 2013; Wang 2007; Brunner et al. 2007). Furthermore, high-refresh sequences of geostationary satellite images have been used to retrieve cloud-top divergence and cloud-top vorticity and subsequently detect supercell thunderstorms (Apke et al. 2016). Textural patterns at cloud top have also been used to infer updraft strength (Bedka and Khlopenkov 2016). In the presence of strong upper-level flow, some overshoots generate above-anvil cirrus plumes (AACP) downstream from the overshooting top as a result of internal gravity wave breaking and are apparent in visible satellite imagery (Wang 2003; Wang et al. 2016; Homeyer et al. 2017; Bedka et al. 2018). AACPs in visible imagery are responsible for cold-U features in satellite infrared imagery, together forming a robust indicator of ongoing or imminent severe weather hazards such as large hail, strong downburst wind gusts, and tornadoes.

Total lightning information is also known to be useful for diagnosing and forecasting intense convection. The electrical energy manifested in lightning flashes is related to the kinetic energy and overall vigor of thunderstorm updrafts. Updrafts provide an environment for mixed-phase precipitation processes and a mechanism for microphysical charge transfer and cloud-scale charge separation, generating large electrical potential differences. An increasing rate of total lightning flashes in a storm is often a good indicator of an intensifying convective updraft (e.g., Schultz et al. 2011).

Corresponding author: John L. Cintineo, john.cintineo@ssec.wisc.edu

DOI: 10.1175/WAF-D-20-0028.1

TABLE 1. Dates, sample size, and fraction of sample that is labeled as "intense" convection, for the training, validation, and test datasets.

	Training	Validation	Test
Dates	6–8 May 2018	1 May 2018	10 May 2018
	11 May 2018	2 May 2018	23 Jun 2018
	13–15 May 2018	3 May 2018	2 Jul 2018
	18–19 May 2018	4 May 2018	
	29 May 2018	5 May 2018	
	31 May–1 Jun 2018	14 Jun 2018	
	8 Jun 2018	15 Jun 2018	
	11 Jun 2018	20 Jul 2018	
	17 Jun 2018		
	19 Jun 2018		
	9 Jul 2018		
	29 Jul 2018		
Sample size	153 364	51 178	18 329
"Intense" class fraction	10.1%	11.6%	14.2%

In severe weather warning operations, an operational forecaster is confronted with far more data than can be manually analyzed. Automated methods can help forecasters manage data overload and can provide insights that may have otherwise gone unnoticed. While automated algorithms that identify targeted features of satellite-based observations of intense convection have been successfully developed and tested (e.g., Schultz et al. 2011; Bedka and Khlopenkov 2016; Apke et al. 2016), no algorithm or system has been able to integrate all of the severe weather-pertinent features into a single product. In an effort to simplify algorithm development and consolidate salient satellite-based features of thunderstorms into a single output, we utilize a deep-learning approach that mimics expert human pattern recognition of intense convection in satellite imagery. The goal of this approach is to quantify convective intensity automatically, saving forecasters time in identifying, diagnosing, and prioritizing threats.

Deep learning is a branch of machine-learning methods based on artificial neural networks with feature learning, or the ability to automatically find salient features in data (e.g., Schmidhuber 2015). Deep-learning models, such as convolutional neural networks (CNN), have the ability to encode spatiotemporal features at multiple scales and levels of abstraction with the ultimate goal of encoding features that maximize performance (Fukushima 1980). In a fully connected neural network, each neuron in layer k is connected to all neurons in the adjacent layers ($k - 1$ and $k + 1$), and each connection is associated with a learned weight. The learned weights are used in linear combinations (the value at neuron j in layer k is a linear combination of the values at all neurons in layer $k - 1$, and the weights in the linear combination are those associated with connections between neuron j and the neurons in layer $k - 1$). Each linear combination is followed by a nonlinear activation function, which allows the network to learn nonlinear relationships.

In a CNN, the neurons in each layer are arranged in a spatial grid, with the same number of dimensions as the input data (e.g., if the input data are 2D satellite grids, the neurons in each layer are in a 2D grid as well). Each neuron has a "receptive field," the subdomain of the previous layer to which it is connected. In general, the subdomain is smaller than the entire domain. That is, neuron j in layer k is connected only to neurons in adjacent layers ($k - 1$ and $k + 1$) that are in the same spatial neighborhood. In this work, neurons in the initial input layer are pixels of an image. Deep-learning models have yielded excellent performance on image recognition tasks for nonmeteorological phenomena (e.g., Krizhevsky et al. 2012; Litjens et al. 2017; Li et al. 2018) and we seek to apply such methods to weather satellite imagery.

There has already been success with deep-learning methods for synoptic-scale front detection (Lagerquist et al. 2019), hail size estimation in numerical weather prediction (NWP) model output (Gagne et al. 2019), and tornado prediction (Lagerquist et al. 2020). To the authors' knowledge, this is the first application of deep learning on weather satellite imagery targeting convection intensity of individual thunderstorms. In this paper, a CNN model was trained in a supervised manner to generate an "intense convection probability" (ICP). One benefit of CNNs, and deep learning in general, is the greatly reduced need for feature engineering (i.e., turning gridded data into scalar features or predictors in machine-learning models), which can be analytically challenging, difficult to optimize for predictive skill, and lacking in formalized evaluation tools. This not only saves considerable time but makes the model more objective by not superimposing scientists' preconceived notions of what features are important in an image, which also presents an opportunity to learn new insights into physical phenomena. The model learns from the training data the salient spatiotemporal features that result in the best fit, using a numerical optimization process called backpropagation (Goodfellow et al. 2016). After discussing the construction of the CNN, we characterize its performance and show that the model learned several features, including a number of features that human experts most often associate with intense convection, as well as a lesser-known feature.

2. Data and methods

a. Meteorological data

GOES-16 Advanced Baseline Imager (ABI; Schmit et al. 2005) and Geostationary Lightning Mapper (GLM; Rudlosky et al. 2019; Goodman et al. 2013) radiance and flash data were collected for 29 convectively active days in the May–July 2018

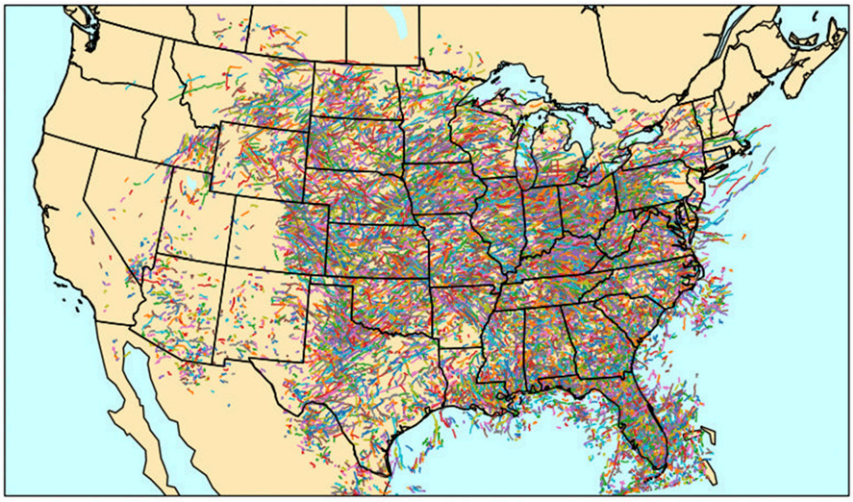

FIG. 1. All 14 746 thunderstorm tracks for the training, validation, and testing datasets. The different colors are used to help distinguish individual storm tracks.

timeframe (Table 1). Dates were selected to include a variety of satellite viewing angles and geography (see Fig. 1), each representing a "convective day" from 1200 UTC of the listed date to 1200 UTC of the next date. The channel 2 (CH02) 0.64-μm reflectance[1] and channel 13 (CH13) 10.35-μm brightness temperature were calculated using attributes from ABI Level-1b files for all ABI contiguous U.S. (CONUS) sector files for each convective date (288 files per day, or 5-min temporal resolution). Please see Table 2 for a summary of the raw input datasets. This paper focuses on these two ABI channels since operational forecasters most often use them to analyze developing and mature convection in satellite data. Future work may assess the impact of other ABI channels. GLM Level-2 files, which contain lightning flash, group, and event point data, were processed with an open-source software package called glmtools (Bruning 2019). This software package was used to create gridded fields for several GLM attributes: flash-extent density (FED), flash-centroid density (FCD), total optical energy (TOE), and average flash area (AFA); please see the example imagery in Fig. 2. The GLM fields were created for the *GOES-16* ABI CONUS sector geostationary projection at 2-km spatial resolution.[2]

To locate thunderstorms in the training data, NOAA/CIMSS ProbSevere files were used. ProbSevere is a machine-learning nowcasting system in the United States for severe weather using radar, lightning, satellite, and NWP data as inputs (Cintineo et al. 2020). The ProbSevere files include the centroid time and latitude/longitude of radar-identified convective cells every 2 min. The ProbSevere thunderstorm objects are based on Multi-Radar Multi-Sensor (MRMS) system imagery (Smith et al. 2016). A "thunderstorm" is defined as a convective cell that was successfully tracked for at least 45 min by the automated procedure utilized by ProbSevere (e.g., Cintineo et al. 2014). The thunderstorm must also have had a flash rate of 2 flashes per minute or greater (the flash rate is the sum of the flashes within the object polygon) at some point during the automated tracking period, as inferred from the Earth Networks Total Lightning dataset used by ProbSevere (e.g., Cintineo et al. 2018). The radar object centroid time and location were used to automatically generate ~64 km × 64 km[3] storm-centered image patches from the ABI and GLM CONUS sector data (at 5-min temporal resolution), resulting in 222 854 image patches from 14 745 different storms. Severe hail, wind, and tornado reports were also gathered from NOAA's Storm Events Database (NOAA 2019) and linked to the storm-image patches via ProbSevere radar objects (e.g., Cintineo et al. 2020). Reports were linked to ProbSevere objects using a ±2-min search window around the report time, with each report being associated with the closest radar object centroid within the temporal search window. The severe reports were not used in labeling or validation, but simply as a way to characterize the dataset (see section 2b) and estimate potential lead time (see section 3a).

b. Data labeling and partitioning

The *GOES-16* image patches were generated for each ABI channel and GLM product listed above. An image patch size of 64 km × 64 km was heuristically chosen to represent the "storm scale." The 64 km × 64 km domain was the same for all channels and resulted in 128 × 128 pixel images for ABI CH02 and 32 × 32 pixel images for ABI CH13 and the GLM channels.

[1] This is the reflectance factor that is not corrected for solar zenith angle. The paper refers to this as "reflectance" throughout, for short.

[2] The GLM regridding was performed at 2 km in this paper to ensure the final grids conveyed the full shape/size of the original GLM pixels as they aligned to the ABI fixed grid.

[3] The patch size in kilometers is approximate, as the ABI channel spatial resolutions are nominal and valid at the satellite subpoint. The actual spatial resolution decreases away from the subpoint.

TABLE 2. Summary of raw input data. The horizontal spacing values are for the *GOES-16* satellite subpoint (the point on Earth directly below the satellite; on the equator at 75°W). Abbreviations: Advanced Baseline Imager (ABI), Geostationary Lightning Mapper (GLM), flash-extent density (FED), flash-centroid density, (FCD), total optical energy (TOE), and average flash area (AFA).

Dataset	Time step	Horizontal spacing
GOES-16 ABI 0.64-μm reflectance	5 min	2 km
GOES-16 ABI 10.35-μm brightness temperature	5 min	0.5 km
GOES-16 GLM FED, FCD, TOE, AFA	5 min	2 km

Images created for ABI CH13 brightness temperature and CH02 reflectance enabled manual storm labeling.[4] Human experts performed the labeling using a custom tool built on the React Javascript library (see Fig. 3 for an example). The experts (three of the coauthors of this paper) labeled all images as either "intense" (22 505 images) or "ordinary" (200 366 images) convection based on the presence, or lack thereof, of features within the patch widely accepted as being associated with strong midlatitude convection (e.g., overshooting tops, cold-U, cloud-top divergence, AACP, high visible texture).

MRMS merged composite reflectivity (MergedRef; a column maximum of radar reflectivity at each horizontal grid point G, based on nearby radars observing point G) was also contoured over the images to provide extra context for humans labeling the intensity of a storm, but the CNN only utilizes *GOES-16* satellite data. In the absence of a clear satellite indicator for intense convection, we looked for corresponding strong reflectivity cores (50–60+ dBZ), giving careful attention not to consider MergedRef too highly when radar beam blockage was present, or the storm was on the edge of MRMS domain. Label selections were linked to a database for expedient cataloging of the dataset. While these images were useful for labeling, they were not the same images used to train the CNN—the actual ABI and GLM numerical-data patches were stored in separate files.

Based on the National Weather Service (NWS)-defined severe criteria of hail diameter $\geq$ 1 in. (25.4 mm), wind gust $\geq$ 50 kt (25.72 m s^{-1}), or the presence of a tornado, 55.5% of the intense class images were from severe storms (irrespective of when a severe report occurred), while only 5.6% of the ordinary class images were from severe storms. This analysis confirms that storms that exhibit one or more of the storm top features (e.g., overshooting tops, cold-U, cloud-top divergence, AACP, high visible texture) targeted by the human experts are much more likely to produce verified severe weather than storms where such clearly defined features were absent.

[4] The manual labeling of small storm-centric image patches was elected in this work, rather than an automated pixel-by-pixel labeling of larger images [often used in semantic segmentation; e.g., Ronneberger et al. (2015)] or automated image-patch labeling, because of uncertainty or inconsistency in radar and reports-based datasets. While pixel-by-pixel manual labeling is possible, it is more labor intensive than assigning one label per small image patch (the approach of this paper).

The 29 days of labeled data were divided into three groups—training, validation, and testing. The groups consisted of 18 (70%), 8 (22%), and 3 (8%) days, respectively (Table 1). The proportion of intense-labeled storms was 10.1%, 11.6%, and 14.2%, for the training, validation and testing sets, respectively. An independent set of dates, as opposed to a random method, was used to minimize collinearity between images in each group. In machine learning, the training set is the sample of data used to fit the model. This is how the model learns and encodes spatial features, using backpropagation to minimize the loss function in the training set. Backpropagation computes the gradient of the loss function with respect to each weight of the CNN. The validation set is used to provide an independent assessment of a trained model, which is useful in selecting hyperparameters (see section 2d). However, by choosing hyperparameter values that optimize performance on the validation set, the hyperparameters can be overfit to the validation set, just like model weights (those adjusted by training) can be overfit to the training set. Thus, the selected model is also evaluated on the testing set, which is independent of the data used to fit both the model weights and hyperparameters.

c. *Model architecture*

CNNs use a multilayered architecture to learn spatial features (e.g., Fig. 4). This architecture is typically broken down into three fundamental types of layers: convolutional layers, pooling layers, and fully connected layers. The convolutional and pooling layers turn input into "feature maps," or transformations of the data. The "maps" received by the first convolutional layer are the ABI and GLM image patches. Maps received by deeper layers have been transformed by one or more convolutional filters and activations, creating abstractions of the data.

Convolution is formally defined by Eq. (4) in Lagerquist et al. (2019), which operates spatially and in a multivariate fashion on a set of input grids, encoding spatial patterns that combine the input variables. Each convolutional filter in the model has a different set of weights, which are initialized randomly. Activation is a nonlinear function applied to the feature maps after each convolutional layer, elementwise. The activations are an important step, as a CNN would only learn linear relationships in the data without applying the activations. The nonlinear activation applied after every convolutional layer in this model is the rectified linear unit, ReLU (Nair and Hinton 2010).

After two sets of convolutions and activations, pooling layers are applied, which downsample each feature map

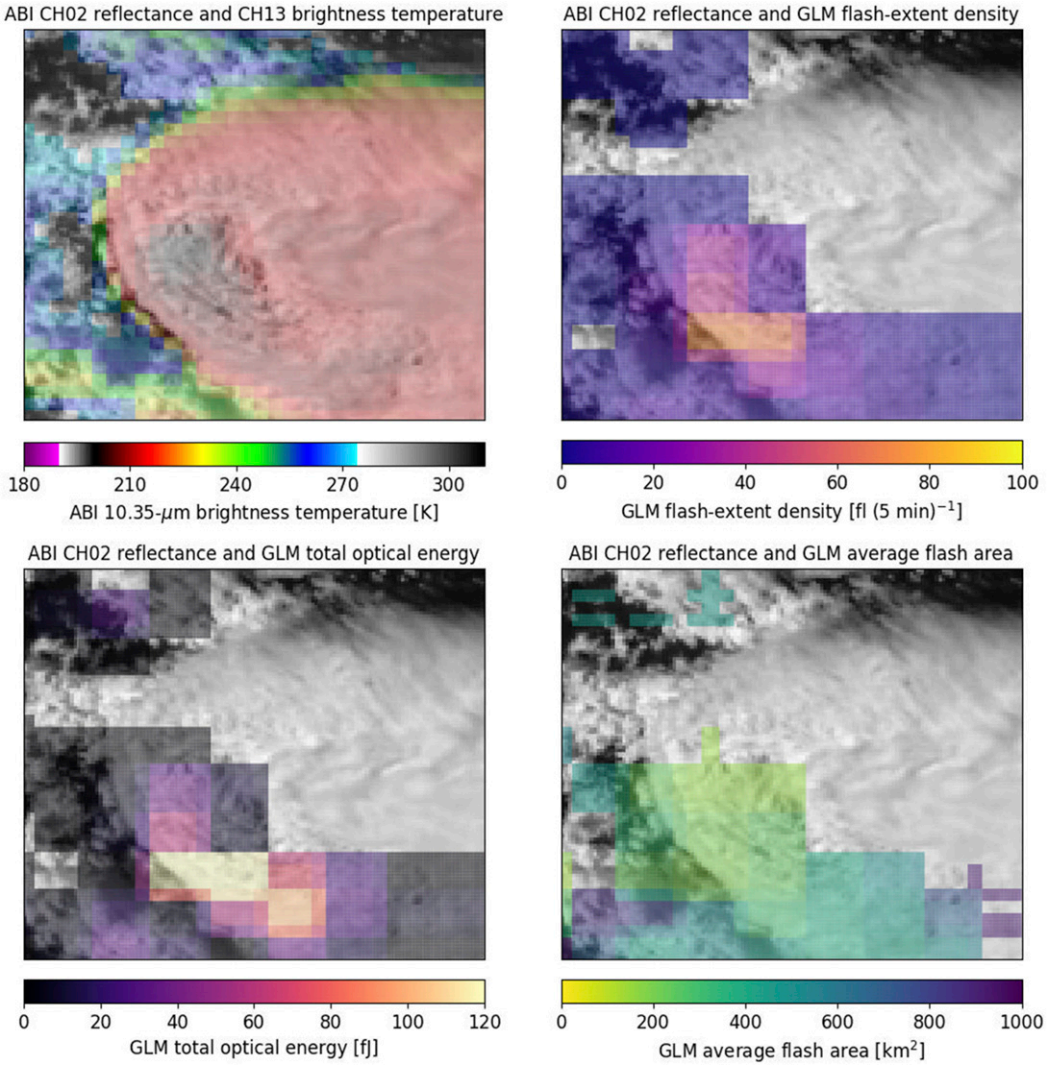

FIG. 2. *GOES-16* ABI and GLM imagery of a supercell in central Texas. ABI 0.64-μm background image overlaid with a (top left) semitransparent 10.3-μm brightness temperature image, (top right) GLM flash-extent density, (bottom left) GLM total optical energy, and (bottom right) GLM average flash area.

independently (e.g., Li et al. 2020). The model of this paper uses a maximum filter with a window size of 2 × 2 pixels, halving the spatial resolution for each pooling (these have become fairly standard choices). This pooling operation enables deeper convolutional layers to learn larger-scale features in the data and helps the model become invariant to small spatial translation in the inputs. The series of convolutions, nonlinear activations, and pooling operations allow the model to learn higher-level abstractions at deeper layers of the network.

After a series of convolution, activation, and pooling layers, the feature maps of the network are flattened into a 1D vector and passed to a series of fully connected layers to create the final predictions. The model of this paper uses ReLU for the activations of the first two fully connected layers and uses the sigmoid function for the activation of the final fully connected layer with a single output, forcing the final prediction to be a probability between [0, 1].

We used the Keras Python API with TensorFlow backend to perform the training and evaluation of CNNs (Chollet 2015). This is a binary classification problem ("intense" or "ordinary" convection are the classes), so the loss function chosen to minimize was binary cross-entropy [Eq. (1)]. The term p_i is the predicted probability of intense convection, y_i is the label (1 if intense, 0 otherwise) for the ith example, N is the number of examples, and ε is the binary cross-entropy, ranging from $[0, \infty)$:

$$\varepsilon = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]. \quad (1)$$

d. *Hyperparameter tuning*

A hyperparameter is a parameter whose value is set before the learning process begins for training a CNN. There are many

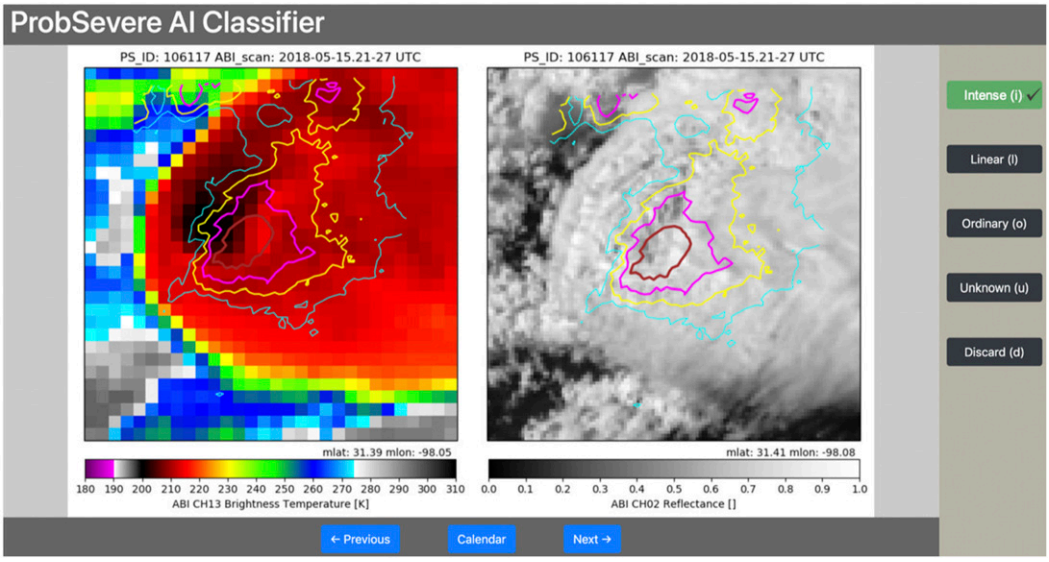

FIG. 3. Example images in the tool that was used to create the labeled dataset of "intense" and "ordinary" convection classes. A different version of these images, without overlays, text, and color bars, was used for training, validation, and testing. Contours are NEXRAD reflectivity from the Multi-Radar Multi-Sensor (MRMS) system. The 30- (cyan), 40- (yellow), 50- (magenta), and 60-dBZ (brown) reflectivity contours are shown. The human-assigned labels are uploaded to a database.

design components to creating a CNN, including the number and types of layers, convolutional filter size, the number of convolutional filters, regularization techniques, image padding techniques, learning rate, mini batch size (the amount of samples the network sees before a weight change is made), activation function, image patch sizes, the number of epochs (passes through the data), and others, not to mention different combinations of input predictors. While many general-purpose CNN architectures exist (e.g., ResNet [He et al. 2016]), we found that starting simple and iteratively building a CNN

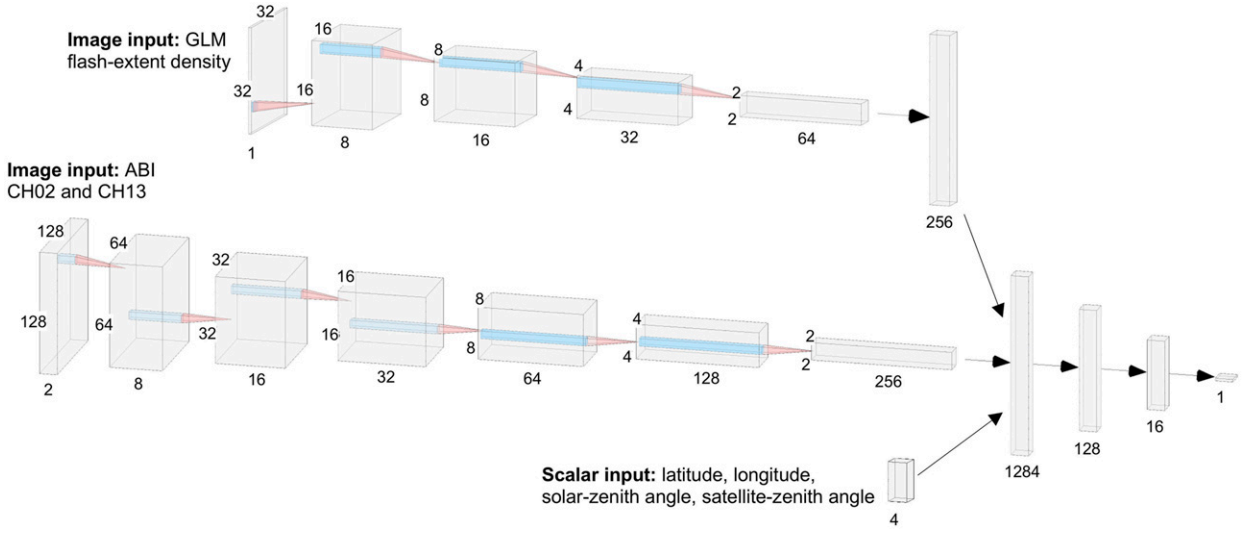

FIG. 4. Schematic of the convolutional neural network described in this paper. The blue boxes and pink pyramids represent 3×3 pixel convolutional filters acting over the feature maps (or input images, initially). The dimensions for the gray boxes indicate the horizontal dimensions (the equal dimensions in each box) and the number of feature maps after 2D-convolutional and maximum pooling layers are applied (or the number of input grids, initially). In the case where one dimension is present, it is the length of the 1D vector. After several blocks of convolutions and poolings, the encoded ABI and GLM features are then "flattened," or made into a 1D vector, and concatenated with the scalar input vector. The concatenated vector is processed through several fully connected layers to generate a probability of intense convection. This image was created at http://alexlenail.me/NN-SVG/AlexNet.html.

TABLE 3. Select hyperparameters used for the training of the convolutional neural network.

Hyperparameter	Value
Loss function	Binary cross-entropy
Learning rate	0.01; reduced by 90% if no improvement in validation loss after 2 epochs
Total number of epochs	14 (early stopping if no loss improvement after 6 epochs)
Batches per epoch	511
Examples per batch	300
Filter window	3×3 pixels for each Conv2D filter
Optimizer	Rectified Adam (RAdam)
Dropout ratio	50% (used for first two fully connected layers only)
Nonlinear activation	Rectified linear unit (ReLU) for all convolutional layers and first two fully connected layers; sigmoid for final fully connected layer.
Padding	Feature maps are zero-padded such that the size of the output feature maps is the same as the size of the input feature maps
Graphics processing unit	One NVIDIA TITAN V

worked best for this problem (e.g., iteratively adding blocks of 2D convolution + pooling layers and other components until performance on the validation data decreased) as opposed to using a more sophisticated architecture. Given the infinite number of CNN hyperparameter combinations, our proposed architecture is perhaps suboptimal, but works well in practice (see Table 3 for the final hyperparameter configuration).

The criterion used to attempt to optimize hyperparameters was the maximum critical success index (CSI) of the validation set. The CSI is the ratio of true positives (''hits'') to the sum of true positives, false positives (''false alarms''), and false negatives (''misses'') for a given probability threshold. It is bound between [0, 1], with 1 representing perfect skill. It is an excellent metric for rare-occurring classes, since it does not reward true negatives. The hyperparameters we attempted to optimize were the: 1) number of convolutional layers, 2) convolutional filter size, 3) number of convolutional filters in the initial convolutional layer, 4) application of the dropout operation to the fully connected layers (Hinton et al. 2012), 5) application of batch normalization to the convolutional layers, 6) application of L_2 regularization (Hoerl and Kennard 2000), and 7) inclusion of multiple GLM fields. Hyperparameter changes that improved the maximum validation CSI (by at least 0.0025) were included in the final model architecture. There is one CSI per probability threshold, so the probability threshold with the maximum CSI was chosen.

Batch normalization (Ioffe and Szegedy 2015) is applied elementwise to each of the feature maps to mitigate the inherent vanishing-gradient problem (see Schmidhuber 2015, section 5.9) in neural networks and speed up learning. However, it did not improve the maximum CSI. The L_2 regularization adds a weight term to the loss function [Eq. (1)] that is the sum of the squared weights in all convolutional layers multiplied by the parameter λ. This method is meant to help the model learn smaller weights and become less sensitive to small changes in the input predictors (i.e., make it more stable to small changes). The values for λ tested were 10^{-4}, 10^{-3}, and 10^{-2}, yet none improved the validation set CSI, perhaps because dropout in the fully connected layers provided sufficient regularization. Some findings that did improve model performance included:

(i) The 3×3 convolutional filters were better than 5×5 filters. The smaller filters may allow the network to learn features at the finest scale in the first convolutional layer, then larger-scale features in deeper layers after pooling has been applied.

(ii) Eight convolutional filters for the initial convolutional layer were better than 16, 32, or 64. This may be because it reduces the number of weights in the CNN, leading to faster convergence.

(iii) Two convolutional layers per block were better than one or three, as two layers allowed the network to learn more complex abstractions at each spatial scale before pooling, whereas one layer did not allow this and three layers per block led to too many weights.

(iv) Dropout applied to the fully connected layers was better than no dropout. Dropout randomly zeroes out fraction F of a layer's outputs, where F is the dropout rate. This is meant to force weights in layer L (with dropout) learn more independently of other weights in layer L and reduce overfitting to the training data. The fractions tested were 0, 0.3, and 0.5 (there was little difference in performance between 0.3 and 0.5). The dropout likely prevented overfitting.

Upon testing various CNN inputs, we found that ABI CH02 reflectance (0.64 μm), CH13 brightness temperature (10.35 μm), and GLM FED, along with the scalar values of satellite-zenith angle, solar-zenith angle, latitude, and longitude, provided the best performance in discerning intense convection in the validation dataset. The inclusion of the TOE, FCD, and AFA from the GLM did not improve performance on the validation dataset. The ABI channels were jointly processed through a set of six convolution and maximum pooling blocks, whereas the FED was processed through a separate set of four convolutional and maximum pooling blocks. The ABI and GLM convolutional bases were then joined with the scalar data and connected to three fully connected layers with 128, 16, and 1 node(s) (Fig. 4). Future work will examine if metrics that are derived from two or more ABI channels (e.g., brightness temperature differences, reflectance ratios, etc.) and/or time series can be used to improve model performance.

One somewhat unique aspect of this model is the combination of two convolutional bases. Initially, FED, CH02 reflectance, and CH13 brightness temperature were separate channels

in one convolutional base (one stack of convolutional and pooling layers), having upsampled FED and CH13 brightness temperature to 0.5-km horizontal grid spacing. The single convolutional base formulation performed poorly compared to a model that excluded GLM. However, when the GLM input was processed using a separate stack of convolutional and pooling layers, the ABI + GLM model performance noticeably improved relative to an ABI-only model (maximum CSI improved by 0.035, or 6.3%). This outcome illustrates that care must be taken when utilizing images from multiple data sources.

e. Model evaluation and interpretation

1) Statistical verification

Standard performance metrics were computed separately for the validation and testing labeled data partitions (i.e., the 64 km × 64 km image patches). The computed metrics include accuracy [Eq. (5)], CSI [Eq. (6)], frequency bias [Eq. (7)], Peirce score (PS) [Eq. (8)], Brier skill score (BSS), and the area under the receiver operating characteristic curve [area under the ROC curve (AUC; Metz 1978]. The ROC curve, performance diagram, and attributes diagram are also presented:

$$\text{POD} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$\text{POFD} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{3}$$

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TP}}, \tag{4}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{5}$$

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{6}$$

$$\text{bias} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FN}}, \tag{7}$$

$$\text{PS} = \text{POD} - \text{POFD}, \tag{8}$$

$$\text{success ratio} = 1 - \text{FAR}. \tag{9}$$

In Eqs. (2)–(7), TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives, defined based on a probability threshold. The CNN produces a single prediction for each image patch, and the probability threshold binarizes the prediction into the "yes" and "no" classes (i.e., for probability threshold p^*, and prediction p, $p \geq p^*$ becomes "yes," and $p < p^*$ becomes "no"). POD is the "probability of detection," "hit rate," "true positive rate," or "recall." POFD is the "probability of false detection" or "false positive rate." FAR is the "false alarm ratio" or "false discovery rate."

The accuracy simply measures how well a given probability threshold is able to discriminate between intense and ordinary convection. It ranges between [0, 1], with 1 being perfectly accurate. The training dataset consists of 10.1% intense-labeled samples, so accuracy can be trivially optimized to 0.899 by always predicting "no." The CSI is accuracy without correct nulls or TNs, and ranges between [0, 1], with 1 being perfect. The frequency bias ranges from [0, ∞), with 1 being perfectly unbiased, values > 1 meaning that the intense label is predicted more often than it occurs, and values < 1 meaning that the intense label is predicted less often than it occurs. The Peirce score [Eq. (8)] is the POD minus the POFD, which ranges from [−1, 1], with 1 being perfect, 0 indicating no skill, and −1 indicating a POD = 0 and a POFD = 1 (i.e., no TPs and the maximum amount of FPs for a given probability threshold).

The BSS [Eq. (10)] examines the Brier score (BS) of the model versus a reference Brier score, which is Eq. (11) evaluated with f_t equal to the frequency of the intense class in the training data. The BSS ranges between (−∞,1], with 1 being perfect and 0 indicating no skill compared to the reference Brier score, while decreasing values (toward −∞) indicate deterioration of skill compared to the reference Brier score:

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{reference}}}. \tag{10}$$

The Brier score itself measures the mean squared probability error [Eq. (11)]:

$$\text{BS} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2, \tag{11}$$

where f_t is the probability that was forecast, o_t is the actual outcome of the event at instance t (0 if ordinary and 1 if intense) and N is the number of forecasts. The Brier score ranges from [0,1] with a perfect score being 0. While the BS and BSS combine all probability forecasts to compute scalar metrics of skill conditioned on the forecasts, an attributes diagram (Hsu and Murphy 1986) provides this information per probability bin. In this way, users can see which forecast probabilities are well calibrated and which are not.

The ROC curve plots the POD versus the POFD, from which the area under the curve can be computed. The ROC curve examines how well the model does at distinguishing between classes (1 = perfect separation; 0.5 = no separation; a random model) and is not sensitive to poorly calibrated model predictions. Whereas the accuracy, bias, CSI, and PS are computed on a single probability threshold, the AUC is integrated over all probability thresholds, giving a more holistic characterization of model performance.

The performance diagram plots the POD versus the success ratio [Eq. (9)] for different probability thresholds, with greater CSI in the top-right corner, corresponding to greater POD and greater success ratio (or lesser FAR). For the ROC curve and the performance diagram, each point corresponds to one probability threshold.

2) Intense convection probability grids

In addition to the statistical metrics described in the previous paragraph, intense convection probability grids were created for a number of additional independent scenes from 2019 (all of the training, validation, and testing data were from 2018). To create the ICP grids, a sliding-window approach was used. Within each scene, moving in both the latitudinal and longitudinal directions, a 64 km × 64 km window was used to extract

the ABI and GLM data patches. The stride of the movement for the sliding window was four 2-km ABI pixels. This creates an oversampling of predicted probabilities, with one value every 8 km, whereas the model was trained on storm-centric 64 km × 64 km patches. Contours of selected probability thresholds were then derived from the resulting grid of ICP and subsequently overlaid on the corresponding ABI imagery. Because the model was trained on storm-centric patches, it also learned the parallax[5] relationship between the satellite data and radar-identified storms. This is evident in small displacements of ICP contours to the south and east of the highlighted storms, which can be observed at the higher satellite viewing angles in storms (as shown in Figs. 13 and 14). The ICP contours can also include sections of the storm that may not be intense, which is due to the fact that each 64 km × 64 km patch generates a single probability; that is, the probability is representative for the entire patch. It should be noted that the verification metrics mentioned in section 2e(1) were computed only for the 64 km × 64 km images of the validation and testing datasets, not for the ICP grids, which would require truth labels at every pixel. Thus, the ICP grids provide a more qualitative yet visual aspect of verification. Nevertheless, this sliding-window approach is one possible technique enabling predictions that would not require a radar network.

3) SALIENCY MAPS AND LAYER-WISE RELEVANCE PROPAGATION

Saliency maps (Simonyan et al. 2014; McGovern et al. 2019) and relevance maps (Binder et al. 2016) are another form of model analysis utilized. The objective is to identify the spatial features within each input image that most influence the model results. The saliency of predictor x at image coordinate (i, j) with respect to the intense-convection prediction p, is $\partial p/[\partial x(i, j)]$. Saliency uses backpropagation to determine how changes in each $x(i, j)$ impact the model's output probability. One disadvantage is that it is a linear approximation around $x_{(i,j)}$, meaning the saliency indicates how the model prediction changes when x is perturbed only slightly. It can be both positive and negative. As a complement to saliency, layer-wise relevance propagation (LRP; Alber et al. 2019) is a framework that also uses backpropagation to identify the most relevant or important pixels; that is, the pixels that contribute the most to a given prediction. Relevance (like saliency) indicates how much each predictor contributes to the positive class only for the hyperparameters we chose (see section 3c). One other important difference between saliency and relevance is that saliency indicates which predictors are most important for changing the prediction, while relevance signals which predictors (or regions of those predictors) are most important for the prediction actually made.

4) PERMUTATION TESTS

Finally, in an effort to rank predictor importance, two permutation tests were applied to the trained CNN: Breiman (2001, hereafter B01) and Lakshmanan et al. (2015, hereafter L15). In B01, samples are permuted (or randomized) one predictor at a time; the computed loss in performance is recorded and compared to the performance on unpermuted data; each predictor is returned to its unpermuted state before the next predictor is permuted. The permutations occur by shuffling spatial maps of a predictor x across examples, so that after permutation, each example is matched to the wrong map of x, but the correct maps of all other predictors. This removes the statistical linkage between the permuted predictor x and the output classification. After each predictor is permuted individually, the most important predictor is the one which decreased the performance the most (i.e., incurred the highest "cost").

The method of L15 carries B01 method a step further, by executing successive permutations. It ranks predictor importance in this way:

(i) The most important predictor (rank of $k = 1$) is obtained by using the permutation method of B01.
(ii) Given the k most important predictors, the $(k + 1)$th most important predictor can be found by keeping the k predictor(s) permuted and permuting each of the remaining predictors, one at a time. The predictor that results in the greatest loss in skill carries the $(k + 1)$th rank and remains permuted for the remainder of the test.

If performance diminishes appreciably when a predictor is permuted, this indicates that the predictor is important. If performance does not decline appreciably, the predictor is either unimportant or some information in the predictor is redundant with information contained in other predictors. The L15 method helps discern correlated predictors. For example, for two very important and highly correlated predictors, x_1 and x_2, the B01 method may rank them both as unimportant (relative to other predictors), since permuting only one destroys very little information. Once x_1 has been permanently permuted in the L15 method, x_2 should immediately be considered important, since the redundant information in x_1 has been removed by permutation. However, there is a chance that this may not happen until later iterations of the L15 algorithm, causing neither x_1 nor x_2 to be considered as highly important relative to other predictors.

3. Results

a. *Verification metrics*

The scalar evaluation metrics are summarized for the validation and testing datasets in Figs. 5 and 6, respectively. The probability threshold used for both datasets was the threshold that maximized CSI on the validation data, which was 51%. The statistical evaluation was also partitioned into "day" and "night" using a solar-zenith angle threshold of 85°; "day" is solar-zenith angle less than or equal to 85°, and "night" is solar-zenith angle greater than 85°. The model CSI was greater at night, which may be a result of the fact that there is more mature convection present at night, which may be easier for the model to distinguish, even in the absence of the CH02 reflectance. This may be a result of reduced GLM detection efficiency during the daytime, as well.

[5] Parallax is a displacement in the apparent position of clouds viewed along two different lines of sight (in this case, lines of sight from the geostationary satellite and the ground-based radar).

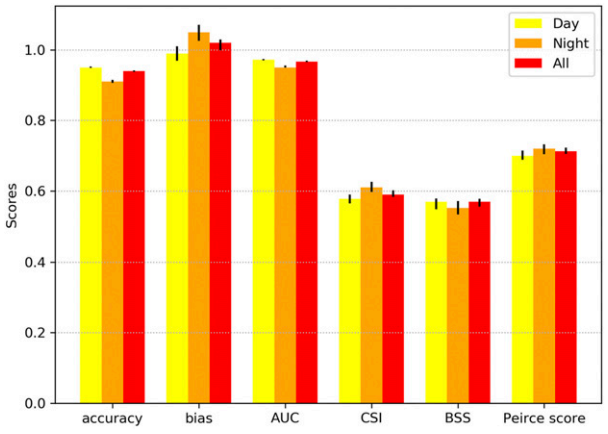

FIG. 5. Summary of scalar verification metrics for the validation data, partitioned by time of day. "Day" signifies the part of the sample with a solar-zenith angle ≤ 85°, whereas "Night" signifies the part of the sample with a solar-zenith angle > 85°. "All" is for the entire sample. These metrics were computed based on a probability threshold of 51% (which maximized validation data CSI). Black bars are 95% confidence intervals determined by bootstrapping 1000 times.

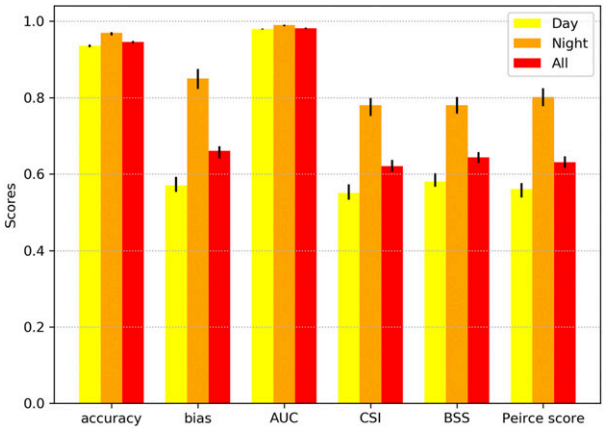

FIG. 6. As in Fig. 5, but for the testing data. The metrics were computed based on a probability threshold of 51% (which maximized validation data CSI).

For the entire validation dataset (combined night and day), the ROC curve shows an inflection point at POFD = 0.10 and POD = 0.95 with a Peirce score > 0.8 (Fig. 7), while the performance diagram (Fig. 8) shows a maximum CSI of 0.59 and bias of 1.01 at ICP threshold = 51%. The attributes diagram in Fig. 9 shows that the model is generally well-calibrated for the validation data, but the CNN exhibits some overforecasting bias between the 40% and 90% probability bins (i.e., the frequency of events in these probability ranges is less than the forecast probability). The testing dataset predictions, while skillful, exhibit a very large underforecasting bias. It is unknown why there is a large difference in the calibration between the validation and testing datasets. This is possibly due to differing frequencies of certain storm morphologies in the two datasets (e.g., supercells, linear storms), but more work is needed to discern if that is the case.

Since severe weather reports were linked to the storm-image patches in the testing dataset, a lead time analysis was performed to assess the ICP's potential to provide an alert prior to the occurrence of severe weather. For the three days in the testing set, 318 independent storms produced severe reports. Lead time to the initial severe report was measured in minutes from the first occurrence of the 50% and 90% ICP thresholds. Of the 318 storms, 153 reached 50% and 126 reached 90% prior to the initial report. Using the bootstrapping technique (Efron and Tibshirani 1986) 5000 times, 95% confidence intervals for the median lead time were created for each ICP threshold. At the 50% ICP threshold, the median lead time to the initial severe report was 24 min (95% confidence interval: 18–30 min); at the 90% ICP threshold, the median lead time was 21 min (95% confidence interval: 18–26 min). While a limited sample, these numbers are comparable to the lead times to initial severe weather reports recorded in Bedka et al. (2018) when measured from AACP occurrence.

b. *Intense convection probability grids*

ICP grids (Figs. 10–14) were created for several independent scenes, using the method in section 2e. The contours created from the grids, when overlaid on ABI imagery, can provide insight on model performance and may be an effective way to visualize results for eventual users of the product. The selected independent cases encompass a range of meteorological conditions, geography, and satellite viewing angles. Additional cases, with animations, are available on the Cooperative Institute for Meteorological Satellite Studies (CIMSS) Satellite Blog (Cintineo 2019). The background ABI image utilized in the ICP grids is the 0.64-μm reflectance and 10.35-μm brightness temperature

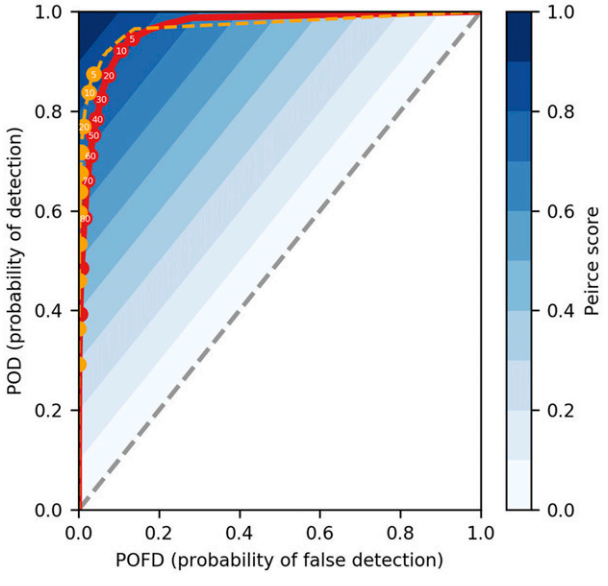

FIG. 7. ROC curve and Peirce score for the validation and testing datasets (solid red and dashed orange lines, respectively). Red and orange circles represent the locations of select probability values (5%, every 10% from 10% to 90%, and 95%). Python code from Lagerquist and Gagne (2019) was used to help create the plot.

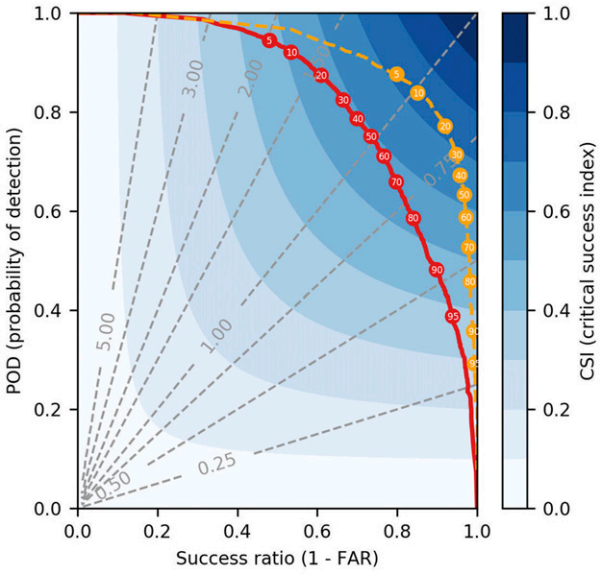

FIG. 8. Performance diagram for the validation and testing datasets (solid red and dashed orange lines, respectively). Intersections with the dashed gray lines indicate the frequency bias. Red and orange circles represent the locations of select probability values. Python code from Lagerquist and Gagne (2019) was used to help create the plot.

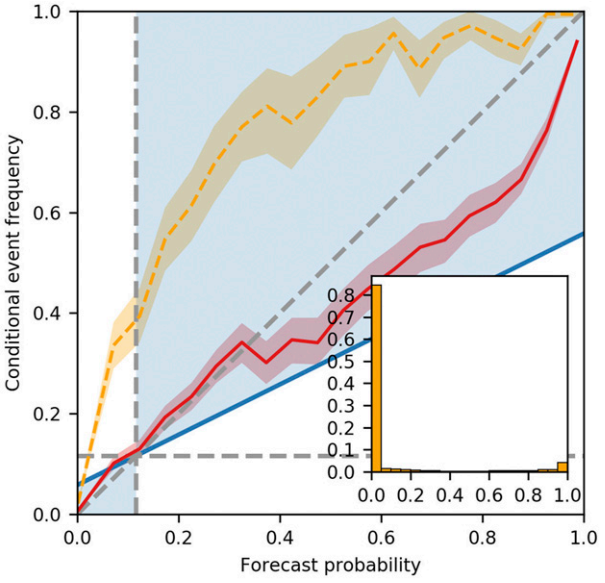

FIG. 9. An attributes diagram for the validation and testing datasets. Solid red and dashed orange lines represent the mean for the validation and testing datasets, respectively, while the shaded red and orange regions denote 95% confidence intervals determined by bootstrapping 1000 times. The inset image shows the frequency of probability forecasts for the testing dataset only. The diagonal gray 1-to-1 line represents perfect reliability or forecast calibration. The horizontal gray line represents the "climatology" or frequency of the intense convection class for the training dataset (also called the line of "no resolution"). The diagonal blue line represents the line of "no skill" with respect to climatology, or where Brier skill score is zero. This line separates the area where forecasts contribute positively to the Brier skill score (shaded blue) and where forecasts contribute negatively to the Brier skill score (white). Python code from Lagerquist and Gagne (2019) was used to help create the plot.

"sandwich" product (unless otherwise stated). Sandwich image composites are created by stacking the reflectance and brightness temperature images with transparency and brightness adjustments, which allows human experts to simultaneously extract textural information from the visible reflectance image and temperature information from the infrared image (Valachová and Setvák 2017). All background images are from *GOES-16* CONUS scans, unless stated otherwise. Reports are plotted on a given image if the time of occurrence is within 60 min after the start of the ABI scan. Figure 10f annotates examples of overshooting tops, the cold-U, and AACP signatures. However, we recommend Homeyer et al. (2017) and Bedka et al. (2018) to readers who desire to become more familiar with the visual identification of these phenomena.

1) WYOMING—10 SEPTEMBER 2019

Figure 10 shows convective storm development in eastern Wyoming between 1941 and 2301 UTC 10 September 2019. At 1941 UTC (Fig. 10a), the 50% ICP value is exceeded in the vicinity of overshooting tops. By 2011 UTC (Fig. 10b), the anvil cloud has expanded greatly, overshooting tops are still present, strong brightness temperature gradients are evident on the anvil edge, the ICP is ≥90% for much of the cloud, and hail and tornado reports are imminent. At 2236 UTC (Fig. 10d), the storm approaching the Nebraska border has an ICP ≥ 90% and is associated with more tornado reports. A developing storm with ICP ≥ 50% is present in the southwest part of the domain at 2236 UTC, which will also turn tornadic. At 2246 UTC (Fig. 10e), the storm to the southwest continues to develop with noticeable overshooting tops. By 2301 UTC (Fig. 10f), the southwestern storm attains an ICP ≥ 90%, while the storm to its northeast still exhibited overshooting top/cold-U/AACP features and ICP ≥ 90%.

2) MISSOURI—26 AUGUST 2019

On 26 August 2019, a strong multicellular line of storms was surging southeastward through the Kansas City, Missouri, metropolitan area around 1601 UTC (Fig. 11a) with a "bubbly-like" texture in the visible reflectance associated with overshooting tops (the ICP of the Kansas City storm was ≥90%). Shortly thereafter, multiple severe wind reports were recorded south of Kansas City, Missouri. By 1801 UTC, two elevated ICP regions with overshooting tops and gravity wave–like patterns are apparent (Fig. 11b). Multiple severe wind reports were associated with both high ICP regions shown in the 1921 UTC image (Fig. 11c). Later, the western storm segment was moving into Arkansas with a strong overshooting top, visual evidence of an AACP, and ICP ≥ 90% (Fig. 11d). The eastern storm weakened, but another storm quickly developed in its wake with a maximum ICP ≥ 50% and a very pronounced overshooting top and thermal couplet (Fig. 11d). Immediately thereafter, the new storm intensified (ICP ≥ 90%; Fig. 11e),

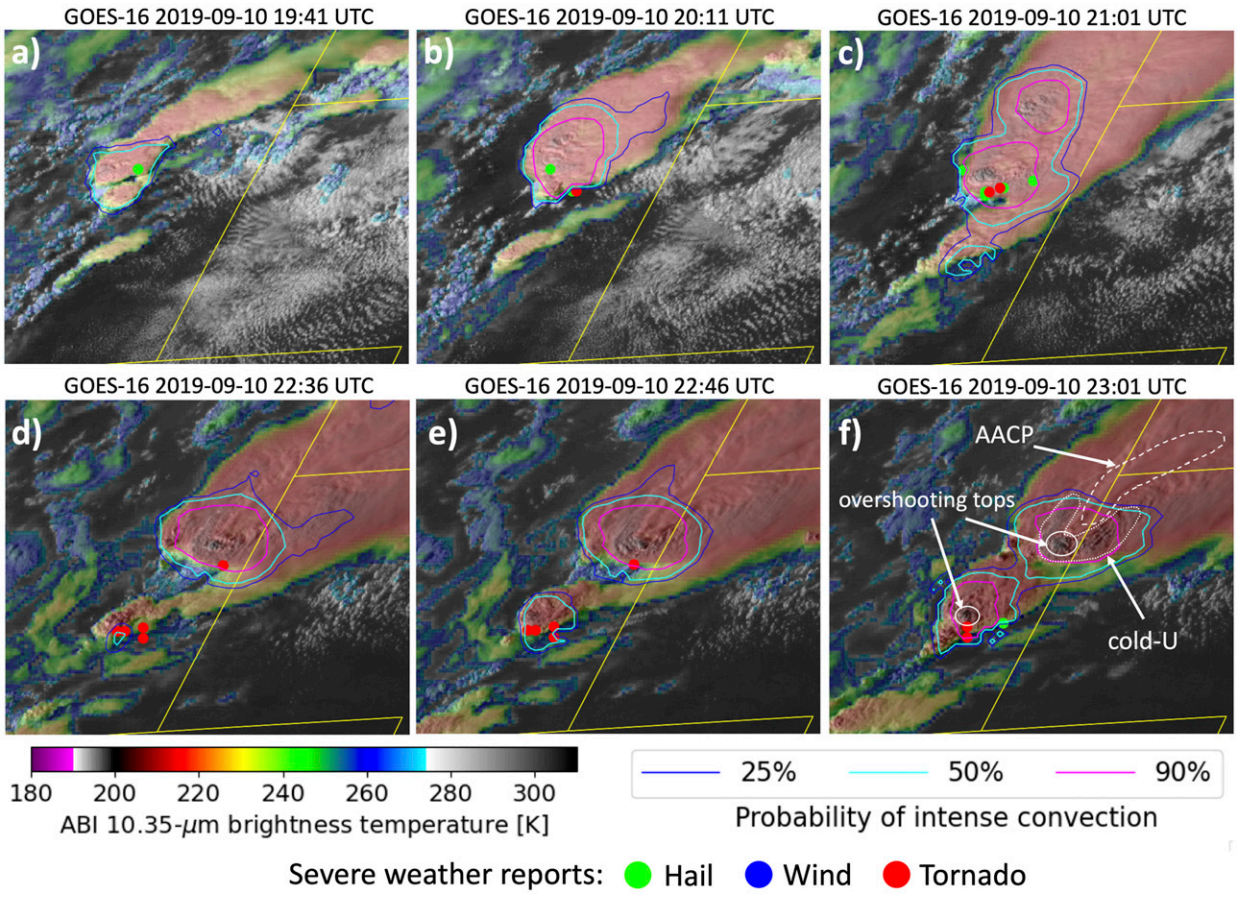

FIG. 10. (a)–(f) A series of intense convection probability contours for storms in Wyoming and Nebraska on 10 Sep 2019. The *GOES-16* ABI visible reflectance and a semitransparent infrared window image are used as the background. Severe weather reports (filled circles) occurred within 60 min after the satellite scan time. Panel (f) depicts examples of the above-anvil cirrus plume (AACP), cold-U signature, and overshooting tops.

with severe hail and wind reports following. The ICP of the storm that moved into Arkansas decreased from above 90% to below 25% (Fig. 11f), consistent with loss of robust textural and thermal patterns. No severe reports were associated with this storm after the ICP dropped below 25%. This example illustrates that the CNN results are consistent with manual interpretation of ABI imagery even when merging anvils from multiple updraft regions complicate the scene.

3) KANSAS/MISSOURI—15–16 AUGUST 2019

On 15–16 August 2019, a cold front initiated very strong storms in northern Kansas. Between 2316 (Fig. 12a) and 0041 UTC (Fig. 12c) the model produces high probabilities (ICP ≥ 50%) in regions associated with overshooting tops and strong brightness temperature gradients on the edge of anvil clouds. In the absence of sunlight, ABI CH13 brightness temperature and the GLM FED are the only image inputs to the CNN (Figs. 12d–i), as the CH02 reflectance becomes a trivial predictor (i.e., it contains a value of zero everywhere). Even in the absence of sunlight, the CNN continues to provide results that are consistent with human interpretation of the imagery, as the model favors regions with overshooting tops and cold-U features, which were generally associated with severe reports (Figs. 12d–i). Robust FED cores were also present, which boosted the ICP, particularly for the storms in Missouri (Figs. 12g–i).

4) ARIZONA—23 SEPTEMBER 2019

In response to moisture and instability associated with a 500-hPa shortwave trough, numerous storms developed in western Arizona on 23 September 2019. At 1631 UTC, the ICP was ≥50% for two of the storms, likely due to the presence of clear overshooting tops and moderate-to-strong brightness temperature gradients around the cloud-top edges near the primary overshoot region (Fig. 13a). By 1706 UTC, the westernmost storm had an expanded area of ICP ≥ 50%, while the eastern storm ICP decreased to <25% as cloud-top temperatures warmed, the textural features softened, and the brightness temperature gradient weakened (Fig. 13b). By 1721 UTC, the western storm intensified (ICP ≥ 90%), consistent with the appearance of a pronounced overshooting top and AACP (Fig. 13c). While features such as overshooting tops, the cold-U

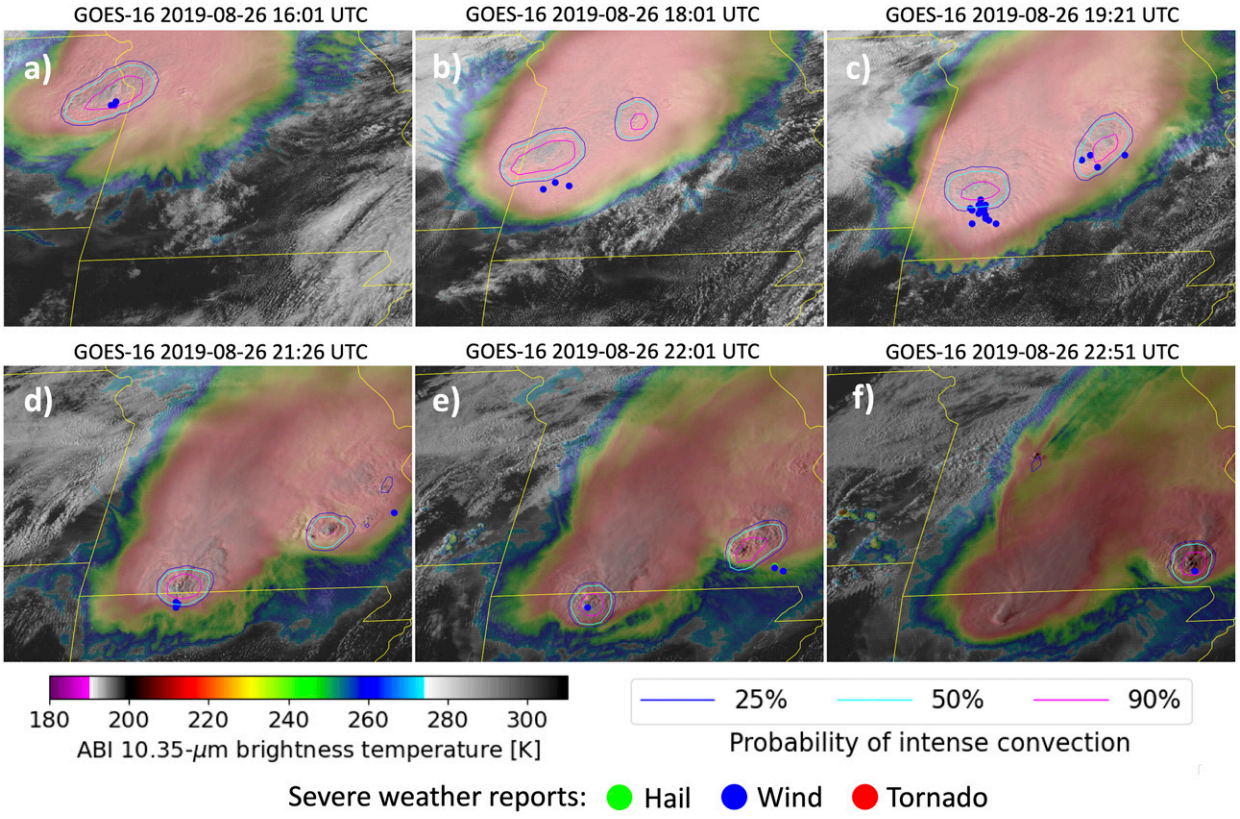

FIG. 11. As in Fig. 10, but for storms in Missouri on 26 Aug 2019. Severe weather reports (filled circles) occurred within 60 min after the satellite scan time.

pattern, and brightness temperature gradients help the CNN produce good predictions, ambiguity in such features may also lead to a bad prediction (e.g., the easternmost storm in Fig. 13a).

5) ALASKA—28 JUNE 2019

At 0249 UTC 28 June 2019, the NWS in Juneau, Alaska, issued the office's first ever severe thunderstorm warning. Since this scene was outside of the GLM field of view, a separate CNN was trained with ABI CH02 reflectance and CH13 brightness temperature images (i.e., no GLM), along with the scalar data discussed in section 2d. The new CNN was deployed on this scene using *GOES-17* 1-min mesoscale ABI scans (but the CNN was trained using *GOES-16* data). Shortly before the storm was warned, it exhibited a cold-ring feature (Setvák et al. 2010), which is more apparent in animations on the CIMSS Satellite Blog (Bachmeier 2019). Despite lower probabilities, the CNN correctly discriminates the intensity of the storm relative to the surrounding convection, with the storm attaining a maximum ICP of 36% at 0243 UTC, while all neighboring convection exhibited ICP $< 5\%$. The lower probabilities may be, in part, due to the absence of the GLM or the very high satellite viewing angles which were not present in the training data. Nevertheless, this example demonstrates that the CNN may be able to generalize reasonably well to new geographic locations and satellites (however, a model without sensor-specific scalar data would still need to be evaluated).

c. *Saliency and relevance maps*

The saliency and relevance were computed for each two-dimensional predictor using a number of samples from the testing dataset. The pixel-wise saliency and relevance values for ten true positive storm samples (Figs. 15 and 16) reveal important features the model has learned. Saliency depicts how changes in a predictor (increasing or decreasing the pixel values) will increase the final probability of intense convection, while the relevance quantifies the degree to which each pixel in each predictor contributes to the predicted probability. The relevance calculations use the "alpha-beta" rule of $\alpha = 1$ and $\beta = 0$ (see Montavon et al. 2019), which does not yield negative relevance scores, but resulted in more coherent output than rules with $\beta > 0$.

Features conducive to intense convection that were identified in the CH13 brightness temperature relevance map for ten true positives (Figs. 15 and 16) include strong overshooting tops (patches A, D, E, F, G, H, and I), portions of thermal couplets and cold-U patterns (patches E and F), strong cloud-edge brightness temperature gradients (patches B, C, D, and H), and warm clear air pixels around anvil clouds (patches F, H, I, and J). While robust overshooting tops and cold-U features have been known for decades to be associated with

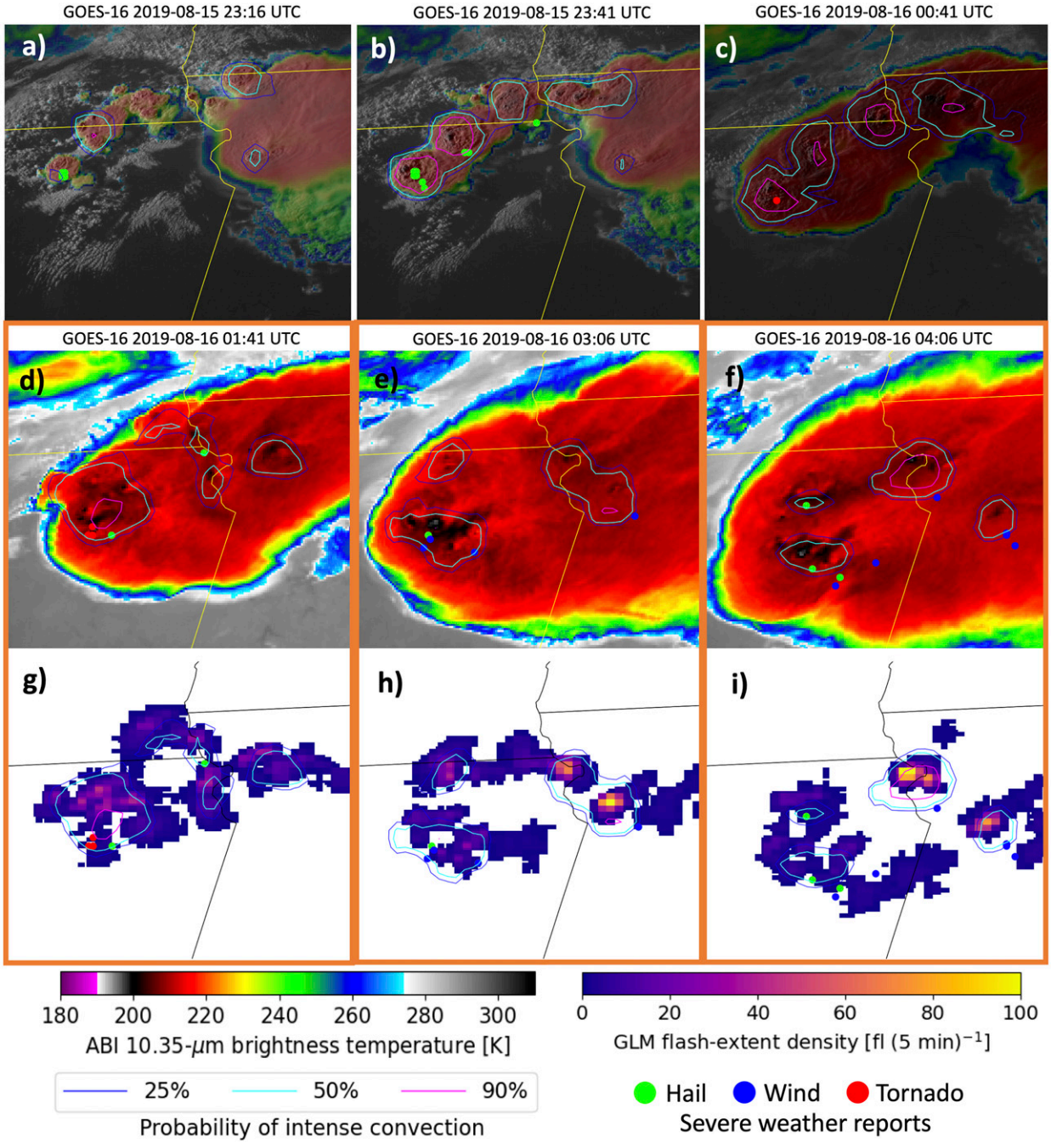

FIG. 12. A series of intense convection probability contours for storms in Kansas and Missouri on 15–16 Aug 2019. (a)–(c) The *GOES-16* ABI visible reflectance and a semitransparent infrared window image are used as the background when sunlight is present. In the absence of sunlight, (d)–(f) the infrared window alone serves as the background, while (g)–(i) the GLM flash-extent density is also shown for the corresponding image in the second row. Each orange rectangle encapsulates the same ABI scan time. Severe weather reports (filled circles) occurred within 60 min after the image time.

intense convection, it is encouraging that the CNN correctly learned and encoded elements of these features. Cloud-edge brightness temperature gradients are less known to be associated with intense convection, yet the model asserts that these are important features in intense storms, even though the human experts did not consciously consider this feature when labeling the images. Furthermore, the model correctly asserts other features known to be associated with intense convection (e.g., overshooting tops), which provides credibility that strong cloud-edge brightness temperature

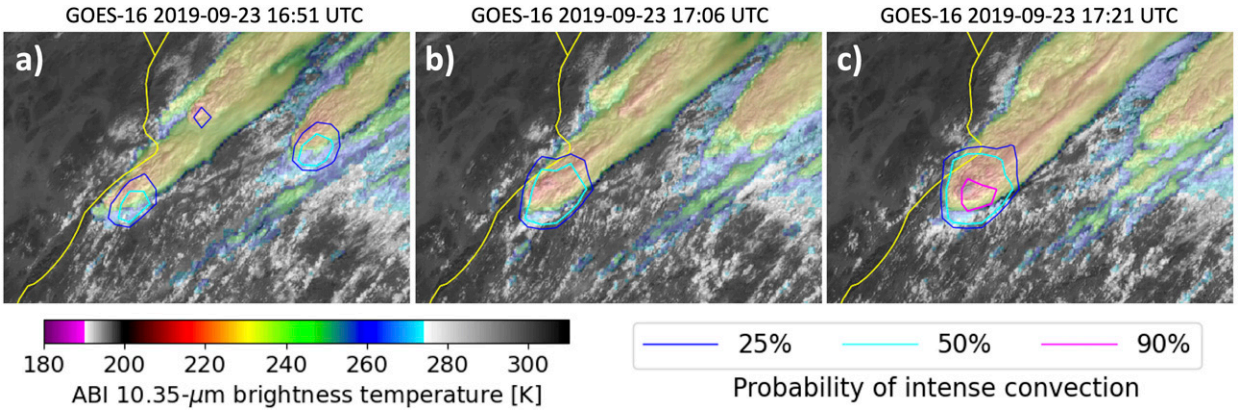

FIG. 13. As in Fig. 10, but for storms in Arizona on 23 Sep 2019. This storm was warned by the U.S. National Weather Service, but no severe hazards were reported.

gradients are not simply an important feature discovered by accident.

The CH13 saliency indicates which pixels to make colder (blue) or warmer (red) in the 10.35-μm brightness temperature to increase the ICP (Figs. 15 and 16). While all storm samples indicate that colder overshooting tops would help, patches C, E, and H indicate stronger cloud-edge brightness temperature gradients would be conducive to higher ICP.

The relevance maps for CH02 reflectance indicate that the CNN identifies ''bubbly-like'' texture features in overshooting tops and cloud tops in general as important (Figs. 15 and 16), as well as less cloudy pixels near the edge of anvil clouds (patches A, B, I, and J)—the latter perhaps highlighting the importance of storm isolation. While not shown, the saliency for CH02 reflectance demonstrated that higher texture in regions within and near overshooting tops (e.g., emanating gravity waves) would increase the ICP of the samples. In other words, the model has learned that increased texture in relatively high-texture regions is important, which is a similar basis for previous visible cloud-top texture rating research (Bedka and Khlopenkov 2016) and agrees well with the conclusion of Sandmæl et al. (2019), that higher texture is correlated with stronger upward motion. More work is needed to evaluate the effect of solar-zenith angle on the contribution of CH02 reflectance to the CNN.

From a lightning-mapping perspective, the relevance maps for the GLM flash-extent density (Figs. 15 and 16) seem to indicate that higher values of flash-extent density are more

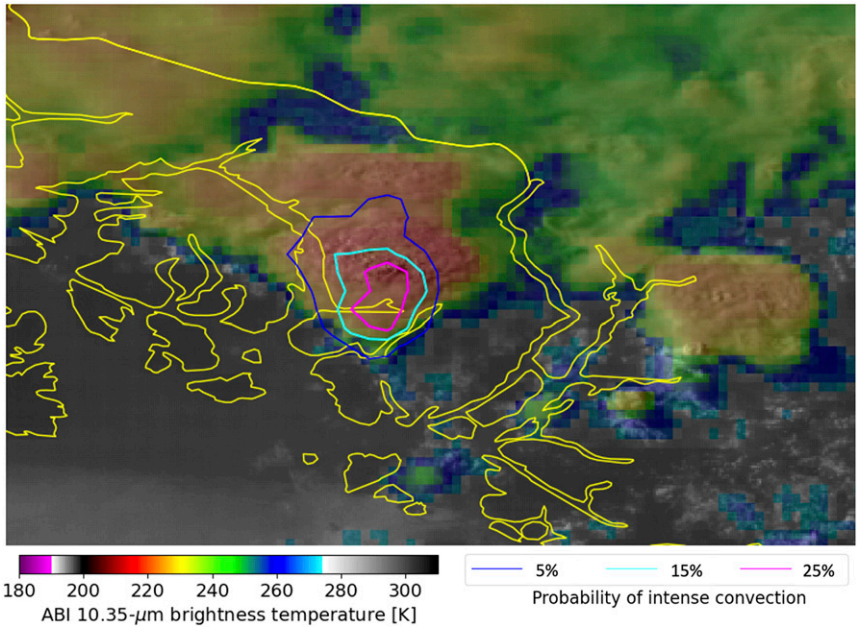

FIG. 14. As in Fig. 10, but for storms in the Alaska Panhandle on 28 Jun 2019. Note that the contoured probabilities are for the 5%, 15%, and 25% thresholds.

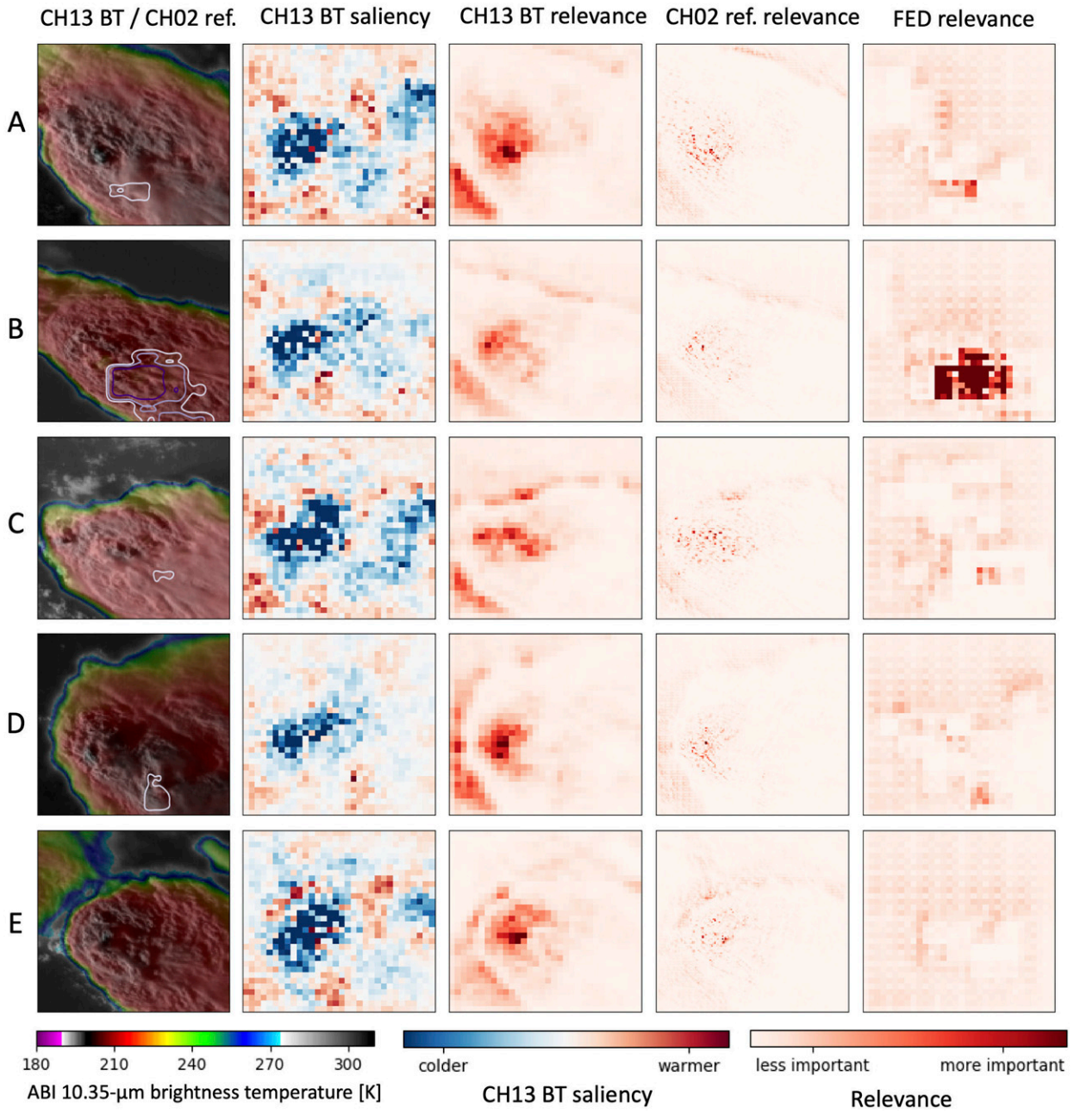

FIG. 15. Saliency and relevance plots for ABI CH13 brightness temperature (BT) and LRP plots for ABI CH02 reflectance (ref.) and GLM flash-extent density (FED) for five true positive storm samples from the validation dataset shown in rows labeled from A to E. The images in the first column are CH13 BT/CH02 reflectance "sandwich" imagery with GLM flash-extent density contours for 10, 20, and 40 flashes per 5 min overlaid in shades of purple. The predicted probability for each image patch was >99%.

relevant than lower values, in general, with marked increases in relevance where flash-extent density is at least 10 flashes per 5 min (see patches B, G, I, and J). The saliency map for flash-extent density is not shown, but was much noisier than the ABI channels, with no clear patterns emerging, making physical interpretation difficult.

From the five false positive storm patches in Fig. 17 (each with a probability > 87%), similar saliency and relevance patterns emerged as being important, such as overshooting tops, brightness temperature gradients, and more textured CH02 reflectance. Compared to the ten true positives, the false positives appear to have weaker overshooting tops in the CH13 brightness temperature and less overall texture in the CH02 reflectance. For the five false positives shown, regions of high relevance were found for the GLM flash-extent density, indicating perhaps that the model erroneously put too much

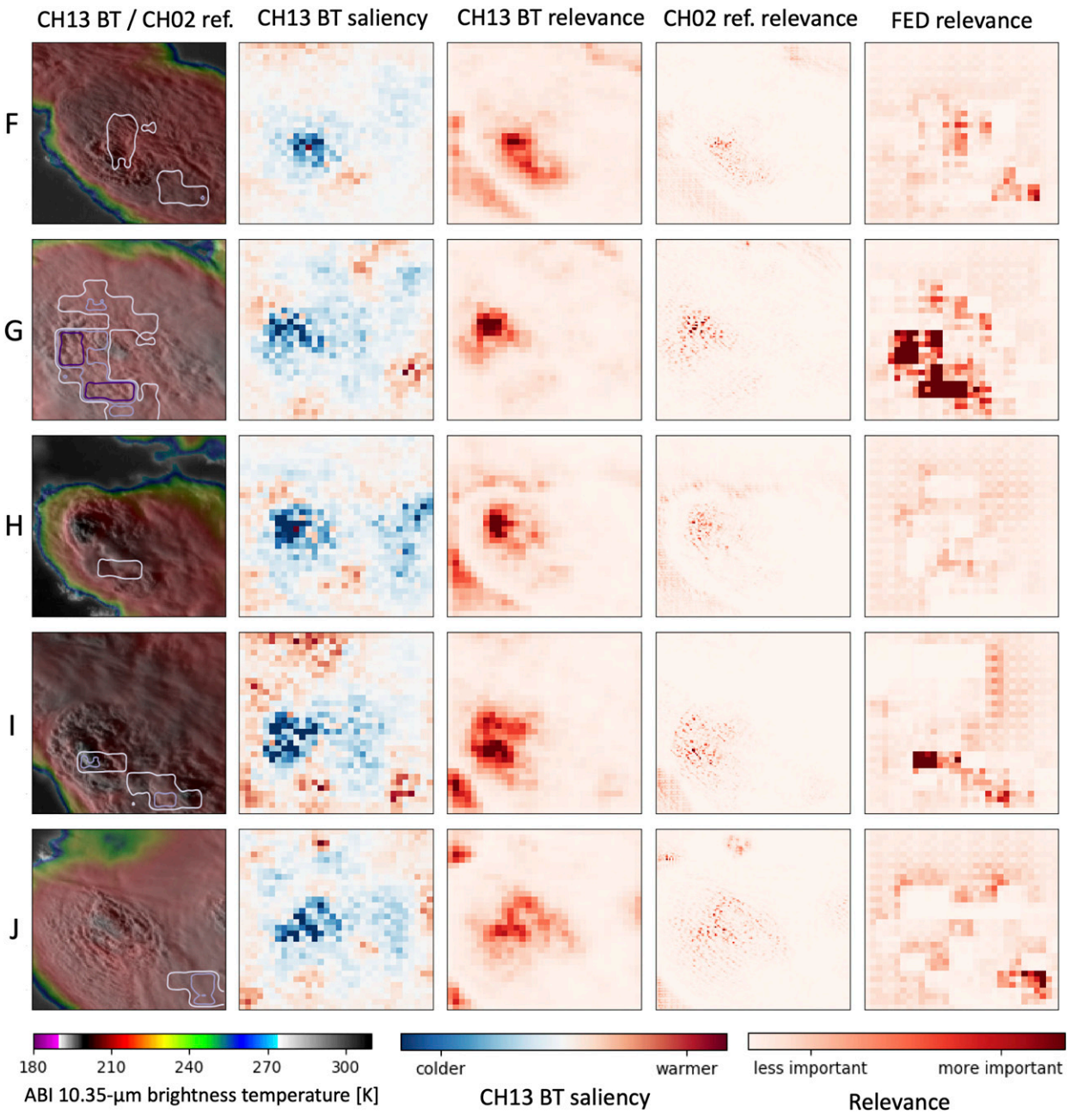

FIG. 16. As in Fig. 15, but for five different true positive samples from the validation dataset. The predicted probability for each image patch was >99%.

importance on these regions of elevated flash-extent density, contributing to the false positive predictions.

Because the hyperparameters chosen for the LRP analysis only show which pixels contribute to the probability of the intense convection class, relevance is near zero everywhere for each predictor for the five false negatives in Fig. 18, as each storm patch had a probability < 1%. Compared to the true positives, these storm patches had less area of cold brightness temperatures, less pronounced (or absent) overshooting tops, and smaller areas of high-textured CH02-reflectance. Two of the patches in Figs. 18b and 18d had diminished GLM flash-extent density, compared with most of the true positives. These patches also appear to be at less mature stages of development than the true or false positives. The CH13 brightness temperature saliency shows that larger and colder cloud-top regions would increase the probability of intense convection.

Interestingly, one feature that the model did not appear to explicitly associate with intense convection is the AACP (Bedka et al. 2018), particularly its unique manifestation in the CH02 reflectance. However, model testing indicates that many

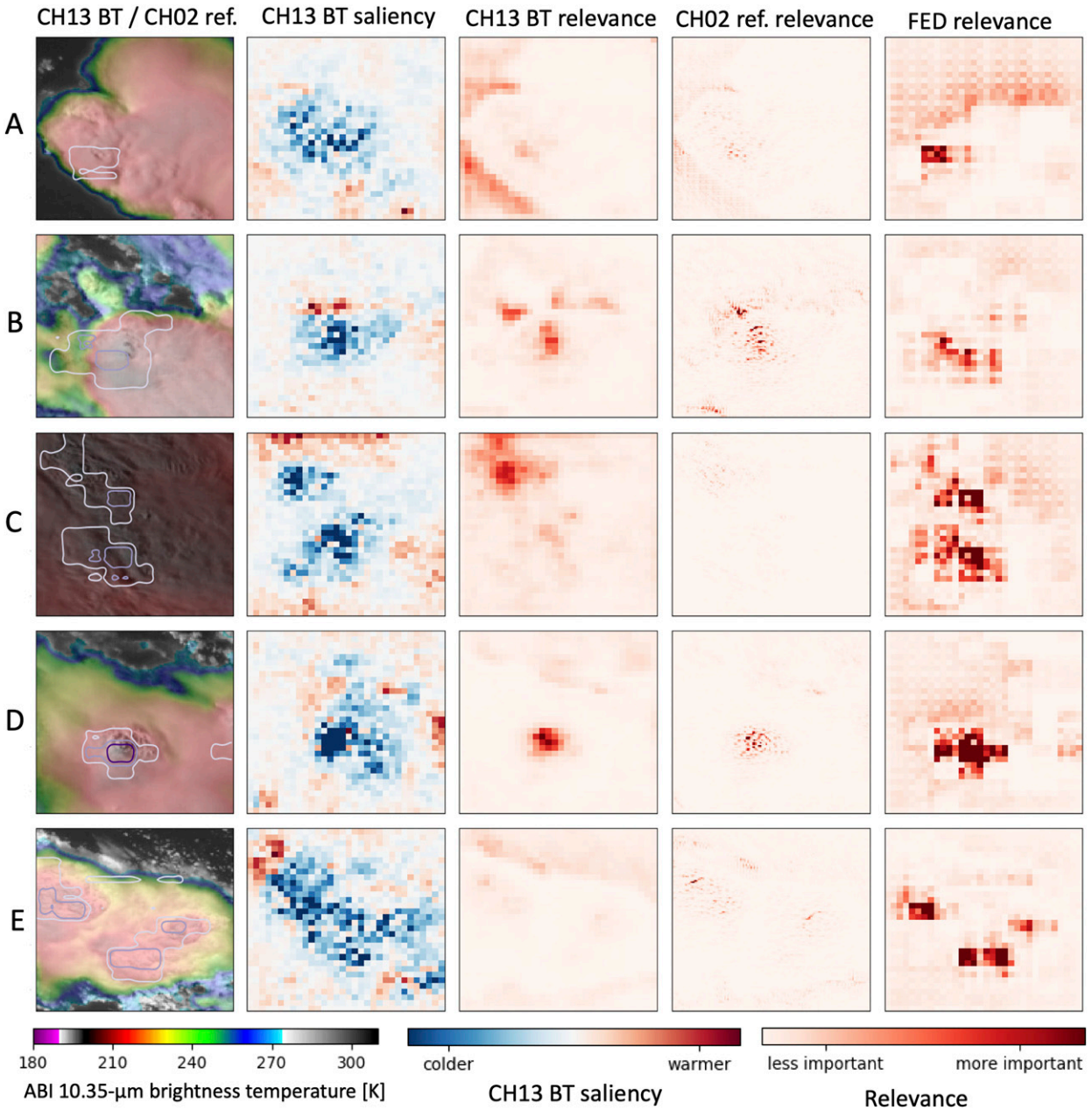

FIG. 17. As in Fig. 15, but for five false positive storm samples from the validation dataset. The predicted probability for each image patch was >87%.

storms with AACPs are correctly identified as intense convection (see Figs. 10–14 and Cintineo 2019). While not explicitly mapped out by the diagnostic tools, the driving force behind the AACP (an intense overshoot) and its infrared presentation (cold-U) are key features identified by the model.

d. Permutation tests

Two permutation tests were performed on the trained CNN using Keras code examples from Lagerquist and Gagne (2019). The cost function used in the permutation test is negative AUC; since the permutation test aims to minimize the cost function, in this case it aims to maximize AUC. Since ABI CH02 0.64-μm reflectance is a trivial predictor after sunset, the permutation tests were performed on storm samples from the validation dataset where the solar-zenith angle was less than 85° (n = 36 900). The B01 method can be thought of as, "the skill as a result of permuting *only* the kth predictor," with the kth-most important predictor resulting in the kth-greatest decrease in AUC. The L15 method can be thought of as, "the skill as a result of permuting the kth predictor *and* each more important predictor."

For both the L15 and B01 permutation methods (see Fig. 19), the "No permutation" bar represents the original

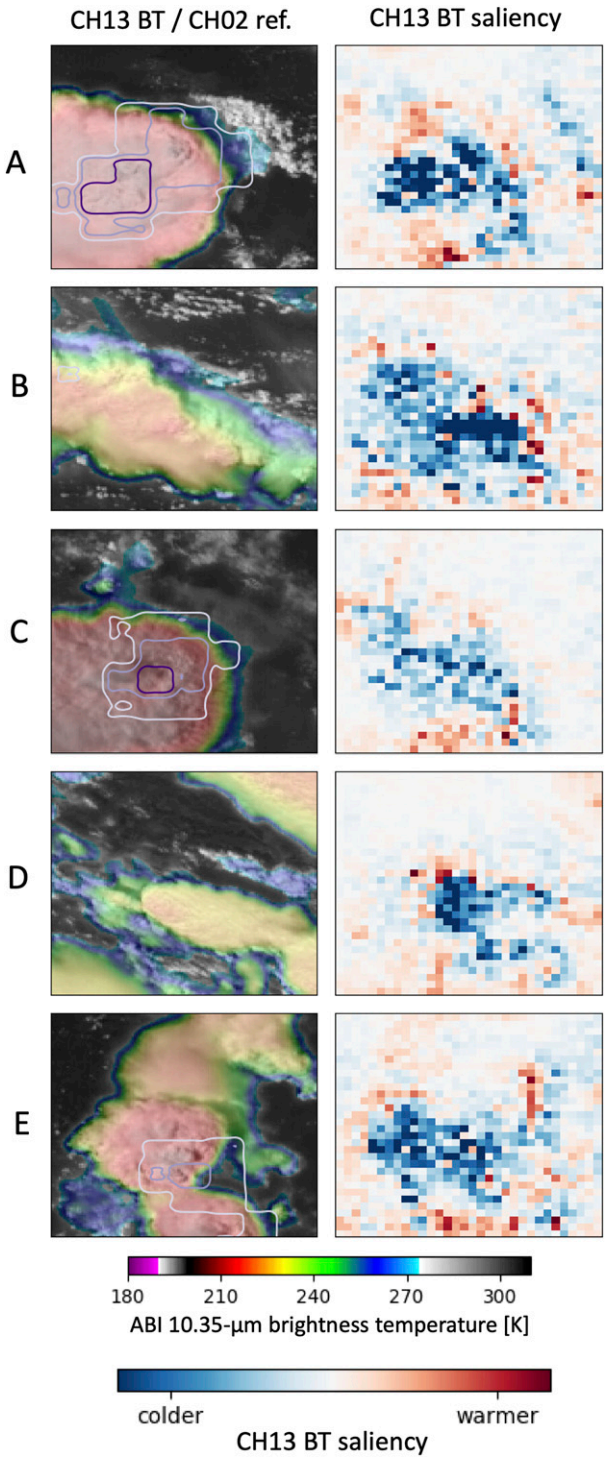

FIG. 18. Saliency plots for ABI CH13 brightness temperature (BT) for five false negative storms. The images in the first column are CH13 BT/CH02 reflectance "sandwich" imagery with GLM flash-extent density contours for 10, 20, and 40 flashes per 5 min overlaid in shades of purple. The predicted probability for each image patch was <1%.

AUC value for the full CNN model for this daytime-only sample (AUC = 0.986). Since the first step of the L15 method is identically the B01 method, it was found that ABI CH13 brightness temperature was the most important predictor for both methods, with an AUC = 0.700 after permutation of that channel. The B01 method found that CH02 reflectance and FED were the next two important predictors, followed by the four scalar predictors. The L15 method found that FED was the 2nd most important predictor, followed by CH02. This is likely due to the high correlation that exists between the ABI channels, relative to the correlation between ABI channels and FED. The satellite-zenith angle was the fourth most important predictor in each test, while the mean latitude and mean longitude were swapped in importance for the two methods. By itself, the B01 test shows that CH02 reflectance contains more information than FED, but the L15 test reveals that once CH13 brightness temperature samples are randomized, the CH02 reflectance does not contain as much additional or independent information as FED.

4. Discussion and conclusions

A machine-learning model that exploits the rich spatial and spectral information provided by the *GOES-16* ABI and the lightning mapping provided by the *GOES-16* GLM was developed with the goal of automatically identifying intense midlatitude convection consistent with human expert interpretation of the satellite images. Over 220 000 images were manually labeled to enable training of a convolutional neural network (CNN), which learned to make skillful predictions of intense convection both day and night.

The CNN learned several features, identifiable using a combination of high-resolution visible reflectance imagery, infrared window imagery, and lightning mapping imagery, that are known to be associated with intense convection. Model diagnostic tools and examples showed that the model learned to recognize overshooting tops, portions of cold-U thermal patterns, cold rings, and robust lightning activity near updraft regions. The model also learned to recognize cloud-edge brightness temperature gradients as important, which is a new and unique result for satellite interpretation of intense convection. Strong gradients in the brightness temperature may arise from a combination of a strong overshooting top associated with the updraft region (producing a very cold minimum in the brightness temperature) and strong upper-level winds limiting upstream propagation of the anvil cloud. The strength of the upper-level winds may be an indication of the amount deep-layer shear, well known to either support or diminish intense convection. However, further analysis of model attributes and associated physical processes is the subject of future work.

A successive rank permutation test revealed that the most important predictors were ABI infrared-window brightness temperature (first), GLM flash-extent density imagery (second), and ABI visible reflectance (third), while the permutation test with replacement (Breiman 2001) ranked the top three predictors as ABI infrared-window brightness temperature (first), ABI visible reflectance (second), and GLM flash-extent density (third). Despite the potential for reduced lightning detection efficiency in

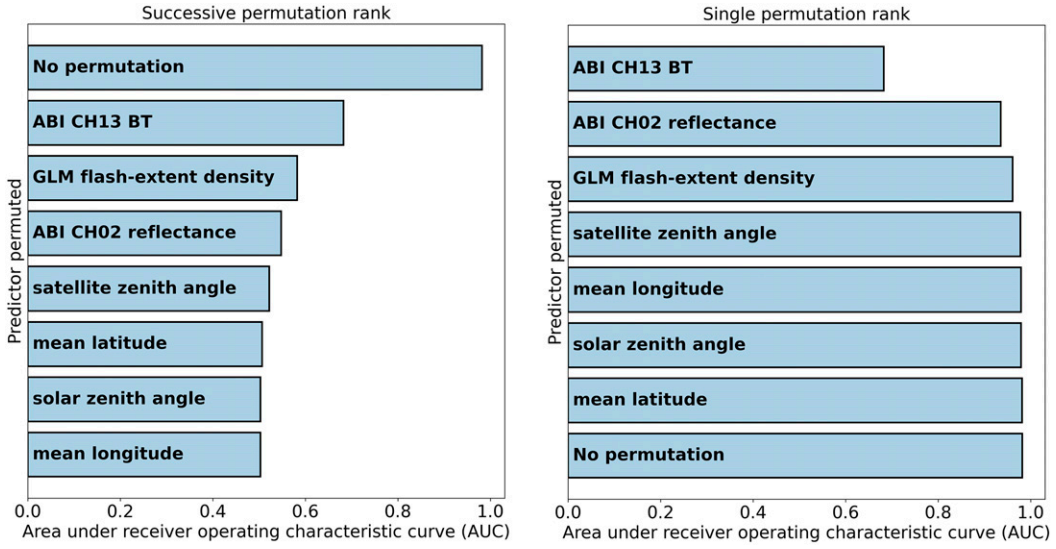

FIG. 19. The (left) successive permutation rank test and (right) single permutation rank test for the trained convolutional neural network of this paper, using 36 900 daytime-only storm samples from the validation dataset. "No permutation" represents the original AUC without any predictors permuted. For both panels, predictors increase in importance moving from the bottom to the top. ABI = Advanced Baseline Imager; GLM = Geostationary Lightning Mapper; BT = brightness temperature.

some optically deep clouds (e.g., Burnett et al. 2018), the GLM added appreciable skill to the model. The model also exhibited gainful lead time to initial severe weather reports in the testing dataset (median of 21 min with a 95% confidence interval of 18–26 min, measured from the 90% probability threshold).

Given the clear importance of the infrared-window brightness temperature, there may be interest in the community to train a model with solely infrared imager channels, for the goal of deploying it on the historical record of spaceborne infrared imager observations to investigate climatological trends in intense convection. However, more work is needed to discern if the infrared-window brightness temperature alone can faithfully identify intense convection, and what the effect of reduced spatial resolution may be (many historical imagers have 4-km horizontal resolution for infrared channels), which could prove significant.

The CNN model output could complement and enhance radar interrogation of storms and could be incorporated into nowcasting applications such as ProbSevere, with the goal of improving severe weather warnings. The CNN model may be particularly valuable for diagnosing and nowcasting convection in regions with limited or no radar coverage (see the example from Alaska, Fig. 14). In addition, the model shows promise for generalizing to other satellite sensors with very similar spectral and spatial attributes to the GOES-R ABI (e.g., *Himawari-8*, Meteosat Third Generation, and *GEO KOMPSAT-2A*), eventually allowing for objective large-scale monitoring of convection, and possibly providing a new framework for the study of convection.

Acknowledgments. The authors acknowledge the National Oceanic and Atmospheric Administration (NOAA) High Performance Computing and Communications Program (HPCC) for supporting this research. We are also grateful to Eric Bruning (Texas Tech University) for his assistance with the glmtools software package, as well as to three anonymous reviewers for their helpful comments, improving the manuscript. The *GOES-16* data used in this study can be freely obtained from NOAA's Comprehensive Large Array Data Stewardship System (CLASS; online at https://www.class.noaa.gov/). The ProbSevere data in this study may be obtained from UW-CIMSS upon request of the corresponding author. The views, opinions, and findings contained in this paper are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. government position, policy, or decision.

REFERENCES

Adler, R. F., and D. D. Fenn, 1979: Thunderstorm intensity as determined from satellite data. *J. Appl. Meteor.*, **18**, 502–517, https://doi.org/10.1175/1520-0450(1979)018<0502:TIADFS>2.0.CO;2.

Alber, M., and Coauthors, 2019: iNNvestigate neural networks! *J. Mach. Learn. Res.*, **20**, 1–8.

Apke, J. M., J. R. Mecikalski, and C. P. Jewett, 2016: Analysis of mesoscale atmospheric flows above mature deep convection using super rapid scan geostationary satellite data. *J. Appl. Meteor. Climatol.*, **55**, 1859–1887, https://doi.org/10.1175/JAMC-D-15-0253.1.

Bachmeier, S., 2019: NWS Juneau, Alaska issues their first-ever severe thunderstorm warning—Based on satellite imagery. CIMSS Satellite Blog, accessed 1 December 2019. https://cimss.ssec.wisc.edu/satellite-blog/archives/34080.

Bedka, K., and K. Khlopenkov, 2016: A probabilistic multispectral pattern recognition method for detection of overshooting cloud tops using passive satellite imager observations. *J. Appl. Meteor. Climatol.*, **55**, 1983–2005, https://doi.org/10.1175/JAMC-D-15-0249.1.

——, E. M. Murillo, C. R. Homeyer, B. Scarino, and H. Mersiovsky, 2018: The above-anvil cirrus plume: An important severe

weather indicator in visible and infrared satellite imagery. *Wea. Forecasting*, **33**, 1159–1181, https://doi.org/10.1175/WAF-D-18-0040.1.

Binder, A., G. Montavon, S. Lapuschkin, K. R. Muller, and W. Samek, 2016: Layer-wise relevance propagation for neural networks with local renormalization layers. *Int. Conf. on* Artificial *N*eural Networks and *M*achine *L*earning—*ICANN* 2016, Barcelona, Spain, Springer, 63–71.

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

Bruning, E., 2019: glmtools. GitHub, accessed 15 July 2018, https://github.com/deeplycloudy/glmtools.

Brunner, J. C., S. A. Ackerman, A. S. Bachmeier, and R. M. Rabin, 2007: A quantitative analysis of the enhanced-V feature in relation to severe weather. *Wea. Forecasting*, **22**, 853–872, https://doi.org/10.1175/WAF1022.1.

Burnett, C., R. F. Garret, W. M. MacKenzie Jr., M. Seybold, J. Fulbright, and J. D. Sims, 2018: GLM observations through optically deep clouds: A case study. *22nd Conf. on Satellite Meteorology and Oceanography*, Austin, TX, Amer. Meteor. Soc., 628, https://ams.confex.com/ams/98Annual/webprogram/Paper333686.html.

Chollet, F., 2015: Keras. GitHub, accessed 10 February 2019, https://github.com/keras-team/keras.

Cintineo, J. L., 2019: The probability of "intense convection" using geostationary satellite data. *CIMSS Satellite Blog*, S. Bachmeier, Ed., Cooperative Institute of Meteorological Satellite Studies, https://cimss.ssec.wisc.edu/satellite-blog/archives/34480.

——, M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecasting*, **29**, 639–653, https://doi.org/10.1175/WAF-D-13-00113.1.

——, and Coauthors, 2018: The NOAA/CIMSS ProbSevere model: Incorporation of total lightning and validation. *Wea. Forecasting*, **33**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

——, M. P. Pavolonis, J. M. Sieglaff, A. Wimmers, and J. C. Brunner, 2020: NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Efron, B., and R. Tibshirani, 1986: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1**, 54–75, https://doi.org/10.1214/ss/1177013815.

Fukushima, K., 1980: Neocognitron—A self-organizing neural network model for a mechanism of pattern-recognition unaffected by shift in position. *Biol. Cybern.*, **36**, 193–202, https://doi.org/10.1007/BF00344251.

Gagne, D. J., II, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

Goodfellow, I., Y. Bengio, and A. Courville, Eds., 2016: Back-propagation and other differentiation algorithms. *Deep Learning*, MIT Press, 200–220.

Goodman, S. J., and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper (GLM). *Atmos. Res.*, **125–126**, 34–49, https://doi.org/10.1016/j.atmosres.2013.01.006.

He, K. M., X. Y. Zhang, S. Q. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, IEEE, 770–778.

Hinton, G., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2012: Improving neural networks by preventing co-adaptation of feature detectors. arXiv e-prints, https://arxiv.org/pdf/1207.0580.pdf.

Hoerl, A. E., and R. W. Kennard, 2000: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42**, 80–86, https://doi.org/10.1080/00401706.2000.10485983.

Homeyer, C. R., J. D. McAuliffe, and K. M. Bedka, 2017: On the development of above-anvil cirrus plumes in extratropical convection. *J. Atmos. Sci.*, **74**, 1617–1633, https://doi.org/10.1175/JAS-D-16-0269.1.

Hsu, W. R., and A. H. Murphy, 1986: The attributes diagram—A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Ioffe, S., and C. Szegedy, 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Int. Conf. on Machine Learning*, Lille, France, JMLR, 448–456.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: ImageNet classification with deep convolutional neural networks. *Commun. ACM*, **60**, 84–90, https://doi.org/10.1145/3065386.

Lagerquist, R., and D. J. Gagne II, 2019: Interpretation of deep learning for predicting thunderstorm rotation: Python tutorial. GitHub, accessed 4 March 2019, https://github.com/djgagne/ams-ml-python-course/blob/master/module_4/ML_Short_Course_Module_4_Interpretation.ipynb.

——, A. McGovern, and D. J. Gagne, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Wea. Forecasting*, **34**, 1137–1160, https://doi.org/10.1175/WAF-D-18-0183.1.

——, ——, C. R. Homeyer, D. J. Gagne II, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Oceanic Technol.*, **32**, 1209–1223, https://doi.org/10.1175/JTECH-D-13-00205.1.

Li, F., J. Johnson, and S. Yeung, 2020: CS231n convolutional neural networks for visual recognition. GitHub, accessed 10 October 2019, http://cs231n.github.io/convolutional-networks/.

Li, Y., H. K. Zhang, X. Z. Xue, Y. A. Jiang, and Q. Shen, 2018: Deep learning for remote sensing image classification: A survey. *Wiley Interdiscip. Rev.*, **8**, e1264, https://doi.org/10.1002/widm.1264.

Litjens, G., and Coauthors, 2017: A survey on deep learning in medical image analysis. *Med. Image Anal.*, **42**, 60–88, https://doi.org/10.1016/j.media.2017.07.005.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Menzel, W. P., and J. F. W. Purdom, 1994: Introducing GOES-I—The 1st of a new-generation of geostationary operational environmental satellites. *Bull. Amer. Meteor. Soc.*, **75**, 757–782, https://doi.org/10.1175/1520-0477(1994)075<0757:IGITFO>2.0.CO;2.

Metz, C., 1978: Basic principles of ROC analysis. *Semin. Nucl. Med.*, **8**, 283–298, https://doi.org/10.1016/S0001-2998(78)80014-2.

Montavon, G., A. Binder, S. Lapuschkin, W. Samek, and K. Müller, 2019: Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,* W. Samek et al., Eds., Springer, 193–210, https://doi.org/10.1007/978-3-030-28954-6_10.

Nair, V., and G. Hinton, 2010: Rectified linear units improve restricted Boltzmann machines. *Proc. 27th Int. Conf. on Machine Learning*, Haifa, Israel, International Machine Learning Society, 807–814.

NOAA, 2019: Storm events database. NOAA, accessed 12 January 2020, https://www.ncdc.noaa.gov/stormevents/.

Purdom, J., 1976: Some uses of high-resolution GOES imagery in mesoscale forecasting of convection and its behavior. *Mon. Wea. Rev.*, **104**, 1474–1483, https://doi.org/10.1175/1520-0493(1976)104<1474:SUOHRG>2.0.CO;2.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. MICCAI 2015: *Medical Image Computing and Computer-Assisted Intervention*, N. Navab et al., Eds., Springer, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

Rudlosky, S. D., S. J. Goodman, K. S. Virts, and E. C. Bruning, 2019: Initial geostationary lightning mapper observations. *Geophys. Res. Lett.*, **46**, 1097–1104, https://doi.org/10.1029/2018GL081052.

Sandmæl, T. N., C. R. Homeyer, K. M. Bedka, J. M. Apke, J. R. Mecikalski, and K. Khlopenkov, 2019: Evaluating the ability of remote sensing observations to identify significantly severe and potentially tornadic storms. *J. Appl. Meteor. Climatol.*, **58**, 2569–2590, https://doi.org/10.1175/JAMC-D-18-0241.1.

Schmidhuber, J., 2015: Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.

Schmit, T. J., M. M. Gunshor, W. P. Menzel, J. J. Gurka, J. Li, and A. S. Bachmeier, 2005: Introducing the next-generation advanced baseline imager on GOES-R. *Bull. Amer. Meteor. Soc.*, **86**, 1079–1096, https://doi.org/10.1175/BAMS-86-8-1079.

——, and Coauthors, 2015: Rapid refresh information of significant events: Preparing users for the next generation of geostationary operational satellites. *Bull. Amer. Meteor. Soc.*, **96**, 561–576, https://doi.org/10.1175/BAMS-D-13-00210.1.

Schultz, C. J., W. A. Petersen, and L. D. Carey, 2011: Lightning and severe weather: A comparison between total and cloud-to-ground lightning trends. *Wea. Forecasting*, **26**, 744–755, https://doi.org/10.1175/WAF-D-10-05026.1.

Setvák, M., and Coauthors, 2010: Satellite-observed cold-ring-shaped features atop deep convective clouds. *Atmos. Res.*, **97**, 80–96, https://doi.org/10.1016/j.atmosres.2010.03.009.

——, K. Bedka, D. T. Lindsey, A. Sokol, Z. Charvat, J. St'astka, and P. K. Wang, 2013: A-Train observations of deep convective storm tops. *Atmos. Res.*, **123**, 229–248, https://doi.org/10.1016/j.atmosres.2012.06.020.

Simonyan, K., A. Vedaldi, and A. Zisserman, 2014: Deep inside convolutional networks: Visualising image classification models and saliency maps. *Workshop at Int. Conf. on Learning Representations*, Banff, Canada, ICLR, 1–8, https://iclr.cc/archive/2014/workshop-proceedings/.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Valachová, M., and M. Setvák, 2017: Satellite monitoring of the convective storms: Forecasters' point of view. Czech Hydrometeorological Institute, 48 pp., https://www.ecmwf.int/sites/default/files/elibrary/2017/17282-satellite-monitoring-convective-storms-forecasters-point-view.pdf.

Wang, P. K., 2003: Moisture plumes above thunderstorm anvils and their contributions to cross-tropopause transport of water vapor in midlatitudes. *J. Geophys. Res.*, **108**, 4194, https://doi.org/10.1029/2002JD002581.

——, 2007: The thermodynamic structure atop a penetrating convective thunderstorm. *Atmos. Res.*, **83**, 254–262, https://doi.org/10.1016/j.atmosres.2005.08.010.

——, K. Y. Cheng, M. Setvak, and C. K. Wang, 2016: The origin of the gullwing-shaped cirrus above an Argentinian thunderstorm as seen in CALIPSO images. *J. Geophys. Res. Atmos.*, **121**, 3729–3738, https://doi.org/10.1002/2015JD024111.